研究概要報告書

資料－6

研究題名	グラフサーチの技法を用いた音声認識アルゴリズム高度化の研究	報告書作成者	好田正紀
研究従事者	加藤正治，門前聖康，山口達也，大関雅和，塚貴明		
研究目的	大語彙・連続音声を対象とする音声認識では，必然的に膨大となる処理量に対して現実的な時間内で実行可能となるように，認識アルゴリズムをより高度化するための研究が本質的・潜在的に重要である． ＨＭＭ（hidden Markov model）の Viterbiアルゴリズムでは、入力音声とＨＭＭのマッチング領域の格子点を節点とするグラフにおいて、状態遷移確率と出力確率の累積値をスコアとして、モデルのタイプに対応した経路展開の規則のもとで、始点から終点に到達するスコア最大の経路を探索する．このことから、ＨＭＭによる音声認識は、本質的にグラフサーチの問題である． ＨＭＭによる音声認識をグラフサーチの問題とみなし、ビームサーチの技法による経路の展開に関して、多くの場合、当該節点までのスコアのみに基づいて枝刈の判定が行われた．当該節点以降の推定スコアも考慮して枝刈の判定を行うと、より大きな効果が期待される．例えば、のforward-backwardサーチでは、forward-passで音素コンテクスト独立ＨＭＭを用いた簡略な認識処理を行い、その結果を推定スコアに利用して backward-passで音素コンテクスト依存ＨＭＭを用いたN-best 候補の探索の高速化をはかっている．ただし、backward-pass の探索法はbest-firstサーチではなくてビームサーチであり、また、推定スコアは最適解を保証していない． best-firstサーチの技法による経路の展開に関しては、 A*探索の考え方に基づく研究が盛んに行われるようになった．連続音声認識に適用したスタックデコーディング法では、長さの異なる単語列の評価関数における正規化係数をヒューリスティックに設定する、しきい値やビーム幅による枝刈も併用する、探索終了判定の条件を緩和する、等により、厳密なA* 探索に必ずしもこだわらない、実用的な探索法が検討されている．tree-trellisサーチでは、forward-passでtrellis サーチにより第一候補を求めた後、そのトレリスを推定スコアに利用してbackward-pass でtreeサーチによりN-best 候補の探索の高速化をはかっている． 本研究の目的は、best-firstサーチの技法により認識処理（第一候補の探索）自体の高速化をはかることにある．これまでに、ＤＴＷ（dynamic time warping）による音声認識において、推定コストを当該節点以降の入力フレームのＶＱ歪に基づいて設定して、経路の展開を標準パターンすべてに対して一括して行う、ＤＰbest-firstサーチのアルゴリズムを提案した．また、推定コストの精度を上げるために、入力音声とＶＱ標準パターンのＤＴＷを経路展開のbest-firstサーチとは逆に終点から始点に向って直接行うことによる推定コスト設定法、及び、この後向きＤＴＷにもbest-firstサーチを利用する推定コスト設定法を提案した． 本研究では、ＤＰbest-firstサーチのアルゴリズム及び推定コスト設定法の考え方をViterbi アルゴリズムに適用し、最大経路スコアに基づく推定スコア設定法、及び、単純な音素ＨＭＭを利用する推定スコア設定法によるViterbi best-firstサーチのアルゴリズムを提案する．Viterbi best-firstサーチは、推定スコアを適切に設定すれば、認識率を低下させずに、認識処理で主要な部分を占める経路展開の計算量が１％以下となり、計算量低減の効果が非常に大きいことを示す．		

様式－９

研究内容

1．Viterbi best-firstサーチ

1.1 Viterbi best-firstサーチの概要

入力音声の第ｉフレーム、単語ｎのＨＭＭの第ｊ状態を示す、トレリス上の節点を(i,j,n)で表す（i=1～Ｉ，j=1～Jn，n=1～N）．(i,j,n)で展開される子節点を、{(i+1,k,n)｜k=j,j+1}とする（図１参照）．これは、図２に示すように、left-to-right モデルで自己ループと隣接する状態への遷移のみをもつタイプのＨＭＭを用いることに対応する．(i,j,n)から(i+1,k,n)への経路を<i,j,k,n>、その経路スコアをｓ<i,j,k,n>で表す．

離散出力確率分布型のＨＭＭを考えて、入力音声の第ｉフレームのシンボルをXi、単語ｎのＨＭＭで状態ｊからｋへの遷移確率をａ(j,k,n)、その遷移でXiの出力確率をｂ(Xi,j,k,n)と表す．Viterbi アルゴリズムでは、経路スコアを

$$s\langle i-1,j,k,n\rangle = \log a(j,k,n) + \log b(Xi,j,k,n) \quad (1)$$

で与えて、トレリス上で経路スコアの累積値が最も大きい経路を探索する．

展開可能な節点のリスト（open list）をＰ、展開済みの節点のリスト（closed list）をＱで表し、それらの初期値を

Ｐ＝ {(0,1,n)｜n=1～N }

Ｑ＝NULL

とおく．Ｐの中からスコア最大の節点を取り出してＱに移し、そこで展開される子節点をＰに追加することを、繰り返す．そして、Ｐから取り出した節点が終点(Ｉ,Jn,n) であるならば、単語ｎを認識結果とする．

1.2 スコア関数

入力音声の第i_1～i_2フレームをＡ(i_1,i_2)、単語ｎのＨＭＭの第j_1～j_2状態をＢn(j_1,j_2)と表す．スコア関数の説明図を図３に示す．

節点(i,j,n) におけるスコア関数は

$$F(i,j,n) = G(i,j,n) + H(i,j,n) \quad (2)$$

と表せる．ここで、G(i,j,n) はトレリス上でＡ(1,i)とＢn(1,j)をViterbi アルゴリズムで対応づけた最適経路上の経路スコアの累積値である．同様に、H(i,j,n)はトレリス上でＡ(i+1,Ｉ)とＢn(j,Jn) をViterbi アルゴリズムで対応づけた最適経路上の経路スコアの累積値である．サーチの過程でＧ(i,j,n) やH(i,j,n)を正確に求めるわけにいかないので、実際には、それらの推定値に基づいてスコア関数を設定する．

節点(i,j,n) におけるスコア関数の推定値を

$$f(i,j,n) = g(i,j,n) + h(i,j,n) \quad (3)$$

とする．ここで、ｇ(i,j,n) はＡ(1,i) とＢn(1,j)を探索範囲内で最も良く対応づけた経路上の経路スコアの累積値である．探索範囲

研究内容

外の経路でさらに大きな累積値を与える可能性が残っているので、g(i,j,n) はG(i,j,n) に等しいとは限らない．一般に、

$$g(i,j,n) \leqq G(i,j,n) \quad (4)$$

である．h(i,j,n) はH(i,j,n) の推定値であり、以下では、推定スコアと呼ぶ．

1.3 推定スコアに対する条件

推定スコアh(i,j,n) が

$$h(i,j,n) \geqq H(i,j,n) \quad (5)$$

の関係を満たせば A*探索となり、最適解が保証される．

さらに単調性の制約条件

$$h(i,j,n) \geqq h(i+1,k,n) + s\langle i,j,k,n\rangle \quad (6)$$

を満たせば、Qに移した節点に対して

$$g(i,j,n) = G(i,j,n)$$

が成り立つ．この場合には、Qの節点は、その後再度展開しなくてよいことが保証される．

２．推定スコアの設定法

2.1 経路スコア

g(i,j,n) の計算では、式(1) の経路スコアを累積する．このことは、Viterbi best-firstサーチの認識実験において一貫している．

一方、推定スコアh(i,j,n) の設定に用いる経路スコアとしては、式(1) に必ずしも拘束されるわけではない．例えば、a(j,k,n) を一定値とする経路スコア

$$s\langle i-1,j,k,n\rangle = \log 1/2 + \log b(Xi,j,k,n) \quad (7)$$

あるいは、a(j,k,n) の項を無視する経路スコア

$$s\langle i-1,j,k,n\rangle = \log b(Xi,j,k,n) \quad (8)$$

を考えることもできる．

2.2 推定スコア

ここで取り上げる推定スコア設定法は、次の６通りである．推定スコア①、③、④は式(5) の A*探索の条件と式(6) の単調性の制約条件をともに満たす．推定スコア⑤、⑥は推定スコアの精度を上げるために時間軸の順序関係を考慮するものである．

＜推定スコア①＞

研究内容

最も安直な推定スコアは、

$$h(i,j,n) = 0 \qquad (9)$$

と設定することである.

<推定スコア②>

節点(i,j,n)までのスコアを利用して、推定スコアを

$$h(i,j,n) = \frac{g(i,j,n)}{i} \times (I-i) \qquad (10)$$

と設定する. これは、経路スコアの累積が、第iフレームまでと同じ割合でそれ以降も続くと仮定したものである.

この推定スコアを実際に用いる場合には、スコア関数が

$$f(i,j,n) = \frac{g(i,j,n)}{i} \times I \qquad (11)$$

となることから、g(i,j,n)をフレーム数iで割った値が最大の節点が展開される.

<推定スコア③>

式(5)の条件を満たすように推定スコアを設定する簡便な一つの方法は、第1～Jn状態内の遷移における経路スコアの最大値を、入力音声の第iフレーム以降の各々について求め、それらを累積して、

$$h(i,j,n) = \sum_{i'=i}^{I-1} \left(\max_{\substack{j'=1\sim(Jn-1) \\ k=j',j'+1}} s\langle i',j',k,n\rangle \right) \qquad (12)$$

とおくことである. 上式の計算に関係する経路の領域を図4(a)に示す.

この推定スコアを実際に用いる場合には、式(12)の右辺を式の表現通りに計算するのではなくて、次の点を考慮すると能率良く計算できる. 一つは、同じ音素が認識対象の語彙中に何度も表れることを考慮して、まず音素毎に音素HMM内の遷移に関して経路スコアの最大値（音素内最大経路スコア）を求めておき、その結果を用いて単語nのHMMを構成する各音素の音素内最大経路スコアを比較する. もう一つは、同じシンボルが入力フレーム中に何度も表れることを考慮して、シンボル毎に上記の音素内最大経路スコアをあらかじめ求めておき、その結果を用いて入力音声の各フレームのシンボルに対応する音素内最大経路スコアを比較する.

<推定スコア④>

同様のもう一つの方法は、第j～Jn状態内の遷移における経路スコアの最大値を、入力音声の第iフレーム以降の各々について求め、

研究内容

それらを累積して、

$$h(i,j,n)=\sum_{i'=i}^{I-1}\left(\max_{\substack{j'=j\sim(Jn-1)\\ k=j',j'+1}} s\langle i',j',k,n\rangle\right) \qquad (13)$$

とおくことである．上式の計算に関係する経路の領域を図４(b) に示す．

推定スコア③、④の設定法はいずれも、入力音声のシンボル列と単語ＨＭＭの状態列の間の時間軸の順序関係を無視している．推定スコアの精度をさらに上げるためには、時間軸の順序関係を考慮することが必要である．

推定スコア設定に、図２(b) に示す２状態１ループの、より単純な音素ＨＭＭを利用することを考える．２状態１ループの音素ＨＭＭを連結した単語ｎのＨＭＭを用いて、入力音声に通常のViterbi アルゴリズムを適用し、最適経路を得る．この最適経路上の第１～ｉフレームのViterbi スコアを$\overline{g}(i,n)$ と表す．

＜推定スコア⑤＞

最適経路から入力音声を音素単位に分割し、第ｉフレームが第ｍセグメントに属するならば、単語ｎのＨＭＭを構成するｍ番目の音素ＨＭＭ（４状態３ループ）内の最大経路スコアを、入力音声の第ｉフレーム以降の各々について求め、それらを累積して、推定スコア③、④と同様にして$h(i,j,n)$ を設定する．推定スコア⑤の計算に関係する経路の領域を図４(c) に示す．

＜推定スコア⑥＞

最適経路上のViterbi スコアに基づいて、

$$h(i,j,n) = \overline{g}(I,n) - \overline{g}(i,n) \qquad (14)$$

とおく．

2.3 他の研究との関係

推定スコア①、②はいずれも、当該節点までのスコアのみに基づくbest-firstサーチであり、$g(i,j,n)$ の値、あるいは、$g(i,j,n)$ をフレーム数ｉで割った値が最大の節点から展開される．推定スコア①によるbest-firstサーチは最も単純な方法、推定スコア②によるbest-firstサーチは連続音声認識の分野で時々用いられる方法であるが、ここでは、本研究で提案する方法との比較のためにそれらを取り上げる．

推定スコア③、④の設定法は、ＤＰbest-firstサーチにおいてＶＱ歪を利用する推定コスト設定法と同様の考え方に基づいて、ここでは、単語内最大経路スコアを利用する方法を提案したものである．推定スコア③は、推定スコア設定のための計算量が小さく、有用であろう．一方、推定スコア④は、推定スコア③より精度が良いものの、推定スコア設定のための計算量が大きいために、実用的では

研究内容

ないといえよう．
推定スコア⑤、⑥は、forward-backwardサーチにおいてより単純なモデルを利用する推定スコア設定法と同様の考え方に基づいて、ここでは、より単純なモデルとして２状態１ループの音素ＨＭＭを取り上げる．

３．特定話者単語音声認識の実験条件
(1) 音声資料、音声分析、符号帳の作成
ＡＴＲ音声データで、男性１名(MHT) が発声した重要語5240語と音素バランス単語 216語の音声資料を用いる．
標本化周波数 12kHzで16ビットに量子化し，フレーム長32ms，分析周期 8msに設定し，ハミング分析窓を用い，$1-Z^{-1}$の特性で高域強調したのち， 1〜16次のＬＰＣケプストラム係数を抽出する．
符号帳は音素バランス単語 216語の音声資料から作成する．符号帳サイズは 256とする．
(2) 音素ＨＭＭの作成
40種類の音素についてそれぞれ、語頭・語中別に、４状態３ループと２状態１ループの２つのタイプの離散出力確率分布型ＨＭＭを作成する．２状態１ループの音素ＨＭＭは、推定スコア⑤、⑥の設定のためにのみ用いる．音素ＨＭＭの学習は、重要語5240語の偶数番目の単語の音声資料（各音素最大 400サンプル）を用いてBaum-Welchアルゴリズムで行う．学習回数は20回とする．
(3) 計算量の評価
経路展開の計算量は、展開された経路数の全経路数に対する割合（%）で評価する．推定スコア設定の計算量は、H(i,j,n) を得るための計算と比較して評価する．

４．実験結果と考察
4.1 実験結果
推定スコア設定法、及び、探索アルゴリズムの種々の組み合わせについて、音素バランス単語 216語の音声資料を用いた単語認識を行い、Viterbi best-firstサーチによる計算量と認識誤り数を求めた．結果を表１に示す．計算量は 216単語の平均値を示す．() 内の数字は 216サンプル中の認識誤り数を示す．経路スコアの欄の(a,b)、(1/2,b)、(-,b) はそれぞれ、a(j,k,n) をそのまま用いる場合、一定値とする場合、無視する場合に対応する．
表１の主な結果を図６にまとめた．図６の横軸は推定スコア設定法の違いを示し、縦軸は経路展開の計算量を、前向きの一方向サーチ、及び、推定スコア設定に反対方向のスコアを利用する双方向サーチの場合について示す．そのときの認識誤り数も示す．
4.2 処理例

研究内容	推定スコア①～⑥のそれぞれを用いる場合について、処理例を図８に示す．図８において、上段のグラフは、展開された経路数の入力フレームによる変化を示す．従って、このグラフと横軸で挟まれた領域の大きさは経路展開の計算量に対応する．推定スコア①を用いる場合の経路展開の計算量が突出して大きいこと、推定スコア⑤、⑥を用いる場合の経路展開の計算量が極めて小さいことがよくわかる．中段のグラフは、展開された経路の属する単語ＨＭＭの種類数（すなわち、単語候補数）の入力フレームによる変化を示す．単語候補数が（結果として）１個になった時点を短い縦線で示す．図８の処理例で推定スコア②～⑥を用いる場合、入力フレームの半ば以降は正解単語ＨＭＭに対して展開された経路のみであることがわかる．下段の図は、正解単語ＨＭＭに対して展開された経路の領域を示す．図中の太線は最適経路を示す． **4.3 考察** 実験結果から次のことがわかる． (1) Viterbi best-firstサーチは、推定スコアを適切に設定すれば、認識率を低下させずに、経路展開の計算量が１％以下となり、計算量低減の効果が非常に大きい． (2) 推定スコア０とおく方法は、最適解が保証されるものの、予想された通り、経路展開の計算量が著しく増加する． (3) 推定スコア②は、認識誤り数が著しく増加し、単語音声認識では有用とはいえない．この結果は、経路スコアの累積が当該節点までと同じ割合でそれ以降も続くという仮定が単純すぎて、現実と合っていないことを示す． (4) 経路展開の計算量と推定スコア設定の計算量の両方を考慮すると、単語内最大経路スコアに基づく推定スコア③が最も良い．この推定スコアは、計算が簡単であるにもかかわらず、精度がかなり良く、h(0,1,n)による単語認識で約55％の認識率が得られている．また、A* 探索の条件を満たすので、最適解も保証される． (5) 推定スコア④は、推定スコア③より精度が良いので経路展開の計算量が小さいものの、推定スコア設定の計算量がH(i,j,n) を得るための計算と同等であるために、実用的ではない． (6) 単純な音素ＨＭＭを利用する推定スコア⑤、⑥は、最適解の保証はないものの、時間軸の順序関係が考慮されるので精度が良く、h(0,1,n) による単語認識では90％以上の高い認識率が得られている．そのため、経路展開の計算量低減の効果が最も大きく、 わずか 0.05～0.2％の計算量である．しかしながら、推定スコア設定において、２状態１ループの音素ＨＭＭを用いて最適経路を得るために大きな計算量を必要とすることが難点である． (7) 推定スコア設定に用いる経路スコアにおけるa(j,k,n)は、一定値とする場合、そのまま用いる場合、無視する場合の順で計算量低減の効果が大きい．a(j,k,n)を一定値とする場合と無視する場合の経路スコアの差はlog 1/2 であるので、推定スコアh(i,j,n) の差は (I-i)log 1/2 となる．それらで計算量低減の効果が異なるのはこの差による．なお、上記の差は負であるので、a(j,k,n)を無視する場合の推定スコアは一定値とする場合の推定スコアより必ず大きくなる．a(j,k,n) を無視する場合に計算量低減の効果が劣るのは、

研究内容	推定スコアが大きくなりすぎて、精度が悪いためである．一方、a (j, k, n) を一定値とする場合には、式(5)の条件を満たすとは限らないので、最適解が保証されないことに注意しなければならないが、推定スコア③においては実用上ほとんど問題なさそうである． (8) 双方向サーチは、推定スコア設定に反対方向のスコアを利用しないと、一方向サーチより計算量が大きくなり、有用とはいえない．一方、推定スコア設定に反対方向のスコアを利用すると、一方向サーチより計算量低減の効果が大きくなり得るが、探索アルゴリズムにおける経路展開の判定処理が一方向サーチと比較して煩雑になることが難点である． ５．むすび ＨＭＭのViterbi アルゴリズムによる音声認識をグラフサーチの観点から検討し、best-firstサーチの技法による経路の展開に関して、最大経路スコアに基づく推定スコア設定法、及び、単純な音素ＨＭＭを利用する推定スコア設定法を提案した．Viterbi best-firstサーチは、推定スコアを適切に設定すれば、認識率を低下させずに、認識処理で主要な部分を占める経路展開の計算量が１％以下となり、計算量低減の効果が非常に大きいことを示した．単純な音素ＨＭＭを利用する推定スコアは、時間軸の順序関係が考慮されるので精度が良いが、推定スコア設定に大きな計算量を必要とする．経路展開の計算量と推定スコア設定の計算量の両方を考慮すると、単語内最大経路スコアに基づく推定スコアが最も良い．この推定スコアは、A* 探索の条件を満たすので、最適解も保証される． 本研究のViterbi best-firstサーチのアルゴリズム及び推定スコア設定法の考え方を、ワードスポッティングや連続音声認識、連続出力確率分布型ＨＭＭを用いるViterbi best-firstサーチにも適用できるようにすることは今後の課題である．

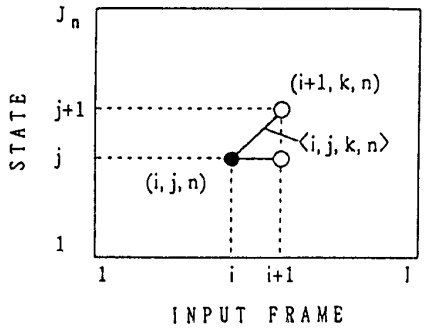


図 1　トレリス上の経路展開

Fig. 1 An expansion of paths on trellis.

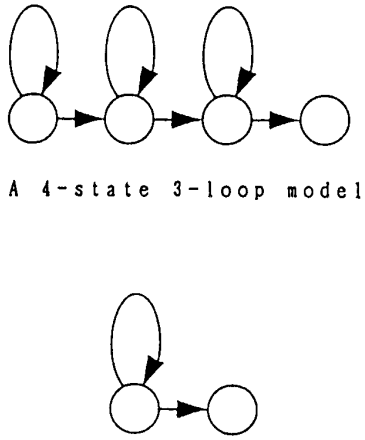


図 2　ＨＭＭ音素モデル

Fig. 2 HMM phoneme models.

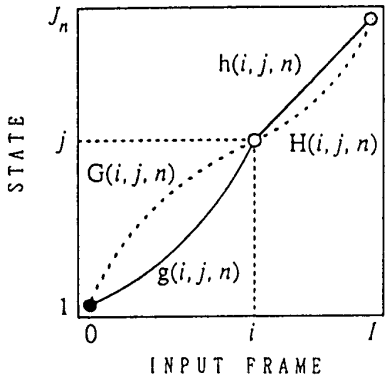
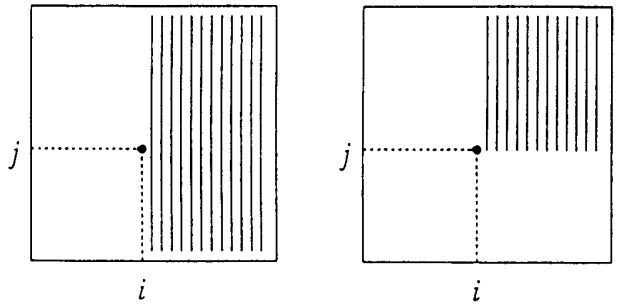


図 3　スコア関数

Fig. 3 A score function.

（注：　フローチャート図，ブロック図，構成図，写真，データ表，グラフ等　研究内容の補足説明に御使用下さい）



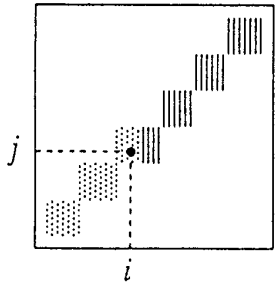


図４　推定スコアの計算における経路の領域

Fig. 4　An area of nodes used for computation of score estimates.

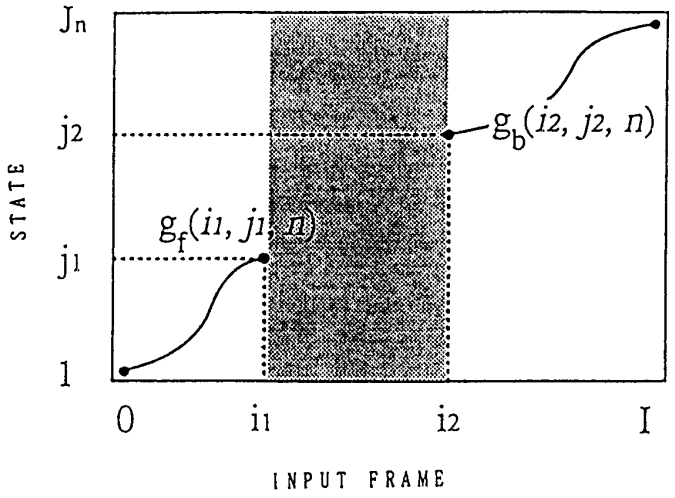


図５　双方向サーチにおける推定スコア設定法

Fig. 5　Score-estimating algorithm in the best-first bidirectional search.

（注：　フローチャート図，ブロック図，構成図，写真，データ表，グラフ等　研究内容の補足説明に御使用下さい）

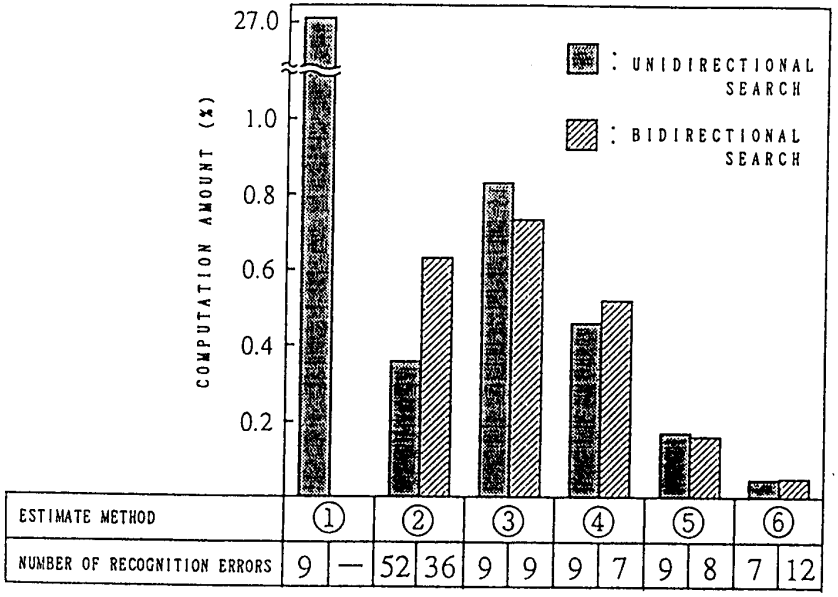


図 6　推定スコア設定法と計算量

Fig. 6 Score-estimating algorithm vs. computation amount.

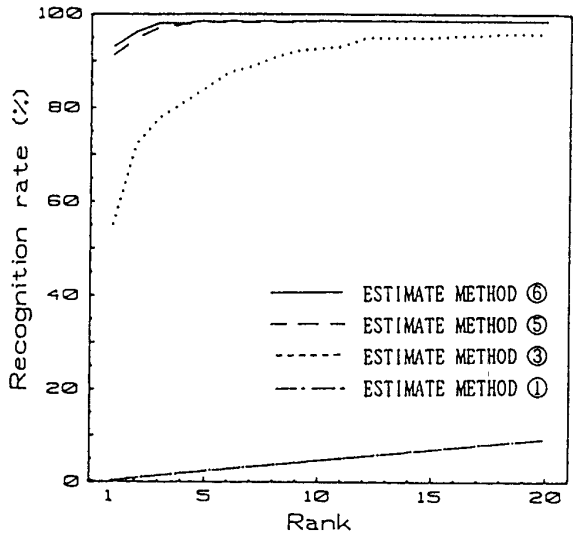


図 7　推定スコア h (0, 1, n) による 216単語認識

Fig. 7　216-word recognition using score estimate h(0, 1, n).

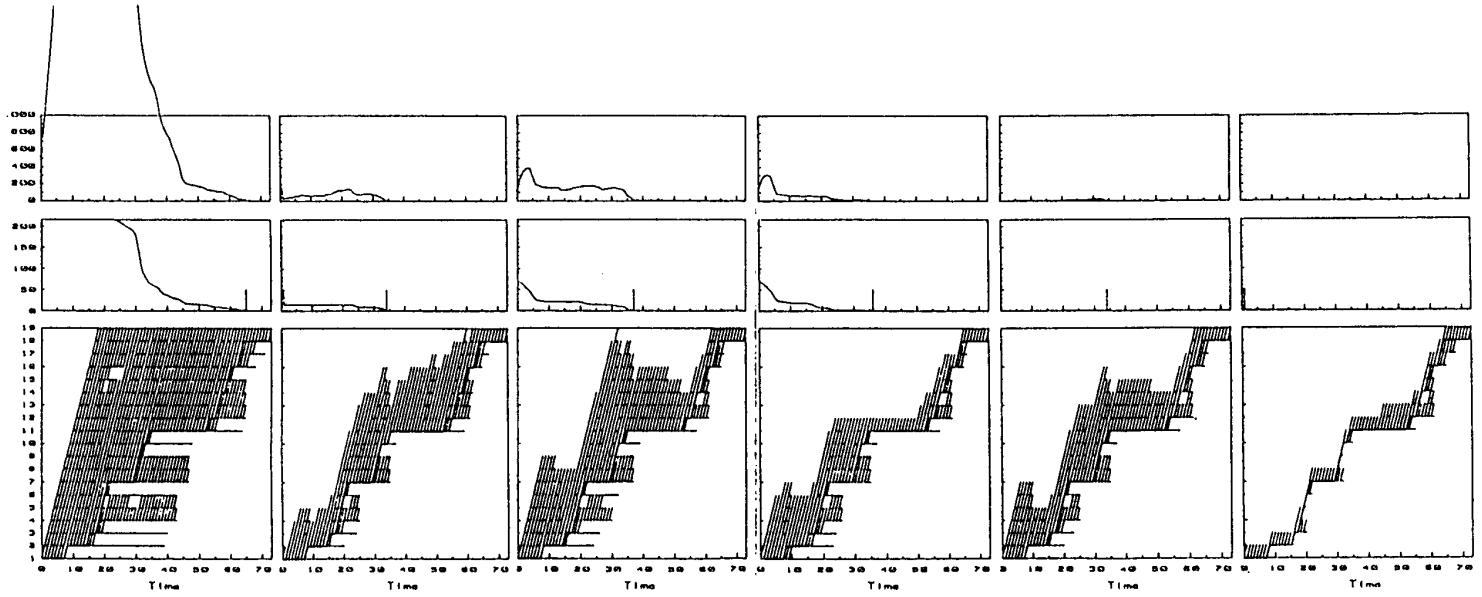


図 8　処理例（一方向サーチ）

Fig. 8　Examples of experimental results (unidirectional search).

（注：　フローチャート図，ブロック図，構成図，写真，データ表，グラフ等　研究内容の補足説明に御使用下さい）

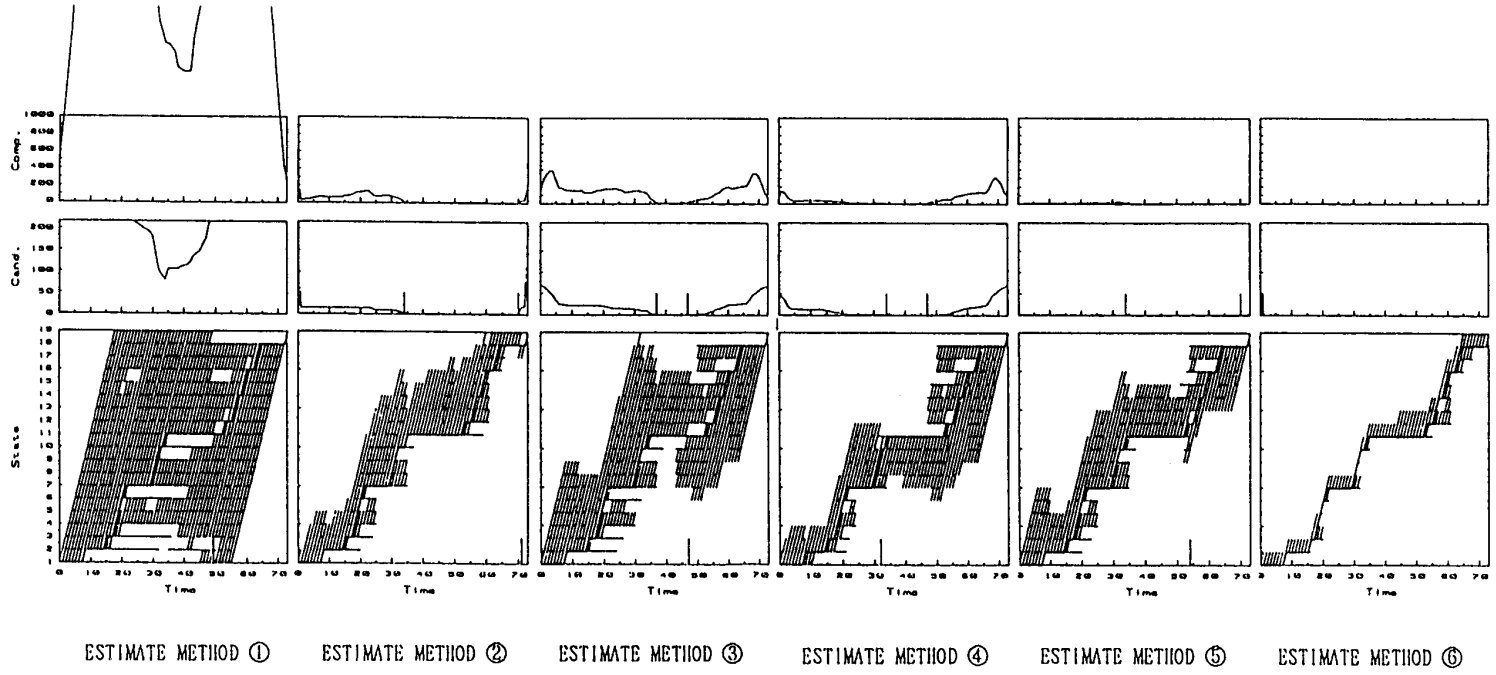


図 9　処理例（双方向サーチ）

Fig. 9　Examples of experimental results (bidirectional search).

（注：　フローチャート図，ブロック図，構成図，写真，データ表，グラフ等　研究内容の補足説明に御使用下さい）

様式－１０

表１　推定スコア設定法、及び、探索アルゴリズムの種々の組み合わせによる216単語音声認識における計算量（％）と認識誤り数

	最適性	時間順序	推定スコア設定法		経路スコア	経路展開の計算量 一方向サーチ 前向き	一方向サーチ 後向き	双方向サーチ	双方向サーチ 逆方向利用	推定スコア設定の計算量
①	有	無視	推定スコア０		——	27.182 (9)	26.643 (9)	52.710 (9)	——	0
②	—		平均対数尤度		——	0.360(52)	0.726(96)	0.641(37)	0.634(36)	
③	有		最大経路スコア	1～Jn	(a , b)	0.834 (9)	0.934 (9)	1.642 (9)	0.729 (9)	2.54
	有				(1/2, b)	0.516 (8)	0.634 (9)	1.020 (9)	0.491 (8)	
	有				(－ , b)	1.613 (9)	1.894 (9)	3.307 (9)	1.356 (9)	
④	有			j～Jn	(a , b)	0.468 (9)	0.521 (9)	0.923 (9)	0.519 (7)	——
	有				(1/2, b)	0.313 (8)	0.376 (9)	0.584 (9)	0.348(10)	
	有				(－ , b)	0.899 (9)	1.008 (9)	1.809 (9)	0.928 (8)	
⑤	—	考慮	単純音素HMM	音素セグメント	(a , b)	0.172 (9)	0.175(10)	0.260 (9)	0.159 (8)	33.3
	—				(1/2, b)	0.102 (8)	0.117(12)	0.150(11)	0.106(13)	
	—				(－ , b)	0.316 (9)	0.354(10)	0.531 (9)	0.282 (8)	
⑥	—			Viterbiスコア		0.048 (7)	0.045(14)	0.050(12)	0.050(12)	

() 内は認識誤り数

（注：　フローチャート図，ブロック図，構成図，写真，データ表，グラフ等　研究内容の補足説明に御使用下さい）