特　集　「音声の知的処理」

音声認識における時系列パターン照合アルゴリズムの展開

Time-Sequential-Pattern Matching Algorithm on Automatic Speech Recognition

中川　聖一*
Seiichi Nakagawa

* 豊橋技術科学大学情報工学系
Dept. of Information and Computer Sciences, Toyohashi University of Technology.

1988年4月8日　受理

Keywords: speech recognition, DP matching, HMM, time warping, dynamic programming, one pass DP, one stage DP, forward-backward algorithm, Baum-Welch estimation algorithm.

1．音声認識におけるパターン照合の問題

単語や文は，音韻とか音節とかの言語の基本構成単位から構成されているので，音韻認識や音節認識が確実にできれば単語音声認識や連続音声認識も確実にできる．しかし，実際には，話者や前後のコンテキストによって音声パターンが変動するので（個人差，調音結合の現象），音韻や音節の認識は容易でない．また，たとえ発声者と発声内容が同一であっても，発声ごとに音声パターンは異なる．特に時間軸上でのパターンの変動がみられる．

時間軸上での変動に対しては，時系列パターンの時間軸上での非線形伸縮を許すパターン照合アルゴリズム，いわゆるDPマッチング法が開発され実用化されている．個人差の問題に対して最も安易な対処方法は，話者を限定することであり，この場合は特定話者方式と呼ばれている．一方，話者を限定しない場合は個人差の正規化や適応化手法が試みられてきた．しかし，満足できる正規化手法は開発されていない．通常は，統計的手法で個人差の問題に対処している．調音結合の問題に対しては，調音結合の影響を受けたままの単語音声パターンを一つの観測パターンとみなして，単語音声どうしで直接照合する方法が成功を収めてきた．しかし，この考え方を大語彙の単語音声や連続音声認識に適用することはできない．なぜなら，単語パターンの作成が容易でないことや単語間にわたる調音結合には対処できないからである．

本稿では，音声認識の基本的アルゴリズムとしてほぼ確立されているDPマッチング法とHMM（Hidden Markov Model，隠れマルコフモデル）について解説する．DPマッチング法は，二つの異なる単語音声の時系列パターンを非線形に伸縮しながら照合する手法である．一方，HMMは，個人差や調音結合（単語単位の認識の場合は，単語内の調音結合は問題ないが）によるパターンの変動を確率モデルとしてとらえる方法である．もちろん，時間長の変動も確率モデルで吸収できる．これらの点から原理的にはDPマッチン

表1　DPマッチングによる単語音声認識結果

語　彙	話者モード	認識率〔%〕
10数字	不特定	99.2
39数字，アルファベット	特　定	79.5
	不特定	79.0
54コマンド語	不特定	96.5
129航空座席，予約単語	特　定	88.0
	不特定	91.0
1109基本単語	特　定	79.2

表2　HMMによる不特定話者の数字音声認識結果

方　法	モ　デ　ル	認識率〔%〕
HMM	離散分布	97.1
	連続分布（1モデル／数字）	99.1
	連続分布（2モデル／数字）	99.4
	5混合対角分布（2モデル／数字）	99.3
DP	12テンプレート／数字	99.4

グ法よりもHMM法のほうが精度よく標準パターンを構成できる．しかし，実際には確率モデルのパラメータ数が多くなるとパラメータの精度よい推定が困難になるから，結局，単語音声を粗っぽいモデル（たとえば，単語音声を時間的に5～10状態）で表現せざるを得ないので，マルチテンプレートによるDPマッチング法と比べて認識精度は現在のところ大差はない．

表1，2はAT&Tベル電話研究所で行われてきた一連の音声認識研究の結果をまとめたものである[1][2]．

2．DPマッチングによる音声認識アルゴリズム[3][34]

2・1　孤立単語音声の認識[4][6]

(a)　単語音声認識の定式化

一つの単語は発声されるたびに継続時間が変わり，その時間構造も線形ではない．発声速度が速くなるに従って，母音の定常部分は短くなるのに対して，子音部やわたりの部分は比較的固有の長さを保っている．このような性質をもった二つの不等長なパターンを比較する場合，時間長の違いによって生ずる本質的でない差を極力排除した距離尺度が望ましい．このような考え方から時間軸に非線形な変換（これを時間変換関数：time warping functionと呼ぶ）を施すことにより，パターン長の差の影響をできるかぎり吸収しようとする方式がDPマッチングと呼ばれているものである．

二つの音声パターン$\boldsymbol{A}$と$\boldsymbol{B}$が，時間標本化された特徴ベクトルの時系列として表現されている場合に問題を定式化しよう．

$$\boldsymbol{A}=\boldsymbol{a}_1\boldsymbol{a}_2\cdots\boldsymbol{a}_i\cdots\boldsymbol{a}_I$$

$$\boldsymbol{B}=\boldsymbol{b}_1\boldsymbol{b}_2\cdots\boldsymbol{b}_j\cdots\boldsymbol{b}_J$$

とする．ここで，$\boldsymbol{a}_i$は$\boldsymbol{A}$の第iフレームを示すものとする（$\boldsymbol{a}_i=\boldsymbol{A}(i)$）．$I, J$はそれぞれ$\boldsymbol{A}$および$\boldsymbol{B}$の時間長でフレーム総数である．

ここで，$\boldsymbol{A}, \boldsymbol{B}$両パターンの時間の対応すなわち時間変換関数は，$i$-$j$平面上の格子点$c=(i, j)$の系列$\boldsymbol{F}$で表現することができる．

$$\boldsymbol{F}=c(1)\,c(2)\cdots c(k)\cdots c(K)$$

ただし，系列点$c(k)=(i(k), j(k))$はi軸では$i(k)$，j軸では$j(k)$の格子点である．二つの特徴ベクトル$\boldsymbol{a}_i$と$\boldsymbol{b}_j$との間では，いくつもの距離を定義できるが，その定義の仕方は一応問題にせず，便宜的にこの二者間の距離を格子点を仲介にして$\boldsymbol{d}(c)=\boldsymbol{d}(i, j)$で表すと，点列$\boldsymbol{F}$に沿ってとった距離の荷重平均（時間正規化距離）において点列$\boldsymbol{F}$を変化させたときの最小値で，二つのパターン$\boldsymbol{A}$と$\boldsymbol{B}$との距離$\boldsymbol{D}(\boldsymbol{A}, \boldsymbol{B})$を定義する．

$$\boldsymbol{D}(\boldsymbol{A}, \boldsymbol{B})=\min_{\boldsymbol{F}}\frac{\sum_{k=1}^{K}\boldsymbol{d}(c(k))\cdot w(k)}{\sum_{k=1}^{K}w(k)} \quad (1)$$

式(1)の分母$\sum_{k=1}^{K}w(k)$は点列$\boldsymbol{F}$を構成する点の数による距離の値の差を補正するための正規化係数であり，その値は$\boldsymbol{F}$に依存する．このとき式(1)の最小化問題は複雑になるので，式(1)の分母を$\boldsymbol{F}$に依存しない定数とすることによりこれを簡単化する．

$$N=\sum_{k=1}^{K}w(k)$$

とおくと，(1)式は，

$$\boldsymbol{D}(\boldsymbol{A}, \boldsymbol{B})=\frac{1}{N}\min\sum_{k=1}^{K}\boldsymbol{d}(c(k))\cdot w(k) \quad (2)$$

となり，最小化する目的関数が加法的になるので，次の(b)で述べるように，その最小化は動的計画法の手法を用いて効率よく解くことができる．

Nを定数化する方法は種々ありうるが，次の二つがよく試みられている．

① 対称形：$w(k)=(i(k)-i(k-1))+(j(k)-j(k-1))$，
このとき，$N=I+J$（図1 (a)）

② 非対称形：$w(k)=i(k)-i(k-1)=1$，
$0\leqq j(k)-j(k-1)\leqq 2$，
このとき，$N=I$（図1 (c)）

(b)　動的計画法による最小化問題の解法

ここでは，式(2)の最小化問題を動的計画法（Dynamic Programming）を用いて効率よく解く方法を

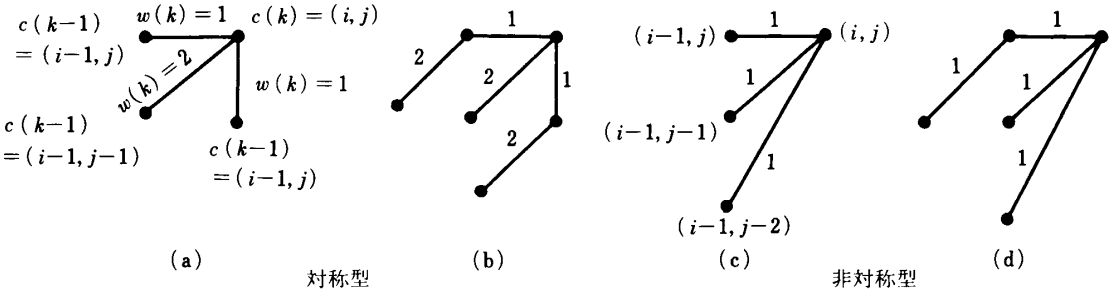


図1　各種DPパスと荷重係数の例

説明する．まず部分点列 $(c(1)\,c(2)\cdots c(k)$ $(c(k)=i(k), j(k))$ であるが，便宜上 $c(k)=(i,j)$ と略記する）に対する次の部分和 $\boldsymbol{g}(c(k))$ を考える．

$$\boldsymbol{g}(c(k))=\boldsymbol{g}(i,j)=\min_{c(1)\cdots c(k)}\left(\sum_{l=1}^{k}\boldsymbol{d}(c(l))\cdot w(l)\right) \tag{3}$$

上式を変形すると，

$$\begin{aligned}\boldsymbol{g}(c(k)) &= \min_{c(l)\cdots c(k)}\left(\sum_{l=1}^{k-1}\boldsymbol{d}(c(l))\cdot w(l)+\boldsymbol{d}(c(k)\cdot w(k))\right)\\ &= \min_{c(k-1)}\left(\boldsymbol{g}(c(k-1))+\boldsymbol{d}(c(k))\cdot w(k)\right)\end{aligned} \tag{4}$$

$c(k)=(i,j)$ であるから，対称形の $w(k)$ を用いることにすると，点列に対する単調連続性の条件から $c(k-1)$ は $(i,j-1)$, $(i-1,j)$ および $(i-1,j-1)$ のいずれかである．故に，式(4)は次のように書ける．

$$\boldsymbol{g}(i,j)=\min\begin{Bmatrix}\boldsymbol{g}(i,j-1)+\boldsymbol{d}(i,j)\\ \boldsymbol{g}(i-1,j-1)+2\boldsymbol{d}(i,j)\\ \boldsymbol{g}(i-1,j)+\boldsymbol{d}(i,j)\end{Bmatrix} \tag{5}$$

したがって，式(2)の最小化問題は次の3段階の計算によって解くことができる．

① 初期設定：$\boldsymbol{g}(1,1)=2\boldsymbol{d}(1,1)=2\boldsymbol{d}(a_1,b_1)$ および $i=1$

② $j_1=\max(1,i-r)$, $j_2=\min(i+r,J)$ として j を j_1 から j_2 まで一つずつ増やしながら式(5)によって $\boldsymbol{g}(i,j)$ を計算する（$2r$ は整合窓長）．

③ i が I に等しくなければ i を1だけ増して②へいく．等しければ，$\boldsymbol{D}(\boldsymbol{A},\boldsymbol{B})=\boldsymbol{g}(I,J)/N$ とする．式(5)を用いる場合は $N=I+J$ である．整合窓は極端な非線形変換を防ぐためのものである．

上記②の計算がフレーム周期（通常10～20 ms）以内で処理できれば，$\boldsymbol{A}$ が入力パターンの場合，実時間で $\boldsymbol{D}(\boldsymbol{A},\boldsymbol{B})$ が計算できることになる．

DPマッチング法を単語音声の認識に適用するには次のようにする．認識の対象となっている単語を n $(n=1\sim N)$，その標準パターンを $\boldsymbol{B}(n)$ で表す．入力 $\boldsymbol{A}$ と各標準パターン $\boldsymbol{B}(n)$ との距離 $\boldsymbol{D}^n=\boldsymbol{D}(\boldsymbol{A},\boldsymbol{B}(n))$ を上記の方法で計算し，$\boldsymbol{D}^{\hat{n}}=\min_n(\boldsymbol{D}^n)$ を与える単語 $\hat{n}$ を $\boldsymbol{A}$ に対する認識結果とする．

このほか，非線形伸縮比を局所的にいつも1/2～2に押さえる方法として傾斜制限つきDPパスを用いる方法がある．例えば，図1（d）の非対称形パスの場合は，

$$\boldsymbol{g}(i,j)=\min\begin{Bmatrix}\boldsymbol{g}(i-2,j-1)+\boldsymbol{d}(i-1,j)+\boldsymbol{d}(i,j)\\ \boldsymbol{g}(i-1,j-1)+\boldsymbol{d}(i,j)\\ \boldsymbol{g}(i-1,j-2)+\boldsymbol{d}(i,j)\end{Bmatrix} \tag{6}$$

$$\boldsymbol{D}(\boldsymbol{A},\boldsymbol{B})=\boldsymbol{g}(I,J)/I$$

となる．

2・2 連続単語音声の認識

(a) 連続単語音声認識の定式化[7][8]

次に動的計画法に基づいたDPマッチング法の連続単語への拡張について述べる．

認識の対象となっている単語を n $(n=1\sim N)$ で，その標準パターンを $\boldsymbol{B}(n)=\boldsymbol{b}_1^n\,\boldsymbol{b}_2^n\cdots\boldsymbol{b}_{J^n}^n$（$J^n$ は標準パターン $\boldsymbol{B}(n)$ の長さ）で示す．また入力パターンを $\boldsymbol{A}=\boldsymbol{a}_1\,\boldsymbol{a}_2\cdots\boldsymbol{a}_I$ とする．x 個の単語からなる単語列 $\boldsymbol{n}=n_1\,n_2\cdots n_x$ を発声したときのパターンを，標準パターン $\boldsymbol{B}(n_1)$, $\boldsymbol{B}(n_2)$, …, $\boldsymbol{B}(n_x)$ を接続して得られるパターン，$\boldsymbol{B}=\boldsymbol{B}(n_1)\oplus\boldsymbol{B}(n_2)\oplus\cdots\oplus\boldsymbol{B}(n_x)$（⊕でパターンの接続を示す）によって近似する．

この $\boldsymbol{B}$ を連続単語の標準パターンと考えて，前節で述べたDPマッチング法によって入力パターン $\boldsymbol{A}$ との間で時間正規化マッチングを実行すると，両者の間の距離が求められると同時に，入力パターン $\boldsymbol{A}$ の単語単位のセグメンテーションも，最適な照合パスをバックトレースすることによって得られる．単語列 $\boldsymbol{n}$（したがって個数 x も）を変化させたとき，両者の間の距離を最小にするような単語列を認識結果とするわけである．このときの距離の最小値を $\boldsymbol{D}'(I)$（I は $\boldsymbol{A}$ の長さ）と書くと，

$$\boldsymbol{D}'(I)=\min_{n,x}(\boldsymbol{D}(\boldsymbol{A},\boldsymbol{B}(n_1)\oplus\boldsymbol{B}(n_2)\oplus\cdots\oplus\boldsymbol{B}(n_x))) \tag{7}$$

を満たす単語数 x および単語列 $\boldsymbol{n}=n_1\,n_2\cdots n_x$ を求めようとするのである．

式(7)を総当たり法で解くと N^x（X：入力パターンとマッチングをとる単語列中の最大の単語数）のオーダのDPマッチングの回数を必要とする．そこで，この最小化問題を2段階に分割して計算量の軽減を図ることを考える．

入力パターン $\boldsymbol{A}$ の部分パターン $\boldsymbol{a}_{l+1}\,\boldsymbol{a}_{l+2}\cdots\boldsymbol{a}_m$ を $\boldsymbol{A}(l,m)$，$\boldsymbol{D}$ および $\boldsymbol{D}'$ に対する正規化前の累積距離を $\boldsymbol{T}$ および $\boldsymbol{T}'$ と書く．非対称形のDPパス（$w(k)=i(k)-i(k-1)=1$）の性質を用いれば，途中の導出過程は略するが，次の漸化的な表現を得る．

$$\boldsymbol{T}'(I)=\min_{1<l<I}\left(\boldsymbol{T}'(l)+\min_n\{\boldsymbol{T}(\boldsymbol{A}(l,I),\boldsymbol{B}(n))\}\right) \tag{8}$$

上式の誘導は $\boldsymbol{A}$ の任意の部分パターン $\boldsymbol{A}(0,i)$ に対して成立するから，上式の I（入力パターン全体の長さ）は任意の i（$<I$）に置き換えることができて，

$$T'(i)=\min_{1<l<I}\left(T'(l)+\min_{n}\{T(A(l,i),B(n))\}\right)$$

$$(i=1,2,\cdots,I) \tag{9}$$

ここで，$T'(0)=0$とする．上式を満たすl, nをそれぞれ$C(i)$, $N(i)$と定義する．すなわち，$N(i)$は入力音声Aの第iフレームで終る全区間でDPマッチングがとれた最適単語列の最後尾の単語名，$C(i)$は$N(i)$の入力音声パターンとのマッチング開始フレームの値($l+1$)に対し，一つ前のlを指し単語境界フレームに該当する．式(9)を解釈すれば，入力音声パターンがiフレームで終端すると仮定した場合，その最適な単語列(単語数x)は入力パターンの第lフレームまでの最適な単語列(単語数$x-1$)と，入力パターンのそれに続く$l+1\sim i$フレームと照合がとれる最適単語を結合させたものであり，式(9)の$T'(i)$はlを可変とした際の最適値を示している．連続単語としての全体の認識結果は，$N(I)$, $N(C(I))$, $N(C(C(I)))$, …として，最適単語列の最後尾単語名から逆順に求めることができる．

(b) 2段DPマッチング法[8]

式(9)は式(7)の最小化問題がやはり動的計画法によって解けることを示している．式(9)の$T(A,B)$の計算は前節で述べたDPマッチング法を使用するので，結局，式(9)は動的計画法を2段階に分けて使用していることになる．この意味から，式(9)を直接解く方法は2段DP法と呼ばれる．

式(9)においてlの範囲は原理的には$1<l<i$であるが，前項で述べた整合窓の条件により，$i-r-\max_{n}(J^n)\leq l\leq i+r-\min_{n}(J^n)$($r$は窓の大きさ，$J^n$は単語$n$の標準パターンの長さ)なる範囲をとればよいことがわかる．さらに，1回のTの計算によって多くの部分パターンとの距離計算が並列的に行われる．すなわち，単語nの標準パターンと入力パターンの第$l+1$フレームからの入力の部分パターンとの累積距離を求める際に，並列的に$T(A(l,i),B(n))$($l+J^n-r\leq i\leq l+J^n+r$)を一度に求めることができるという事実に注意すると，2段DPマッチング法は式(9)から予想されるよりかなり少ない計算量で連続単語の認識を可能にしていることがわかる．

(c) $O(n)$ DPマッチング法[9], One Pass DP法[10], One Stage DP法[11]

式(9)は次のように変形できる．

$$T'(i)=\min_{n}\left(\min_{1<l<i}\{T'(l)+T(A(l,i),B(n))\}\right) \tag{10}$$

$T(A(l,i),B(n))$は，孤立単語の認識の項で述べたように，初期設定を，$g^n(l+1,1)=d^n(l+1,1)$, $g^n(l'+1,1)=\infty$ ($l'>l$)として以下の漸化式を求めていくことによって得られる(ただし，$w(k)$は非対称形)．

$$g^n(h,k)=\min\begin{Bmatrix}g^n(h-1,k)\\ g^n(h-1,k-1)\\ g^n(h-1,k-2)\end{Bmatrix}+d^n(h,k) \tag{11}$$

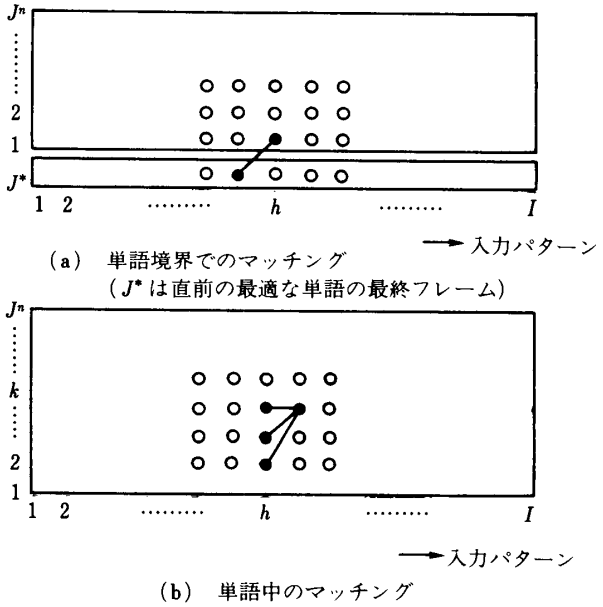
$$T(A(l,i),B(n))=g^n(i,J^n)$$

しかし，この場合，初期設定を$g^n(l+1,1)=T'(l)+d^n(l+1,1)$, $0<l<i$として，$T'(l)$をgの中へ組み込めば式(10)は，

$$T'(i)=\min_{n}(g^n(i,J^n)) \tag{12}$$

となり1段の動的計画法を表現することになる．しかもiに関しては2段DP法の項で述べたように，一度のTの計算に並行して種々のiについて並列的に計算していくことができる．こう考えることによって，2段DPマッチングより効率のよいアルゴリズムが誘導できる．図2はこのマッチング法のようすを示している．

この方法の計算量は，入力パターン長I，平均標準パターン長J，標準パターン数Nに比例することとDP(Tの計算)の回数が単語数に比例することから$O(n)$ DP法[9]と呼ばれている．この方法は，迫江らによるクロック伝搬法[12]の特殊な場合とも解釈できる．また，Bridleによっても独立に考案され，Neyによって明確なアルゴリズムが示され，それぞれOne Pass DP法[10]，One Stage DP法[11]とも呼ばれている．



(a) 単語境界でのマッチング
(J^*は直前の最適な単語の最終フレーム)

(b) 単語中のマッチング

図2 $O(n)$ DPマッチング法，One Pass DP法におけるマッチング処理

つまり，孤立単語の DP マッチングによる認識の計算量と同じオーダで連続単語の認識が可能となる．

このほかにも種々のアルゴリズムが開発されている．Level Building 法は，第 n レベルの処理として，入力パターンと（$n-1$）単語列との最適な照合結果を初期値とし各単語の照合を連結していき，n 単語列としての最適な照合結果を得るもので，2 段 DP 法よりも計算量は少ない[13]．また，拡張連続 DP 法は，式(9)の第 2 項をワードスポッティングアルゴリズムで能率よく求める方法で，近似解（ただし，厳密解と認識精度は同等）となるが計算量はさらに少ない特徴がある[14]．

2・3 オートマトン制御 DP マッチング[15]

自然言語文の音声で代表される連続音声は，協力的で明確に発声された場合は，上述の連続単語音声の場合と同様にパターンマッチング法が適用可能である．ただ単語間の接続に関して構文的・意味的制約がある点が異なる．

連続単語音声の場合は，単語間の接続に関しては何も制約がなくあらゆる単語が接続可能であった．連続音声の場合は，接続できる単語と接続できない単語を区別して処理しなければならない．この関係は状態遷移図で表すことができ，一般に有限状態オートマトンで表現できる．

図 3 は簡単な英文を生成する有限状態オートマトンである．たとえば，状態 ART では，boy, girl, little, big の 4 単語が接続可能である．二重丸印の状態は最終状態を表し，この状態に達すると，入力は終端しうることを示す．

有限状態オートマトン α は次のように定義できる．

$\alpha = <K, \Sigma, \delta, q_0, F>$

K：状態 q_i の有限集合，$\{q_i\}$

Σ：入力単語 n の有限集合 $\{n\}$

δ：状態遷移関数，$K\times\Sigma\rightarrow K$，$\{\delta(q_i, n) = q_j\}$

q_0：初期状態，$q_0 \in K$

F：最終状態の集合，$F \subset K$

単語の接続条件がオートマトン α で表現される場合，単語列 $n_1, n_2, \cdots$ がこのオートマトン α に受理されるという条件のもとに，式(9)と類似した以下の式を最小化する最適な単語列 $\boldsymbol{n}$ を求めるとよい．

$$T'(q_k, i) = \min_{l, q_j, n} (T'(q_j, l) + T(A(l, i), B(n))) \quad (13)$$

ただし，$q_k = \delta(q_j, n)$．上式を満たす $\hat{n}, \hat{l}, \hat{q}_j$ によって $N(q_k, i)$, $C(q_k, i)$ が定義され，新たに状態遷移の結果保存のために $Q(q_k, i)$ を定義算出しておく必要がある（$N(q_k, i) = \hat{n}$, $C(q_k, i) = \hat{l}$, $Q(q_k, i) = \hat{q}_j$）．

認識結果は，$q' = \arg\min(T'(q, I))$ として $N(q', I)$, $N(Q(q', I), C(q', I))$, …のようにしてオートマトンで受理される単語列のうち最適単語列が最後尾単語名から逆順に求められる．

この定式化の具体的なアルゴリズムは前述の 2 段 DP 法や $O(n)$ DP 法に容易に組み込むことができる（オートマトン制御 2 段 DP 法[15]，オートマトン制御クロック同期伝播形 DP 法[12] などと呼ばれている）．

3. Hidden Markov モデルによる音声認識アルゴリズム[16]-[21]

3・1 基本モデルと認識アルゴリズム

(a) 基本モデル

DP マッチングは，時系列パターンの時間的構造の変動に対しては強力であるが，話者の個人差などに起因するスペクトルそのものの変動に対しては弱い．パターンマッチングに基づく音声認識の立場からは，これらの変動成分を統計量によって表現する方法がある．この考え方は，時系列パターンの時間的構造の変動に対しても適用できる．DP マッチングにおいては，単語に対して標準的な時系列パターンを標準パターンとしているが，HMM(Hidden Markov Model) による方法は，各単語を標準的な確率状態遷移機械で表現する[17]．このモデルのパラメータは，スペクトルや時間的な変動を最もよく吸収するように学習サンプルによって求められる．

図 4 は，Bakis モデルと呼ばれ，最もよく用いられている HMM の例を示している．もちろん，状態数は適当に決めておく必要がある．状態数を多くすれば，きめ細かい単語パターンの表現ができるが，モデルの

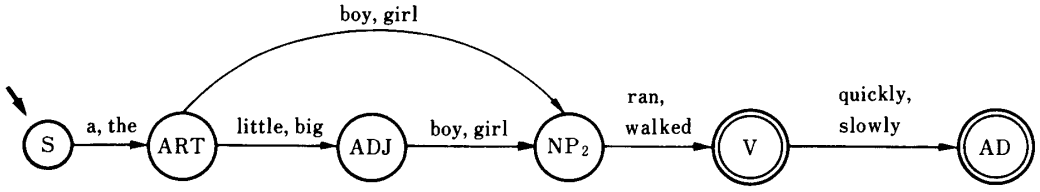


図 3 簡単な英文を生成する有限状態オートマトン

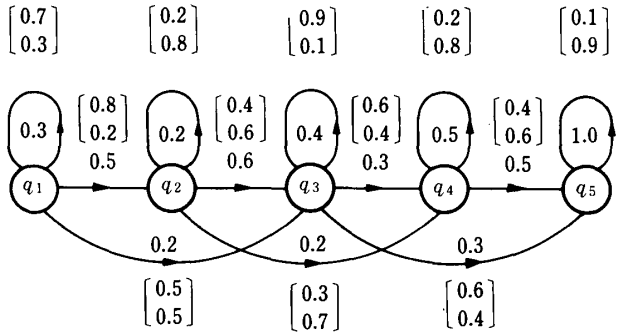


図 4 Bakis モデルの例
([] 内の数値は，シンボル a, b の出力確率)

パラメータ数が多くなりパラメータ推定の精度が悪くなる．図4の状態遷移のアーク上の数値は状態遷移確率を表している．これは，状態数を S とすると $S\times S$ の行列によって表現できる．状態 q_i から状態 q_j への遷移確率を a_{ij} と書くと図4の例では，

$$A=\{a_{ij}\}=\begin{vmatrix} 0.3 & 0.5 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.6 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{vmatrix}$$

となる．通常，音声パターンには，時間的な非可逆性の性質があるから，$i>j$ なら $a_{ij}=0$ である．

観測パターンが，有限個 (K個) のシンボルの一つとして表現できる場合は (通常ベクトル量子化などの手法を用いる)，離散分布モデルと呼ばれる．状態 q_i から状態 q_j の遷移でシンボル k のパターンが観測 (出力) される確率を $b_{ij}(k)$ と表す．この $\{b_{ij}(k)\}$ をシンボル出力確率行列と呼ぶ．

簡単な例として，出力シンボルを $\{a, b\}$ の二つに限られた場合の例を図4の状態遷移のアーク上の〔 〕内に示す．この出力確率が，遷移元の状態によってのみ決定され状態遷移に独立なモデルも考えられる．すなわち，$b_{ij}(k)=b_{ik}(k)$．この場合は，出力確率行列は $S\times K$ の行列で表現できる．

各状態 q_i の初期確率を π_i で表す．図4の例では，$\pi_1=1$，$\pi_i=0$ ($i>1$) である．また，最終状態の集合を F と記す．図4の例では $F=\{q_5\}$ である．

(b) 認識アルゴリズム

今，$\boldsymbol{y}=y_1y_2\cdots y_T$ を出力シンボルの観測系列としよう．たとえば，10ms ごとの音声特徴パラメータ (スペクトルとかケプストラム) の時系列パターンが考えられる．このとき，この HMM の $\boldsymbol{M}$ によって $\boldsymbol{y}$ が生起する確率 $P(\boldsymbol{y}\mid\boldsymbol{M})$ を求める．これを各単語に対応する HMM に対して求め，最大確率を与えるモデルに対応する単語を認識結果とする．

$\boldsymbol{q}=q_{i0}q_{i1}\cdots q_{iT}$ を状態遷移系列 (ただし，$q_{iT}\in F$) とすれば，

$$P(\boldsymbol{y}\mid\boldsymbol{M})=\sum_{i_0,i_1,\cdots,i_T}P(\boldsymbol{y}\mid\boldsymbol{q},\boldsymbol{M})\cdot P(\boldsymbol{q}\mid\boldsymbol{M}) \tag{14}$$

$$P(\boldsymbol{q}\mid\boldsymbol{M})=\prod_t P(q_{it}\mid\boldsymbol{q}_{i0}^{t-1},\boldsymbol{M})$$
$$=\prod_t P(q_{it}\mid\boldsymbol{q}_{it-1},\boldsymbol{M})$$
(マルコフモデルの性質) (15)

$$P(\boldsymbol{y}\mid\boldsymbol{q},\boldsymbol{M})=\prod_t P(y_t\mid q,\boldsymbol{y}_1^{t-1},\boldsymbol{M})$$
$$=\prod_t P(y_t\mid q_{it-1},q_{it},\boldsymbol{M}) \tag{16}$$

$$P(\boldsymbol{y}\mid\boldsymbol{M})=\sum_{i_0,i_1,\cdots,i_T}\prod_t P(q_{it}\mid q_{it-1},\boldsymbol{M})\cdot P(y_t\mid q_{it-1},q_{it},\boldsymbol{M})$$
$$=\sum_{i_0,i_1,\cdots,i_T}\pi_{i0}a_{i0i1}b_{i0i1}(y_1)\,a_{i1i2}b_{i1i2}(y_2)\cdots a_{iT-1iT}b_{iT-1iT}(y_T)$$

ここで，$\boldsymbol{q}_{i0}^{it-1}=q_{i0},q_{i1},\cdots,q_{it-1}$,

$$\boldsymbol{y}_1^t=y_1y_2\cdots y_t \tag{17}$$

となる．なお，最後の式中のそれぞれの確率は，モデル $\boldsymbol{M}$ での条件確率である．簡単な例として図4でシンボル系列 aba が出力される確率を求めてみよう．この場合は，考えるべき状態遷移系列は，図4からもわかるように，左上隅から右下隅に至る経路に対応するもので次の6通りである．$q_1\to q_1\to q_3\to q_5$，$q_1\to q_2\to q_3\to q_5$，$q_1\to q_2\to q_4\to q_5$，$q_1\to q_3\to q_3\to q_5$，$q_1\to q_3\to q_4\to q_5$，$q_1\to q_3\to q_5\to q_5$．それぞれの確率は，

$P_1=0.3\times0.7\times0.2\times0.5\times0.3\times0.6=0.00378$
$P_2=0.5\times0.8\times0.6\times0.6\times0.3\times0.6=0.02592$
$P_3=0.5\times0.8\times0.2\times0.7\times0.5\times0.4=0.0112$
$P_4=0.2\times0.5\times0.4\times0.1\times0.3\times0.6=0.00072$
$P_5=0.2\times0.5\times0.3\times0.4\times0.5\times0.4=0.0024$
$P_6=0.2\times0.5\times0.3\times0.4\times1.0\times0.1=0.0012$

となる．故に，$P(aba\mid M)=P_1+P_2+P_3+P_4+P_5+P_6=0.04522$ となる．

一般に $P(\boldsymbol{y}\mid\boldsymbol{M})$ の値は，以下に述べる Forward アルゴリズムで求められる．

Forward 変数 $\alpha(i,t)$ を観測シンボル $\boldsymbol{y}_1^t$ を出力して状態 q_i にある確率とすれば，

$$\alpha(i,0)=\pi_i \text{ for } i=1,2,\cdots,S$$
$$\alpha(i,t)=\sum_j \alpha(j,t-1)\cdot a_{ji}\cdot b_{ji}(y_t) \tag{18}$$
$$\text{for } t=1,2,\cdots,T;\ i=1,2,\cdots,S$$

$$P(\boldsymbol{y}\mid\boldsymbol{M})=\sum_{i,q_i\in F}\alpha(i,T) \tag{19}$$

一方，$P(\boldsymbol{y}\mid\boldsymbol{M})$ を厳密に求めないで，モデル $\boldsymbol{M}$ が観測シンボル系列 $\boldsymbol{y}$ を出力するときの最も可能性の高い状態系列 $\boldsymbol{q}=q_{i0}q_{i1}\cdots q_{iT}$ 上での出力確率を使うことも考えられる．図4の例では，$q_1\to q_2\to q_3\to q_5$ (P_2) がこれに相当する．この対数尤度 L は，次の動的計画法を適用することによって求められる．

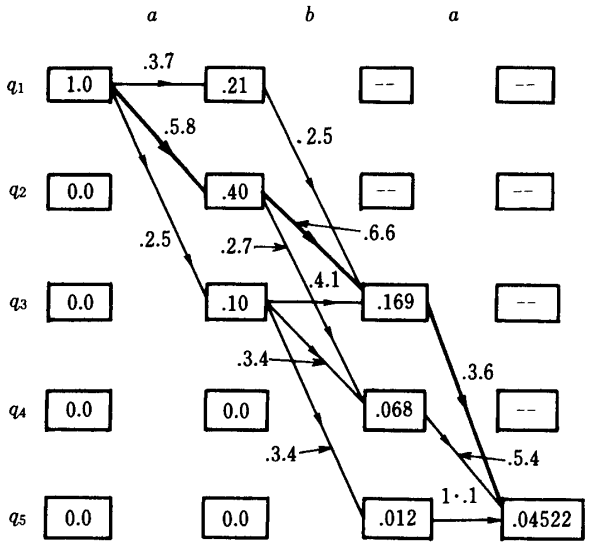


図 5 図4に対する $\alpha(i,t)$ の計算例
(--は $\alpha(5,3)$ を求めるのに無関係なので省略した)

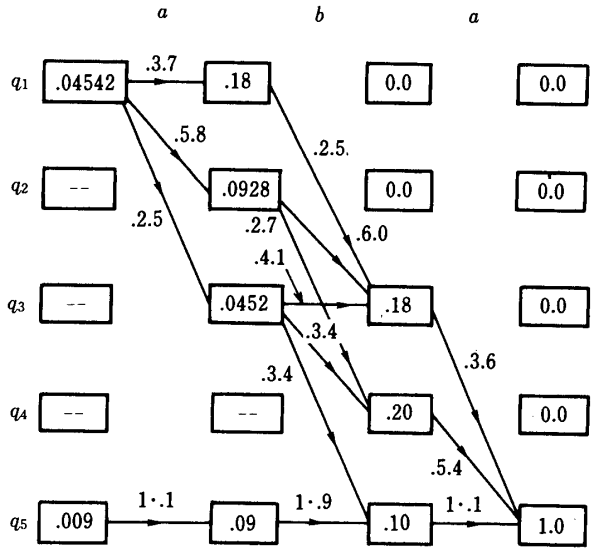


図 6 図4に対する $\beta(i,t)$ の計算例
(--は $\beta(1,1)$ を求めるのに無関係なので省略した)

$$f'(i,0)=\log\pi_i \quad \text{for } i=1,\ 2,\ \cdots,\ S$$

$$f'(i,t)=\max_j\{f'(j,t-1)+\log a_{ji}\cdot b_{ji}(y_t)\} \quad \text{for } t=1,\ 2,\ \cdots,\ T;\ i=1,\ 2,\ \cdots,\ S \tag{20}$$

$$L=\max_{i:q_i\in F} f'(i,T) \tag{21}$$

これは Viterbi アルゴリズムと呼ばれている．この方法は，厳密に確率を求める方法と比較して，計算量が少ないにもかかわらず認識精度は同等であることが実験的に確かめられている[18]．

図5に図4に対する $\alpha(i,t)$ の計算例を示す．図中太線は最適状態遷移系列を示している．

3・2 Baum-Welch のパラメータ推定法 (Forward-Backward アルゴリズム)[16]-[21]

新たに次のパラメータを導入する．Backward 変数 $\beta(i,t)$ を，時刻 t に状態 q_i にいて以後観測シンボル ${y_{t+1}}^T$ を出力する確率，$\gamma(i,j,t)$ をモデル M が ${y_1}^T$ を出力する場合において，状態 q_i からシンボル y_t を出力し，時刻 t で状態 q_j へ遷移する確率とする．

このとき，

$$\beta(i,T)=\begin{cases}1.0 & q_i\in F\\ 0 & q_i\notin F\end{cases}$$

$$\beta(i,t)=\sum_j a_{ij}\cdot b_{ij}(y_t)\cdot\beta(j,t+1) \quad \text{for } t=T,\ T-1,\cdots,1;\ i=1,\ 2,\cdots,S \tag{22}$$

$$\gamma(i,j,t)=\alpha(i,\ t-1)\cdot a_{ij}\cdot b_{ij}(y_t)\cdot\beta(j,t)/P(\boldsymbol{y}|M) \tag{23}$$

図6に図4に対する $\beta(i,t)$ の図4に対する計算例を示す．これから，パラメータ π_i, a_{ij}, $b_{ij}(k)$ を次の再推定を繰り返すことによって求めることができる．

$$\hat{\pi}_i=\frac{\sum_j\gamma(i,j,1)}{\sum_t\sum_j\gamma(i,j,1)}$$

$$\hat{a}_{ij}=\frac{\sum_{t=1}^{T}\alpha(i,\ t-1)\cdot a_{ij}\cdot b_{ij}(y_t)\cdot\beta(j,t)}{\sum_t\alpha(i,\ t)\cdot\beta(i,\ t)}=\frac{\sum_t\gamma(i,j,t)}{\sum_t\sum_j\gamma(i,j,t)} \tag{24}$$

$$\hat{b}_{ij}(k)=\frac{\sum_{t:y_t=k}\gamma(i,j,t)}{\sum_t\gamma(i,j,t)} \tag{25}$$

ただし，ここで注意を要することは，訓練サンプル数が n 個ある場合，これらの全サンプルについて上式を用いて1回パラメータを更新し，これを収束するまで更新を繰り返さなければならないことである．1サンプルごとによるパラメータ更新ではパラメータの収束が保証されない．

Baum-Welch の再推定プログラムが正常に動作しているかどうかのデバッグのためには次の項目のチェックが有効である．

① $\sum_j a_{ij}=1$, $\sum_k b_{ij}(k)=1$

② $\sum_{i\in F}\alpha(i,T)=\sum_i\beta(i,0)\pi_i$ (26)

③ $\hat{P}(\boldsymbol{y}|\hat{\boldsymbol{\lambda}},\ M)\geqq P(\boldsymbol{y}|\boldsymbol{\lambda},\ M)$ (27)

ここで，$\hat{\boldsymbol{\lambda}}$ は HMM の M のパラメータで，$\boldsymbol{y}$ による $\boldsymbol{\lambda}$ からの再推定値である．また，上述の再推定アルゴリズムの $\alpha(i,t)$, $\beta(i,t)$ は，計算途中で非常に小さな値となりアンダーフローの問題が生じるので，スケーリングを行う必要がでてくる[16]．

3・3 基本モデルの拡張[16]

(a) 観測シンボルが連続分布をとる場合（連続分布モデル）

今までは，図4の例のように，観測シンボルが有限個のうちの一つが出力される場合を述べてきた．しかし，特徴パラメータをベクトル量子化などの手法で有限個のシンボルに変換するのは量子化誤差などを伴い好ましくない場合が多い．特に不特定話者の認識の場合は，認識精度の低下を招く．

出力シンボルが，無限個の連続分布をなす場合，$b_{ij}(k)$ を離散的確率分布から連続的分布に変更する必要がある．一般に $b_{ij}(k)$ を多次元正規分布 $N(\boldsymbol{\mu}_{ij}, \Sigma_{ij})$ で近似される場合が多い．この場合，

$$\hat{\boldsymbol{\mu}}_{ij} = \frac{\sum_t \gamma(i,j,t)\cdot \boldsymbol{y}_t}{\sum_t \gamma(i,j,t)} \tag{28}$$

$$\hat{\Sigma}_{ij} = \frac{\sum_t \gamma(i,j,t)(\boldsymbol{y}_t-\boldsymbol{\mu}_{ij})(\boldsymbol{y}_t-\boldsymbol{\mu}_{ij})^t}{\sum_t \gamma(i,j,t)} \tag{29}$$

として再推定できる，ただし，実際には $b_{ij}(k)$ は一つの多次元正規分布で表現できないことも多いので混合分布（mixtured distribution）で近似することが試みられている[22]．

(b) 継続時間分布モデルの導入[24][25]

基本 HMM では状態 i に対して状態 i に τ 時刻（τ フレーム入力）とどまる確率は，

$$d(\tau) = a_{ii}^{\tau-1}(1-a_{ii})$$

で与えられ，τ と共に指数関数的に減少し，通常の音声現象と一致しなくなる．そこで，状態 i に τ 時刻とどまる確率 $d_i(\tau)$ を離散分布モデルとして推定する方法が提案されている[25]．このとき，

$$\hat{a}_{ij} = \frac{\sum_t \sum_{\tau \leqq t} \alpha(i,t-\tau)a_{ij}d_j(\tau)\prod_{k=1}^{\tau} b_{ij}(y_{t+1-k})\beta(j,t)}{\sum_t \alpha(i,t)\beta(i,t)}$$

$$\hat{b}_{ij}(h) = \frac{\sum_t \sum_{\tau \leqq t} \alpha(i,t-\tau)a_{ij}d_j(\tau)c_h(\tau)}{\sum_t \sum_{\tau \leqq t} \alpha(i,t-\tau)a_{ij}d_j(\tau)\cdot\tau} \times \frac{\prod_{k=1}^{\tau} b_{ij}(y_{t+1-k})\beta(j,t)}{\prod_{k=1}^{\tau} b_{ij}(y_{t+1-k})\beta(j,t)}$$

$$\hat{d}_j(\tau) = \frac{\sum_t \sum_i \alpha(i,t-\tau)a_{ij}d_j(\tau)\prod_{k=1}^{\tau} b_{ij}(y_{t+1-k})\beta(j,t)}{\sum_t \alpha(i,t)\beta(i,t)}$$

で与えられる．ただし，$c_h(\tau)$ は $\{y_{t+1-\tau}, \cdots, y_t\}$ の τ 個の出力シンボルのうち h と等しい個数を示す．

3・4 連続単語音声認識への適用[16][26]

HMM による連続音声認識には，DP マッチングによる連続音声認識アルゴリズムがそのまま適用できる．ただし，注意を要することは，連続単語へ適用するためには，2段 DP 法以外は厳密な確率計算法（Baum-Welch のスコアリング）は用いることができず，Viterbi アルゴリズムによるスコアリングを使用する必要があるということである[3][16]．

図7は，HMM による連続単語認識のモデルを示している．このモデルでは単語間の接続に際しては何ら制限はない．もし，単語間の接続にオートマトン（正規文法）による制限を課す場合は，図8のようなモデルを考えればよい．ここで，状態 A，B，C は，合流（confluent）状態と呼ばれるもので，正規分布の非終端記号に対応している．図の例では，$A \to V_1B$, $C \to V_2B$ の書換え規則がある場合を示している．

このとき，前向きパスアルゴリズムは次のように表現できる．

$$\alpha_Y{}^v(1,t+1) = \max_{\{A \to vY\}} \{\alpha_Y{}^v(1,t+1), \alpha(A,t)\}$$

$$\alpha_Y{}^v(j,t+1) = \max_{1 \leqq i \leqq J^n} \alpha_Y{}^v(i,t)\, a_{ij}{}^v b_{ij}{}^v(y_{t+1})$$

$$\alpha(B,t+1) = \max_{\{X \to vB\}} \alpha_B{}^v(J^n, t+1)$$

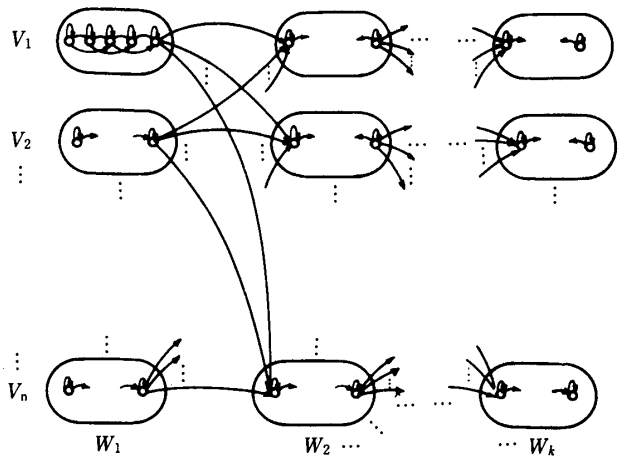
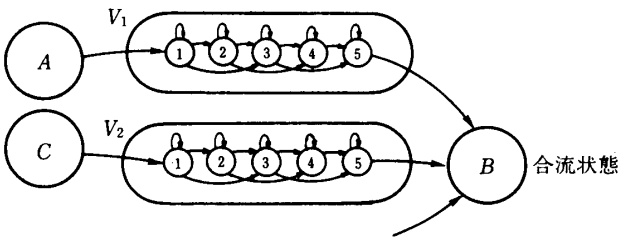


図 7 HMMによる連続単語音声認識モデル



$A \longrightarrow V_1B$

$C \longrightarrow V_2B$

図 8 HMMによるオートマトン制御連続単語音声認識モデル

ここで，J^nは各単語のHMMの最終状態を示している．もちろん，このほかに，単語の接続時点の記憶，前状態の記憶，単語名の記憶が必要であり，これらの記憶方法はDPマッチング法における場合と同一である．

4. 今後の展開

4・1 文脈自由文法および係り受け規則の組込み[16]

2・3節で述べたオートマトン制御DPマッチング法を文脈自由文法（プッシュダウンオートマトン）制御に拡張することが試みられている．文頭から文末方向へのEarleyのアルゴリズムによる下降型構文解析法[27]やCYKアルゴリズムによる上昇型構文解析法[28]をDPマッチングによる連続音声認識に組み込む方法が提案されている．

また，音声認識特有の解析法として最も確からしいところから左右両方向へ構文解析を進める方法（島駆動方式）もDPマッチングに組み込まれ，文頭から文末に向かって(left-to-right)構文解析する方法との比較が行われている[29]．

係り受け規則を満たす最適な文節列を選択する音声認識アルゴリズムも提案されている[30]．しかし，二文節間の係り受け規則は二分木の構文解析木で表現でき，CYKアルゴリズムと本質的に同じである[31]．また，二文節間の係り受け規則だけではあいまいな構文解析結果を多く生じるので，格構造などの導入が必要であろう[32]．

4・2 言語の確率モデルの導入[16]

われわれは自然言語音声を理解する場合，話題についての膨大な知識を用いて，発話内容や発話単語を推定し，理解の焦点を絞っている．また，文法的に許される文はどれも同じ確率で生じるものではない．このような現象を確率モデルでとらえる方法が試みられている．一つは音韻・音節・単語連鎖などの三つ組の生起確率を用いる方法，他の一つは確率文法を用いる方法である．これらは，話題の推移確率などを含めて一つの確率モデルとして組み込まれていくであろう．

4・3 ニューラルネットワークの導入[16]

音声認識などのパターン認識は，記号処理とか知識処理とかでなく神経回路に組み込まれたメカニズムで行われている可能性が強い．その一つの例が，DPマッチングやHMM法でありニューラルネットワークである．ニューラルネットワークも音声認識に適用されているが，時系列パターンの照合に関してはこれからの研究に待たなければならない．また，言語処理部においても意味主導型のシステムはパターン間の連想を中心とするものが多く，ここにもニューラルネットワークの適用がなされよう[33]．

◇参考文献◇

(1) Atal, B. S. and Rabiner, L. R. : Speech Research Direction, *AT&T Tech. Journal*, Vol. 65, Issue5, pp. 75-88 (1986).

(2) Rabiner, L. R., Juang, B. H., Levinson, S. E. and Sondhi, M. M. : Recognition of isolated digits using Hidden Markov Models with continuous mixture densities, *AT&T Tech. Journal*, Vol. 64, No. 6, pp. 1211-1234 (1985).

(3) 坂井利之（編）：情報基礎学詳説，コロナ社 (1983).

(4) Vintsyuk, T. K. : Speech discrimination by dynamic programming, *Cybernetics*, Vol. 4, No. 1, pp. 52-57 (1968).

(5) 迫江，千葉：動的計画法を利用した時間軸正規化に基づく連続音声認識，音響誌，Vol. 21, No. 9, pp. 438-490 (1971).

(6) Sakoe, H. and Chiba, S. : Dynamic programming optimization for spoken word recognition, *IEEE Trans.*, Vol. ASSP-26, No. 1, pp. 43-49 (1978).

(7) Vintsyuk, T. K. : Elementwise recognition of continuous speech composed of words from a specified dictionary, *Cybernetics*, No. 2, pp. 361-372 (1971).

(8) Sakoe, H. : Two-level DP-matching—A dynamic programming based pattern matching algorithm for connected word recognition, *IEEE Trans.*, Vol. ASSP-27, No. 6, pp. 588-595 (1979).

(9) 中川聖一：パターンマッチング法による連続単語および連続音節の音声認識アルゴリズム，信学論誌，Vol. 66D, No. 6, pp. 637-644 (1983).

(10) Bridle, J. S., *et al.* : An algorithm for connected word recognition, *Proc. Int. Conf. ASSP*, pp. 899-902 (1982).

(11) Ney, H. : The use of a one-stage dynamic programming algorithm for connected word recognition, *IEEE Trans.*, Vol. ASSP-32, No.3, pp. 263-271 (1984).

(12) 迫江，亘理：クロック同期伝搬型DP法による連続音声認識の検討，音響学会音声研資，S81-65 (1981. 12).

(13) Myers, C. S. and Rabiner, L. R. : A level building dynamic time warping algorithm for connected word recognition, *IEEE Trans.*, Vol. ASSP-29, No. 2, pp. 284-297 (1981).

(14) 中川聖一：拡張連続DP法による連続認識アルゴリズム，信学論誌，Vol. 67D, No. 10, pp. 1242-1249 (1984).

(15) Sakoe, H. : A generalized two-level DP matching algorithm for continuous speech recognition, 信学論誌, Vol. 65E, No. 11, pp. 649-656 (1982).

(16) 中川聖一：確率モデルによる音声認識，電子情報通信学会

(1988).
(17) Jelinek, J. : Continuous speech recognition by statistical methods, *IEEE Proceedings*, Vol. 64, No. 4, pp. 532-556 (1976).
(18) Rabiner, L. R., Levinson, S. E. and Sondhi, M. M. : On the application of vector quantization and Hidden Markov Models to speaker-independent isolated word recognition, *BSTJ*, Vol. 62, No. 4, pp. 1075-1105 (1983).
(19) Levinson, S. E. : Structural methods in automatic speech recognition, *IEEE Proceedings*, Vol. 73, No. 11, pp. 1625-1650 (1985).
(20) Rabiner, L. R. and Juang, B. H. : An introduction to hidden Markov models, *IEEE ASSP Magazine* (1986. 1).
(21) 大河内正明：Hidden Markov Model に基づく音声認識, 音響学会誌, Vol. 42, No. 12, pp. 936-941 (1986).
(22) Liporace, L. A. : Maximum-likelihood estimation for multivariate observations of Markov process, *IEEE Trans.*, Vol. IT-28, No. 5, pp. 729-734 (1982).
(23) Juang, B. H. and Rabiner, L. R. : Mixture auto regressive hidden Markov models for speech signal, *IEEE Trans.*, Vol. ASSP-33, pp. 1404-1413 (1985).
(24) Levinson, S. E. : Continuously variable duration hidden Markov models for automatic speech recognition, *Computer Speech and Language*, 1, pp. 29-45 (1986).
(25) 橋本, 中川：HMM 法による連続音声のセグメンテーションの検討, 音響学会講論集, 3-P-2 (1988.3).
(26) Fallside, F. and Woods, W. A. (ed.) : Computer Speech Processing, Prentice Hall (1985).
(27) 中川聖一：文脈自由文法のフレーム同期型構文解析法による連続音声認識, 信学論誌, Vol. 70D, No. 5 (1987).
(28) Ney, H. : Dynamic programming speech recognition using a context-free grammar, *Proc. ICASSP*, pp. 69-72 (1987).
(29) 大黒, 中川：Left-to-right & top down 構文解析法と island-driven & bottom-up 構文解析法による連続音声認識の比較・検討, 信学会言語処理とコミュニケーション技報, MLC 87-12 (1987).
(30) 尾関和彦：文節ラチスから最適係り受け構造を選択する多段決定アルゴリズム, 信学論 (D), J70-D, 12, pp. 2621-2629 (1987).
(31) 中川聖一：連続音声認識のための係り受け解析法と文脈自由文法解析法との統一化, 音響学会講論集, 3-5-10 (1987. 10)
(32) 中川, 伊藤：音節標準パターンと逆時間向き係り受け解析法を用いた日本語文音声の認識, 信学論 (D), J70-D, 12, pp. 2469-2478 (1987).
(33) 西田豊明：コネクショニストモデルによる自然言語処理—可能性を探る—, 情報処理学会, 自然言語処理シンポジウム論文集 (1988. 1).
(34) Sankoff, D. and Kruskal, J. (ed.) : Time Warps, String Edits and Macromolecules ; The Theory and Practice of Sequence Comparison, Addison-Wesley (1983).

著者紹介

中川 聖一

昭和 51 年京都大学大学院博士課程修了. 同年, 京都大学工学部情報工学科助手. 昭和 55 年豊橋技術科学大学情報工学系講師, 昭和 58 年助教授. 昭和 60～61 年カーネギーメロン大学客員研究員, 工学博士. 音声情報処理, 自然言語処理の研究に従事. 昭和 52 年度電子通信学会論文賞受賞. 著書「情報基礎学詳説」(共著) コロナ社 (昭和 58 年),「確率モデルによる音声認識」電子情報通信学会 (昭和 63 年).