音声認識とDP

迫江 博昭

1. まえがき

ダイナミックプログラミング(DP)が音声認識の分野で重用されるようになって15年になる．この間，多数の研究がなされ，幾多の論文が発表され，DPは音声認識の基本的な手法としての地位を確立した．今日では脱DPを模索するのが，一部研究者の命題となっているほど，この分野でのDPの存在は大きい．しかし，研究の中心となった人たちがORとは縁が遠かったため，DPそのものの研究グループには，音声認識とDPとのかかわりは，ほとんど知られていなかったようである．

著者自身も，DPの入門書を一読して，その考え方を自分にわかる範囲で把握したあとは，もっぱらその応用を考えるのみで，DPそのものの理解を深めるための努力はしなかった．元祖 R. Bellman 氏に対しても，伝説上の人物といった認識しかなく，発表した大半の論文でも氏の著作を引用するのを省略してしまっている．DPはそれほど有名であったというのが，故人となられたベルマン氏に対する言いわけである．

本稿では音声認識に対するDPの応用を解説する．筆者はこの研究の歴史のはじまりから深く関係し，現役研究者としての一生をこのテーマにささげたものである．このため解説に筆者の主観が相当程度入ることになると思う．また，いくつかの術語に，DPの教科書とは異なったものを用いる．これらは音声認識の分野の方言として定着してしまったものである．

さこえ　ひろあき　日本電気㈱　C&Cシステム研究所

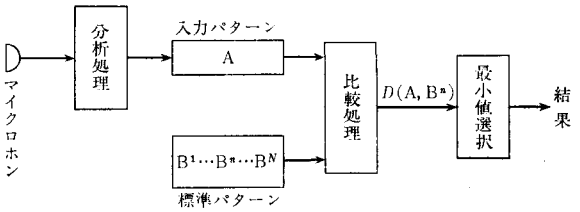

図 1　パターンマッチングによる音声認識

2. 音声認識入門

音声認識のモデルとしては種々のものが提案されているが，最も基本的なものを図1に示す．パターンマッチング法にもとづくものである．入力音声波形は分析部で周波数分析され，短時間スペクトラムの時系列パターンに変換される．図2に数字“3”のパターンの例を示す．数式的には

$$A=a_1a_2\cdots\cdots a_i\cdots\cdots a_I \qquad (1)$$

と，ベクトルの時系列として表現する．a_i は時刻 i でのスペクトラム特徴を示すベクトルである．単語名を番号 n で示すこととし $\{n|n=1,2\cdots\cdots N\}$ なる N 個の単語セットを考える．単語 n には標準パターン $\{B^1,B^2\cdots\cdots B^n\cdots\cdots B^N\}$ が用意されている．これらは(1)のAと同じくベクトルの時系列である．代表的なものを，添字を省略して

$$B=b_1b_2\cdots\cdots b_j\cdots\cdots b_J \qquad (2)$$

	振幅レベル	周波数 (Hz) 250				1000				2000					3000	4000	5000
/S/	135	9	5	5	5	5	7	7	7	7	7	7	7	11	11	54	92
	179	8	7	7	7	5	5	7	7	7	7	7	7	8	9	56	95
	247	7	5	6	5	5	5	6	8	8	8	6	6	9	11	58	98
	290	5	4	4	4	5	6	7	8	9	7	7	6	10	11	58	98
	321	4	5	4	6	6	7	6	9	10	7	7	8	10	10	55	92
	380	5	6	5	5	5	7	6	9	9	7	8	10	14	14	43	95
	437	7	5	5	4	5	7	7	8	9	8	7	8	14	14	44	96
	500	8	6	5	5	5	6	7	7	8	9	7	7	12	12	51	93
	640	7	7	5	5	6	5	6	6	6	8	6	7	13	13	52	94
	818	13	12	8	6	5	5	6	8	8	6	4	7	14	11	54	80
	857	18	16	17	21	15	13	10	9	7	4	4	7	11	6	41	47
/a/	1244	11	12	29	46	37	31	13	5	4	3	4	8	9	5	15	13
	1642	6	9	28	49	47	44	14	3	4	4	4	6	9	7	8	5
	1750	5	7	24	46	48	52	17	4	4	4	5	5	7	7	7	5
	1673	5	6	22	45	44	53	23	5	3	4	5	6	6	5	7	5
	1365	6	8	22	42	41	54	25	6	3	6	7	5	5	4	7	5
	816	11	14	22	36	38	46	20	6	5	9	11	8	7	4	6	4
/N/	459	20	20	20	23	28	22	11	7	6	13	17	15	21	11	9	6
	391	24	21	18	17	21	12	8	7	6	14	18	16	30	16	11	7
	352	27	22	18	15	19	10	8	7	5	13	19	16	29	18	11	8
	308	29	24	19	15	19	10	8	7	5	13	19	16	27	18	11	8
	282	35	25	18	15	19	9	7	6	5	12	19	16	25	17	11	8
	248	41	26	19	14	19	10	7	6	5	12	18	14	23	14	13	8
	170	54	28	19	15	19	9	7	6	4	9	13	12	21	12	13	10

時間軸 ↓

各周波数成分の強さ

図 2　音声パターンの例(数字“3”)

と記す.

未知の入力パターンAが与えられると，これらの標準パターンとのあいだで比較処理（マッチング)が行なわれ，相異の度合いを示す尺度として，距離 $D(A, B^n)$ が算出される．これらの距離が最小となる単語名 $n=\hat{n}$ が認識結果として出力される．

以上は最も簡単な離散単語の認識の場合で，入力パターンも標準パターンも孤立発声した単語パターンとして扱われる．利用する各人の音声パターンを標準パターンとして登録する型のものを特定話者型と呼ぶ．この型のものでは，本人の標準パターンを用いることによって声質の個人差の影響を避けるのである．不特定の話者を対象とする型では，多数の人の音声パターンの中から，代表的なパターン(通常各単語ごとに複数個)をクラスタリング手法で選び，標準パターンとする．

2個以上の単語を連続発声したものの認識のためには,図1のモデルを若干修正する必要がある．これについては，4章で説明する．

3. 離散単語認識における DP

3.1 音声パターンの時間軸歪

パターンの比較をどう行なうかは，音声認識の性能を大きく左右する．音声パターンに生ずる各種の変動を十分考慮した方式を用いないと良好な精度は期待できない．主要な変動要因として時間軸歪がある．同一人が同じ単語を発声しても，まったく同一の速度で発声するとは限らない．このため，図3に示すように，入力パターンAと標準パターンBとのあいだに歪が発生する．このため時間的に対応するベクトル $\boldsymbol{a}_i$ と $\boldsymbol{b}_i$ との距離の総和を求めるといった比較方法は，きわめて根拠のない方法となる．以前は線形な伸縮によって長さを揃えてからマッチングするという方法（線形伸縮法）が試みられたが不十分であった．音声パターン中の各部は独立に伸縮するので全体としては複雑な非線形伸縮となる．線形伸縮は1次近似に

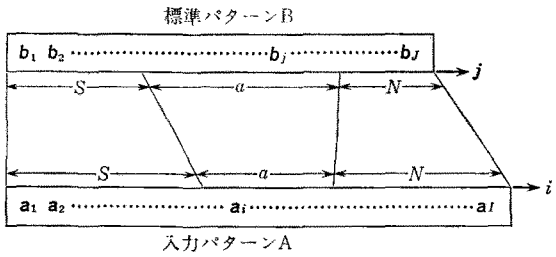

図 3 音声パターンの時間軸歪

しかならなかったのである．

3.2 DPマッチング

時間軸歪の問題をDPの導入によって解決できた．アルゴリズムには，いくつかの分化があるがDPマッチングと総称されている．

図4は入力パターンAを横軸に，標準パターンBを縦軸に配したものである．この図で，標準パターンの時間軸 j を関数 $j(i)$ によって入力パターンAの時間軸 i に対応づけ，時間軸歪を除くことを考える．

関数 $j(i)$ をどういう基準で定めるとよいのであろうか．そこで，時間軸歪がマッチングにおよぼす影響を定性的に考えてみる．AとBとがまったく同じパターンであるならばその間の距離は0である．時間軸歪が生じ，その度合いが大きくなると，両者のあいだの距離も大となる．したがってAとBとの距離が0となるように写像 $j(i)$ を定めればよい．一般には，AとBのあいだには時間軸歪以外の他の変動要因による歪も入っているので，距離0とはならないが，距離を最小にすると時間軸歪を除去できたと考えてよい．すなわち

$$D(\mathrm{A},\mathrm{B})=\min_{j=j(i)}\left[\sum_{i=1}^{I} d(i,j)\right] \qquad (3)$$

なる最小化問題として定式化できる．ここに $d(i,j)=\|\boldsymbol{a}_i-\boldsymbol{b}_j\|$ で，ベクトル $\boldsymbol{a}_i$ と $\boldsymbol{b}_j$ との距離である．

図4と(3)式を並べると，おのずからDPの適用が連想される．すなわち図4の各 (i,j) 座標に距離 $d(i,j)$ が対応しているとき，それらの総和が最小になるようにするという最小コスト問題となっているのである．

具体的にDPを適用する前に，写像の関数 $j(i)$ の性質について考えてみる．ここでは一時，時間軸 i,j は連続的と考えて説明する．

1) $j(i)$ は連続関数．(時間から時間への写像だから当然)
2) $j(i)$ は単調増加．(現象の前後関係を保つため)
3) $j(1)=1,\ j(I)=J$．(始端や終端が切れてしまわないため)
4) $j(i)\sim i$．(実際には時間軸歪があまり極端になることはないという事実に対応)

これらの性質が(3)式の最小化問題に対する制約条件となる．ただし，実際には離散系であるので，これらの条件は一部近似的に実現される．

次節に入る前に次の点を注意しておく．パターンマッチングの目的には(3)式の最小値 $D(\mathrm{A},\mathrm{B})$ が求められればよいので，最適な写像 $j(i)$ の決定は省略する．$D(\mathrm{A},\mathrm{B})$ はAとBとのあいだの時間歪を除去しても，なお残る距離であって，時間正規化距離と呼ばれることもある．(一般的に言っ

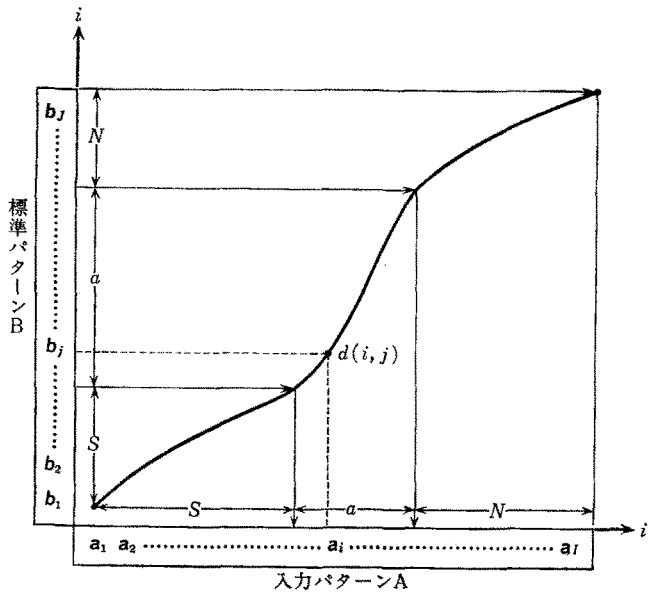

図 4 最適経路問題への変換

て距離の公理を満足するものではないが，慣用的に距離と呼んでいる）

3.3 実行アルゴリズム

前節の(3)式の最小化問題を解くのであるが，条件1)，2)，3)，4)の実現の仕方によって種々の実行アルゴリズムが分化している．以下には最も簡単なものを示す．

図5の(i,j)格子点で次のDP計算を行なう．

○初期条件

$$g(1,1)=d(1,1) \tag{4}$$

○漸化式

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

○整合窓条件

$$j-r\leqq i\leqq j+r \tag{6}$$

これによって距離((3)式の最小値)は

○距離

$$D(\mathrm{A},\mathrm{B})=g(I,J) \tag{7}$$

と定まる．

漸化式(5)は時刻iにおける点(i,j)に至る経路として，図5中に示す3種を許すという形になっている．これは前節の1)，2)の性質を離散座標系で近似的に実現したものである．3)の性質は(4)式と(7)式で実現されている．4)は整合窓条件(6)式によって0次近似として実現されている．

(6)式の条件は，制約条件によって問題が簡単化されるというDPの好ましい性質にマッチしたものになっている．漸化式(5)の計算を，図5の整合窓の範囲内に限定し，処理量を低減している．

表1 DPマッチングの効果

方　法	誤 認 識
線形伸縮法	5.9%
DPマッチング	0.8%

認識対象は地名50語

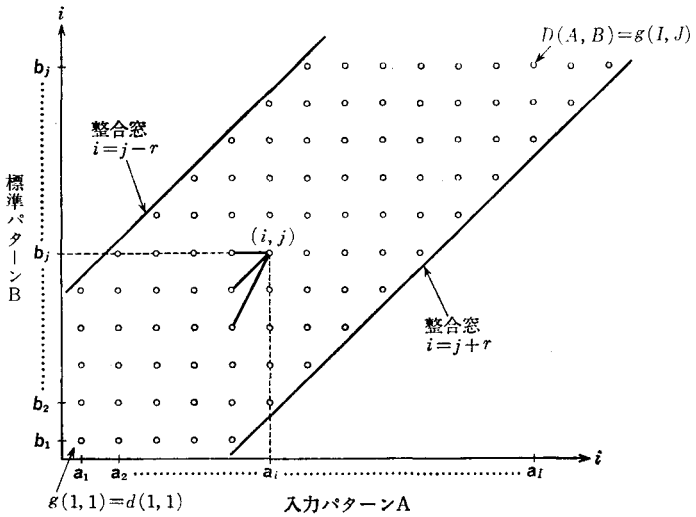

図5 DPの計算

また音声認識の分野では，音声パターンの終端$i=I$が不明な場合がある．この場合には，終端開放の条件で(5)式を計算すると

$$J-r\leqq I\leqq J+r$$

の範囲の各終端候補Iに対して(7)式により距離$D(\mathrm{A},\mathrm{B})$が逆列に評価できる．これもDPならの性質であるが，後に述べる連続単語認識などで，有効なものである．

以上のようなDPの適用によって，非線形な時間軸歪み問題が実用レベルとして解決された(1970)．表1にDPマッチングと，代表的な従来手法である線形伸縮法との比較実験の結果を示す．誤認識が約1/7に低減されるという効果が得られている．

図6にDP漸化式の変形の例を示す．(a)が図5(すなわち(5)式)のものである．(c)は写像$j(i)$の傾斜を近似的に$1/2\leqq j'\leqq 2$と制限した型であり，前節の条件4)を1次近似まで実現したものである．この型のものが，認識精度の点で最も良いことが実験的に確認されている．表1のデータは，この型を用いた場合のものである．

以上で離散単語認識のDPマッチングの説明を終る．不可能を可能にしたと言えるほどDPの効果は顕著であった．(3)式あるいは図4の最適経路問題に対するDPの効用は多くの読者に自明の

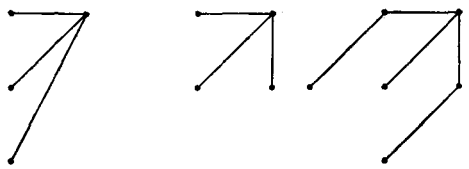

図 6　経路条件の変形

ことと思う．DPによって導びかれた(5)式の漸化式には距離 $d(i,j)$ の計算，最小値選択，加算，と3種の処理が含まれているが，計算量的にはベクトル $\boldsymbol{a}_i$ と $\boldsymbol{b}_j$ の距離を計算するための処理が大半を占める．すなわち，DPの効用があまりにも大きいので，DPマッチングのための計算は，ベクトル間距離計算という，DPの責任外の処理のために，ほとんどが費やされるのである．

4.　連続単語認識における DP

いくつかの単語を連続して発声したものを認識しようとする場合にはセグメンテーションの問題が発生する．連続した単語と単語の境界が不明なので，これを決定しないと前章のような単語単位でのパターンマッチングが適用できない．このためセグメンテーションに関する研究が数多くなされたが，境界そのものがきわめて曖昧模糊なものであるため，信頼性のある手法は確立されていない．極端には "21"(/ni itʃi/)のように同じ音 i がつづいて，境界決定が原理上不可能な場合もあるのである．

このため，図7に示すように，入力パターンを分解するかわりに，単語標準パターンを接続して連続標準パターンを合成し，パターンマッチングを行なうという原理を採用した．4桁の連続数字(棒読み)の例で説明すると，まず "0"～"9" の標準パターン B^0～B^9 を用意する．入力音声パターンAとして，連続数字 "5892"(実際には未知)が入力されたとする．4桁の数字 "0000" から "9999" までの数字（一般に $n(1),n(2),n(3),n(4)$ とする）の連続標準パターン $B^{n(1)}\oplus B^{n(2)}\oplus B^{n(3)}\oplus B^{n(4)}$ を合成して入力パターンAとDPマッチングを行なう．距離 $D(A, B^{n(1)}\oplus B^{n(2)}\oplus B^{n(3)}\oplus B^{n(4)})$ をそれぞれ求め，最小となる $n(1),n(2),n(3),n(4)$ を決定し，これを認識結果とする．一般的には次の最小化問題となる．

$$\min_{\{n(x)\}}[D(A, B^{n(1)}\cdots\cdots\oplus B^{n(x)}\cdots\cdots\oplus B^{n(k)})] \quad (8)$$

この原理によると確かにセグメンテーションの問題を回避できる．しかし，新たに順列 $\{n(x)\}$ に関する最小化という問題が生じた(距離 D を算出するための写像関数 $j(i)$ に関する最小化が埋め込まれているが，これは前章で解決ずみである)．この最小化問題を総当り法で解くと，k 数字連続の場合，10^k 種の連続標準パターンに関してDPマッチングのくりかえしが必要となる．そこで再びDPの適用を試みる．

入力パターンAの $i=l+1$ から $i=m$ までの部分を部分パターン $A(l,m)=a_{l+1}\cdots\cdots a_m$ と定義する．(8)式を解くための補助変数として，入力パターンAの中に $(k-1)$ 個の区分点 $l(1)\cdots\cdots l(x)\cdots\cdots l(k-1)$ を仮定する．これによって入力パターンAは $A=A(l(0), l(1))\cdots\cdots\oplus A(l(x-1), l(x))\cdots\cdots\oplus A(l(k-1), l(k))$ と，k 個の部分パ

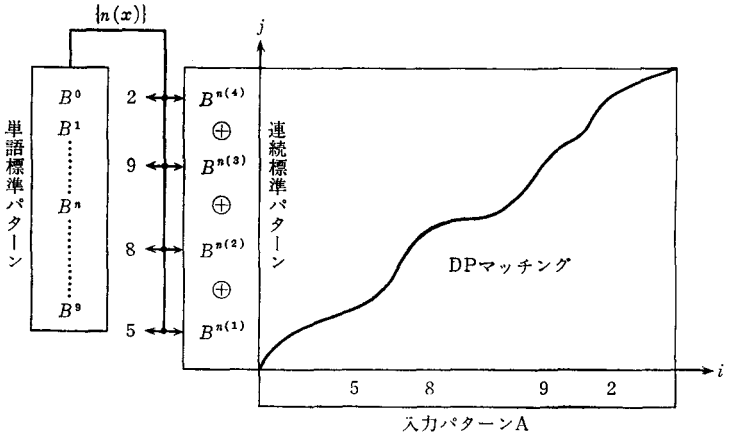

図 7　連続標準パターンの合成とDPマッチングによる連続単語認識の原理

ーンの接続として表現される．ここに $l(0)=0$, $l(k)=I$. これを(8)式に代入して整理すると

$$\min_{\{l(x)\}}\left[\sum_{x=1}^{k}\min_{n(x)}\left[D(\mathrm{A}(l(x-1),\ l(x)),\mathrm{B}^{n(x)})\right]\right] \quad (9)$$

と2段階の最小化に分解される．内側の最小化は $n(x)$ に関する単純比較なので，一般的に次のように計算できる．

$$\hat{D}(l,m)=\min_{n}\left[D(\mathrm{A}(l,m),\mathrm{B}^{n})\right] \quad (10)$$

この最小値を与える単語名

$$\hat{N}(l,m)=\operatorname{argmin}_{n}\left[D(\mathrm{A}(l,m),B^{n})\right] \quad (11)$$

は，部分パターン $\mathrm{A}(l(x-1),\ l(x))$ が1個の単語であると仮定して認識したときの結果である．

(10)式を用いると(9)式は

$$\min_{\{l(x)\}}\left[\sum_{x=1}^{k}\hat{D}(l(x-1),l(x))\right] \quad (12)$$

となるので，次のようなDPによって計算できる．

○初期条件

$$T(0,0)=0$$

○漸化式

$$T(x,m)=\min_{l<m}\left[\hat{D}(l,m)+T(x-1,l)\right] \quad (13)$$

この(13)式を $x=5, m=I$ まで計算する．この間，(13)式の l の最適値を x, m に対応づけて記憶しておくとバックトラックによって(12)式の最適解

$$\{\hat{l}(x)\,|\,x=0,1,\cdots,k\}$$

が定まる．これをもとに(11)式の $\hat{N}$ のテーブルを参照することにより

$$\{\hat{N}(\hat{l}(x-1),\hat{l}(x))\}$$

として認識結果が得られる．

表 2　2段DPマッチングによる連続数字認識実験結果

連続桁数			1	2	3	4	認識率
発声音	A	誤り数	0	0	0	0	100 %
	B		0	0	1	0	99.8%
	C		0	0	0	1	99.8%
	D		0	2	0	2	99.2%
	E		0	0	0	5	99.0%
平均認識率			100%	99.6%	99.9%	99.2%	99.6%

テストデータは1～4桁数字を100種ずつ発声したもの

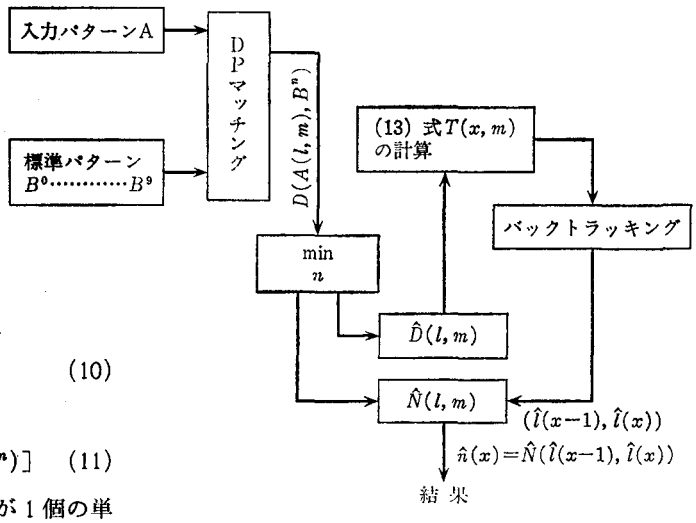

図 8　2段DPマッチングのブロック図

以上の処理のブロック図を図8に示す．この方法では(10)式等の距離 $D(\mathrm{A}(l,m),\mathrm{B}^n)$ を計算するために前章で述べた単語単位でのDPが必要とされ，これに加えて(13)式のDPが計算される．これにちなんで2段DPマッチング法と呼んでいる．この方法では3.3節のDPの並列性が有効に利用できる．(10)，(11)の計算のための距離 $D(\mathrm{A}(l,m),\mathrm{B}^n)$ は各 (l,m) に対して独立に算出する必要はない．始端 l を定めてDPマッチングを行なうと，整合窓内の m に対して並列的に $D(\mathrm{A}(l,m),\mathrm{B}^n)$ が得られるからである．

2段DPマッチング法によって，連続単語認識がはじめて可能になった(1975)．表2に連続数字認識に適用した実験結果を示す．その後，種々の改良が行なわれたが，主として計算効率の向上を目的としたもので，連続標準パターンとのマッチングをDPによって処理するという原理は今も生きている．

5. ハードウェア，LSI

DPの処理を実行するための専用のハードウェアが作られ，LSI化まで実現しているのは音声認識の分野だけではなかろうか？　実用音声認識装置となると，コンピュータ・シミュレーションと異なって，実時間動作が必須である．このため，

やや強引にハードウェア化が進められたとも言える．

1978年に発売され，世界で初のＤＰ式音声認識装置となった日電のＤＰ-100には，約400個のＩＣを使ったＤＰプロセッサが使用された．その後，半導体技術が進歩し，現在では１個のＬＳＩでほぼ同等の機能が実現されている．図９はその拡大写真である．マイクロプログラム制御を採用したので，ＤＰ特有の回路構成といった特徴は見えないが，マイクロインストラクション等にＤＰ向けの工夫がなされている．

6. あとがき

以上，音声認識の分野におけるＤＰの応用の一端を紹介した．ここで述べたものは単語を単位としてパターンマッチングを行なう方式で技術的に完成度が高く，実用化されているものである．今後の方向として，大語彙化をめざして音素や音節を単位とした認識が必要と考えられ，研究が進められている．この中でもＤＰは基本的手法の１つとして位置づけられている．

音声認識の他に文字認識においてもＤＰが利用されている．これらを含めて，日経エレクトロニクス誌(1983年11月７日)に解説しているので興味をおもちの方は参照されたい．

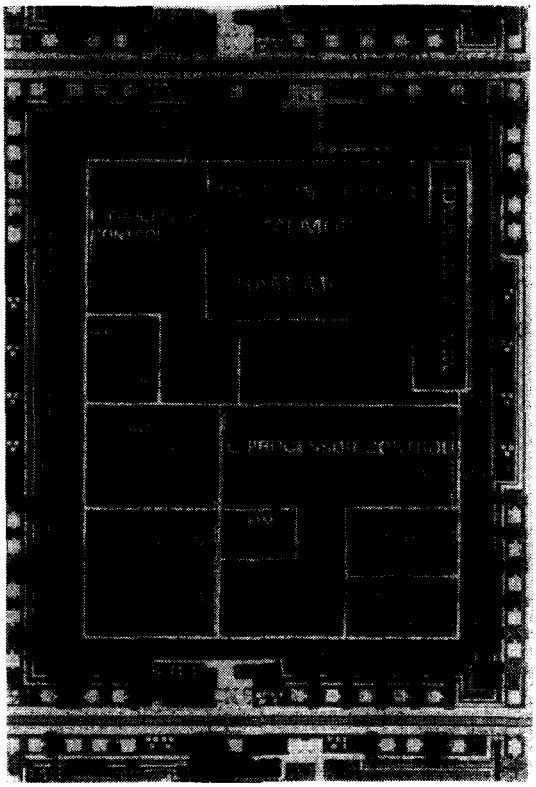

図９　ＤＰマッチングＬＳＩ

1970年以来，音声認識の分野でＤＰが果してきた役割は非常に大きい．特に実用面での評価は高く，筆者らが何となく用いた“ＤＰマッチング”という技術名が，多くの研究論文はもとより，各社の製品カタログでも技術的な裏づけを示すためのＰＲ用語として，引用されている．筆者を含む多くの研究者が，ＤＰによって大きな仕事をすることができた．故ベルマン教授に誌上を借りて心からお礼を述べたい．