MMS 23－2
(1976. 1. 20)

DPを用いた連続単語音声認識システム

鶴田七郎　迫江博昭　千葉成美
（日本電気　中央研究所）

1. まえがき

コンピュータ利用が拡大されるにつれて、マン・マシン・システムにおける音声認識の必要性は、情報入力の手段として、1)人間にとって、より負担が軽い、2)最も自然な形で実現できる、等の観点より近年急速に高まりつつある[1]。人間が行なうと同等の音声認識が究極的目的であるが、現在の技術をもってしてもこの目標に到達したとは到底いえない段階である。しかしながら、何らかの制限、例えば認識できる語数、あるいは特定の発声者への限定等を付加した場合、認識率、経済性共に優れた実用的見地から期待のもてる成果が得られてきている。一方、音声認識の実用的な利用形態を検討してみると多くの場合、情報の入力速度の向上あるいは、使い易さ等の点から、データ等を連続発声した、いわゆる連続単語が認識できることが望ましい。筆者らは、連続単語の認識が可能な認識法として、動的計画法（DP）を利用した時間正規化能力を有するパタンマッチング法であるDPマッチング法[2]を開発し、実時間で動作する認識システムを開発してきた[3][4][5]。この程このシステムに改良を施し、連続音声に対して、より安定に認識動作をする2段DPマッチング法[6]に基づいた実時間認識システムを実現した。

本文では、はじめに2段DPマッチング法の原理と認識アルゴリズムを述べ、続いて、試作した実時間システムの構成について述べ、更にこのシステムを用いて行なった評価実験の結果を述べる。

2. 2段DPマッチング法と連続単語認識アルゴリズム[7]

原理　音声パタンを特徴ベクトルの系列として表現する。

$$A = a_1, a_2, \cdots a_i, \cdots a_I \quad ; \quad B = b_1, b_2, \cdots b_j, \cdots b_J \tag{1}$$

パタンBの時間変動を関数 $j=j(i)$ によってモデル化し、関数 $j=j(i)$ は
i)時間軸の連続性，単調性，ii)音声パタン全体の保存，iii)時間変動量の制約
を満足するものとする。この変動モデルを用いるとパタンBの時間軸をAの時間軸に投影して、両パタンの類似性を比較することが可能になる。関数 $j(i)$ を最適に定め、AとBの間の類似度を次の様に定義する。

$$S(A,B) = \max_{j=j(i)} \left[\sum_{i=1}^{I} g(a_i, b_j) \right] \tag{2}$$

ここに、$g(a_i, b_j)$ は a_i と b_j の類似の度合を示す量である。(2)式の最大化はDPによって実行されることより、このマッチング法をDPマッチングと称する。次に単語 n ($n=1,2,\cdots N$) の標準パタンを(3)式で示す。

$$B^n = b_1^n, b_2^n, \cdots, b_j^n, \cdots, b_{J_n}^n \tag{3}$$

一方、連続単語音声パタン（入力パタン）をCで示し、

$$C = c_1, c_2, \cdots c_i, \cdots c_I \tag{4}$$

標準パタンB^nとB^mの接続パタンを$B^n \oplus B^m$と定義する。なお⊕は接続を意味する。K個のパタン$B^{n(1)}, B^{n(2)}, \cdots B^{n(x)}, \cdots B^{n(K)}$を接続したパタンを$\bar{B}$で示す。

$$\bar{B} = B^{n(1)} \oplus B^{n(2)} \oplus \cdots \oplus B^{n(x)} \oplus \cdots \oplus B^{n(K)} \tag{5}$$

ここで連続音声の入力パタンCと$\bar{B}$との類似度$S(C, \bar{B})$を含まれる単語の個数Kと標準パタンの種類$B^{n(x)}$に関して最大化し、最適パタメタ$K=\hat{K}$, $n(x)=\hat{n}(x)$ $(x=1, 2, \cdots \hat{K})$を決定すると、パタンCは、$\hat{n}(1), \hat{n}(2), \cdots, \hat{n}(\hat{K})$であると判定できる。すなわち、上述の内容は、

$$T = \max_{K, n(x)} \left[S(C, \bar{B}) \right] \; ; \; \hat{K}, \hat{n}(1), \hat{n}(2), \cdots \hat{n}(\hat{K}) = \arg \max_{K, n(x)} \left[S(C, \bar{B}) \right] \tag{6}$$

で表現することができる。ここで$\arg_n \max [\cdot]$は［ ］内の最大を与える変数nを算出することを意味する。しかし、Cの単語数Kが既知であるとしても$\bar{B}$はN^K種可能であって、(6)式の計算量は膨大となる。このため(6)式の計算を以下に述べる単語単位での処理と全体としての処理の2段階に分割して計算量の低減を図る。

今、入力パタンCの$i=l+1$より始まって、$i=m$で終る部分区間として部分パタン$C(l, m)$を定義する。ここでlを始端、mを終端と称する。また$(K-1)$個の区分点を仮定してK個の部分パタンに分割する。

$$C = C(1, l(1)) \oplus C(l(1), l(2)) \oplus \cdots \oplus C(l(x-1), l(x)) \oplus \cdots \oplus C(l(K-1), I) \tag{7}$$

但し、$1 < l(1) < l(2) < \cdots \quad < l(K-1) < I$ (8)

一方(2)式で定義した類似度は、入力パタンの分割に対して加法的である。この性質を使い(6)式に(7)式を代入することにより、

$$T = \max_{K, n(x)} \left[\max_{l(x)} \left[\sum_{x=1}^{K} S(C(l(x-1), l(x)), B^{n(x)}) \right] \right] \tag{9}$$

となる。この最大化及び総和の操作の順序を変更するとn(x)(単語の種類)に関するもの(第1段)と、l(x)(区分点)に関する(第2段)最大化に分解され次の様になる。

$$T = \max_{K, l(x)} \left[\sum_{x=1}^{K} \max_{n(x)} \left[S(C(l(x-1), l(x), B^{n(x)}) \right] \right] \tag{10}$$

この様に(6)式の最大化を2段階に分解した結果DPマッチングは単語単位で実行すればよいことになる。以上が2段DPマッチング法による連続単語認識の原理である。

アルゴリズム

i) 部分マッチング　部分パタンC(l, m)を入力パタンAと考えてDPマッチングを，l<mなる全ての(l, m)に対して，

$$\text{部分類似度}\quad \hat{S}(l,m)=\max_{n}\left[S(C(l,m),B^{n})\right] \tag{11}$$

$$\text{部分判定結果}\quad \hat{N}(l,m)=\arg\max_{n}\left[S(C(l,m),B^{n})\right] \tag{12}$$

を計算しテーブルに記憶する。しかし，部分パタンC(l, m)が単語nであるとすると時間変動モデルのiii)の制約により

$$l+J^{n}-\gamma\leqq m\leqq l+J^{n}+\gamma \tag{13}$$

なる関係が成立する。従って(11)，(12)式は，この範囲でのみ計算すればよい。すなわち，C(l, m)とB^{n}の間のDPマッチングは，

初期条件

$$g(l+1,1)=s(c_{l+1},b^{n}_{1}) \tag{14}$$

漸化式

$$g(i,j)=s(c_{i},b^{n}_{j})+\max\begin{bmatrix}g(i-1,j)\\ g(i-1,j-1)\\ g(i-1,j-2)\end{bmatrix} \tag{15}$$

制約条件（整合窓）　$l+j-\gamma\leqq i\leqq l+j+\gamma$ (16)

類似度　$S(C(l,m),B^{n})=g(m,J^{n})$ (17)

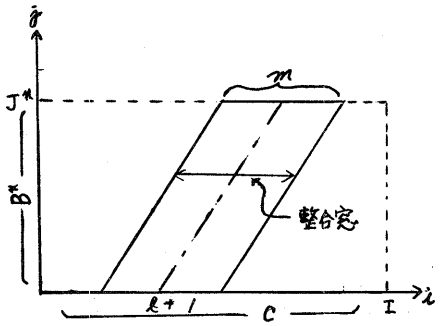


図-1 単語単位のマッチング

のDPによって実行される（図-1）。ここで(15)式の計算がj=J^{n}まで全て終了した時点では，g(m, J^{n})が並列的に，(13)式の範囲で揃っていることになり，(11)式の計算に必要なDPマッチングはN×Iのオーダに低減される。上述のことより，(11)，(12)式が算出される範囲は次の様になる。

$$l+\min_{n}\left[J^{n}\right]-\gamma\leqq m\leqq l+\max_{n}\left[J^{n}\right]+\gamma \tag{18}$$

ii) 全体マッチング　部分類似度テーブルをもとにして，

$$T=\max_{K,l(x)}\sum_{x=1}^{K}\left[\hat{S}(l(x-1),l(x))\right] \tag{19}$$

なる最大問題を計算し，最適な変数K=$\hat{K}$，l(x)=$\hat{l}(x)$ (x=1, 2, …$\hat{K}$)を求める。これは次に示すDPにより実行できる。

初期値　$T(0)=0$ (20)

漸化式　$$T(m)=\max_{l}\left\{\hat{S}(l,m)+T(l)\right\},\quad(m=1,2,\cdots,I)\qquad(21)$$

制約条件　$$m-\max_{n}\left[J^{n}\right]-r\leqq l\leqq m-\min_{n}\left[J^{n}\right]+r\qquad(22)$$

(19)式の最適変数 $\hat{k}$, $\hat{l}(x)$, $(x=1,2,\cdots,\hat{k})$ は(21)式を計算して得られる最適な l を $l(m)$ の形式でテーブルに記憶しておくことにより、$l(I)$ より逆登って計算出来る。

iii) 判定処理　部分判定結果のテーブルを参照して

$$\hat{n}(x)=\hat{N}(\hat{l}(x-1),\hat{l}(x)),\quad(x=1,2,\cdots,\hat{k})\qquad(23)$$

を認識結果とする。

3. 実時間認識システム[5]

本システムの装置構成は図-2に示したものであり、ミニコン(NEAC-M4/n)、DPプロセッサ(DPP)、音声分析器(SPA)、コマンド入力用キーボード、及びCRT-ディスプレイから構成されている。各装置は写真に示す如く、可搬性のある標準ラックに実装されている。写真に於いて、右側のラックにはミニコン、DPP、SPAが実装されており、認識装置として、約30単語の認識を実行できる。左側のラックには増設用メモリー、CRT-ディスプレイ,キーボード、それに、ソフトウェアシステムジェネレーション用のディスクカートリッヂが実装されている。この増設メモリーは標準パタンの種類を増加するために使用し、これにより約80単語まで登録することができる。次に各ブロックの処理概要について述べる。

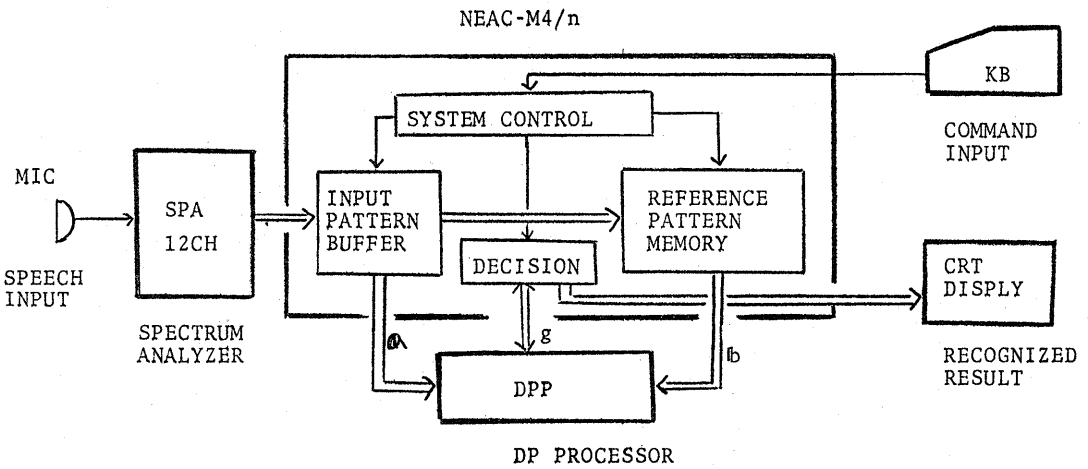
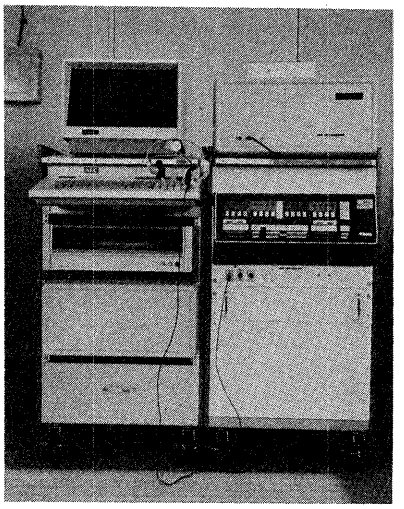


図-2　認識システムの構成

3-1 音声分析器（SPA）

音声分析器はマイクロホンから入力される音声をスペクトラム包絡の時系列パタンに変換する機能を有する。実際には、アクティブバンドパスフィルタと平滑フィルタから成る12個のフィルタバンクにより周波数分析を行う。各フィルタバンクの出力（以下チャンネルと称す）はディジタル化され、更に発声レベルの変動の影響を除くために、各チャンネルの総和により正規化される。正規化された出力は、ある一定の周期（以下ではフレーム周期と称す）で、ミニコンに転送される。1フレームの内容は前節で述べた、特徴ベクトルを表現していることになる。分析器の主な性能を表－1に示す。



写真－1　システムの概観

周波数範囲	200～6,900 Hz
分析チャンネル数	12個
分析フィルタ	75 dB/oct, 3dBでクロスオーバー
直線検波器	0.1～3 KHz, 05dB, 直線性50 dB
LPF	f_c=5, 10, 20, 40Hz, -6dB Bessel
A/D変換器	12bit Straight binary
分析周期	最小 17mS

表－1　SPAの主な性能

3-2 DPプロセッサ（DPP）

前述のDPマッチングを専用に計算するプロセッサであり、前節の(14)～(17)式において、ベクトル間の類似度 $d(C_i, l_{b_j})$ を C_i と l_{b_j} の間の絶対値距離 $d(C_i, l_{b_j})$ として、DPマッチングが実行される。これに伴なって、(15)、(21)式の最大化操作はすべて最小化に置換して実行される。DPマッチングは

$$g(i,j)=d(C_i,l_{b_j})+\min\begin{bmatrix}g(i-2,j-1)+d(C_{i-1},l_{b_j})\\ g(i-1,j-1)\\ g(i-1,j-2)\end{bmatrix} \quad (24)$$

$$d(C_i,l_{b_j})=\sum_{n=1}^{12}|C_{in}-b_{jn}| \quad (25)$$

C_{in}, b_{jn} は各特徴ベクトルの要素である

OP	00	Set Dimension of Vector N
	01	Set No. of Window NW
	02	Initial Mode Change
	03	Control Reset
	04	α Write Start
	05	l_D Write Start
	06	g Write Start
	07	g Read Start
	08	DP Start
	09	DPP Ready

表－2　DPPの命令セット

を、j を一定にし、整合窓内の i に対し、i の増加する方向に順次実行し整合窓内一段を終了すると次に $j=j+1$ として同様に演算を実行し、$j=J''$ まで実行後には(13)の範囲に関する(17)式の類似度がすべて求まることになる。実際には、この動作をカテゴリ数(N)だけ実行して、(18)式の範囲の部分類似度が求まる。DPPの内部構成は(25)式の距離を計算する距離計算部と(24)式の漸化式の処理を実行する漸化式計算部より構成されており、各計算部をパイプライン的に動作させて、DPマッチングを高速に実行している。またDPPの各計算部の動作及びミニコンとのデータ転送は表－2に示す命令セットにより、すべてミニコンにより制御される。

DPPの処理時間は整合窓数、分析チャンネル数によって決定される。整合窓数を15、分析チャンネル数を12とすると、DP計算は1段当り約50μsで実行される。標準パタンの平均長さを20とすると、約1msで標準パタン当りの部分類似度が算出できることになる。実際にはミニコンとの転送時間が加わるので、約2.5ms程度の時間を要している。実時間認識のためには、1フレーム入力毎に、各標準パタンとの間のDPマッチングが全て終了していることが要求される。フレーム周期を20msとし、処理量を低減するために、部分区間(l, m)の変化を2フレーム刻みで行うとして、本システムでの処理時間を算出すると標準パタンの数が約16個程度までは実時間認識が可能である。

3-3 ミニコンの処理内容

ミニコンは音声パタンの読込み、標準パタンの登録、蓄積、部分類似度、分部判定結果の蓄積、全体マッチングの処理、最終判定、結果の表示及びシステム全体の制御等の機能を分担している。以下、主な機能の内容を説明する。

- 音声パタンの読込みとしては、十分長い入力パターンを許容するために、図-3に示すような、有限長の入力パタンバッファをリング状に使用している。入力パタンバッファは周波数方向と時間方向との2次元構成から成り、リング状に動作されるように制御される。(図-3)
- 標準パタン登録に関しては、表示部に発声すべき単語名を表示し、同時に発声許可表示により、発声可能なタイミングを表示する。また、標準パタンに関する各種情報を保持するために、カテゴリコード、格納番地、パタンのフレーム数等の辞書をもっており、登録時に作成され、認識動作時に参照される。
- 実時間動作を可能にするために、入力と認識の同時動作が要求される。この動作を制御するために、入力音声の始端が検出されると同時に、認識動作を行い、認識中の音声(連続発声された幾つかの単語から成り、一呼気で発声される単位；オー音声)が終了した場合には、次に発声される音声(オー音声)の検出や読込みを、認識

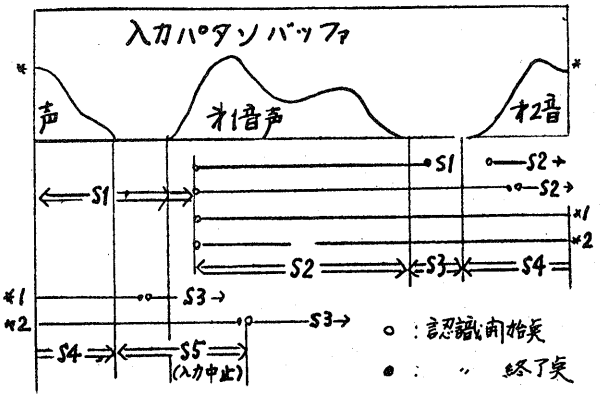


図-3 入力パタンバッファと状態

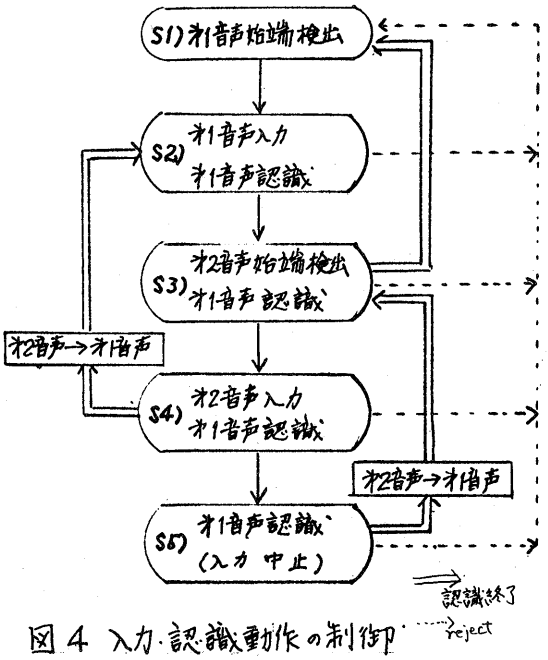


図4 入力・認識動作の制御

と同時に行うことが必要となる。このために本システムでは図-4に示すように、各状態に関しての、有限状態オートマトンによって同時動作を制御している。(図-3) また入力パタンの始点番地等は又段のFIFOメモリーにスタックして順次能率良く処理している。

4. 評価実験および結果

音声認識が実際に利用される種々の形態を考えた場合、その性能を規定するものとしては、

1) 認識率
2) 処理可能な入力速度
3) 対周囲雑音の影響
4) 使い易さ
5) 疲労度

等の要因が考えられる。各要因のうち1)～3)に関しては、ある程度客観的に評価を行なうことができよう。また4), 5)等の評価は人間的な種々の要因が入るため、主観的評価を行うことが必要になる。

ここでは、本認識システムの性能評価の第一歩として、連続数字を対象にした1)～3)に関する実験を行なった。

4-1 入力速度に関する予備実験

入力速度に関し、連続単語発声の有効性を、また連続発声する場合、何桁程度が適当であるかを確認するために、簡単な予備実験を行った。

男性3人により、60個の数字をそれぞれ、1桁づつ区切って発声した場合と、2桁から6桁まで連続に発声した場合について、入力に要する時間を測定した。この際、発声者には何等コメントをよえず、自然に発声してもらった。結果を図-5に示す。

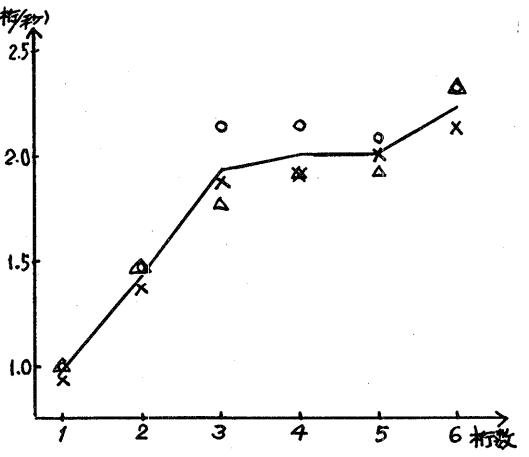


図-5 連続数字桁数と入力速度

図-5から、1桁づつ区切って発声した場合に較べて明らかに連続発声した場合、入力速度が向上していることが判る。更に3桁以上連続発声することにより、約2倍の入力速度向上が得られる。3桁以上連続発声しても、入力速度に大した差異はなく、桁数が多くなると、読みにくくなることから3～4桁の連続発声が、入力速度、発声の仕易さ等から適当であると思われる。

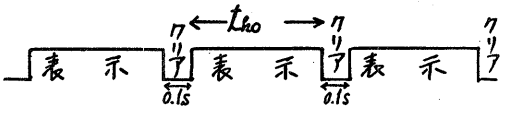


図-6 表示タイミング

4-2 発声速度に関する認識実験

入力速度を向上するためには、速く発声する必要があるが発声速度があ

る程度以上速くなると調音結合や発音のなまりがはなはだしくなり、標準パタンとのずれが生じ、認識率が低下することが予想される。
本システムがどの程度の発声速度まで安定に動作するかを確認するために、種々な発声速度に対する認識実験を行なった。

前述の予備実験より、実験に使用した音声は連続3桁数字を対象にして行なった。発声者としては、本システムを使い慣れている2名を対象とした。発声速度を客観的に測定するために、ディスプレイ装置に発声すべき3桁数字を図-6に示すタイミングにより次々に表示し、発声者は表示される速度に従って順次読み上げる方法でサンプルを集録した。尚、集録は騒音レベル約65dB(A)程度のコンピュータルームにて接話型マイクロホンを使用して行った。発声速度としては、ゆっくりした感じ(t_{h0}=2.3秒)から非常に速く発声した感じ(t_{h0}=1.1秒)までの5段階のものを使用した。実験は各発声速度に対する110サンプル(330桁)を用いて行い、標準パタンは、普通の発声速度の各人、各数字2個使用して行なった。結果を表-3, 図-7, に示す。表-3において( )内は速く発声した標準パタンを使用した場合の結果であり、図-7では、点線で示した。

入力速度 / 発声者	1.25 (桁/秒)	1.43	1.66	2.00	2.50
A	1	1	0	1(0)	4(1)
B	1	2	1	8(2)	14(4)

表-3 入力速度(桁/秒)と誤り数

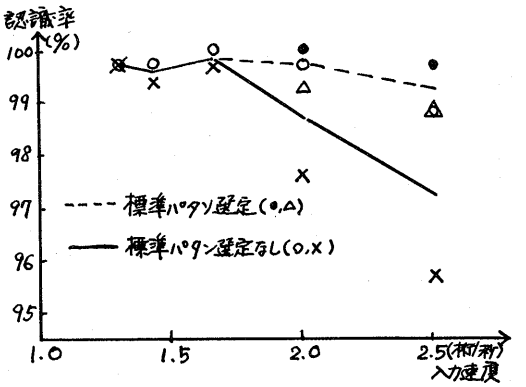


図-7 入力速度と認識率

4-3 周囲雑音に関する認識実験[8]

周囲雑音に対するシステムの性能を評価するために、種々の騒音レベル下におけるサンプルに対して認識実験を行った。実験条件としては、通話実験室に於いて、天井の5つのスピーカから独立に室内雑音発生器を使用し、65dB(A)(静かな状態)から95dB(A)(非常にうるさい状態)の騒音レベルの騒音を発生した。
実験サンプルは、接話型マイクロホンを使用し(口と鼻から約5cm離して固定)、各騒音レベルに対して4-2の実験で使用したものと同じ3桁数字を50個(150桁)発声したものを集録した。尚、発声速度は普通に発声したものを使用した。標準パタンとしては、65dB(A)におい

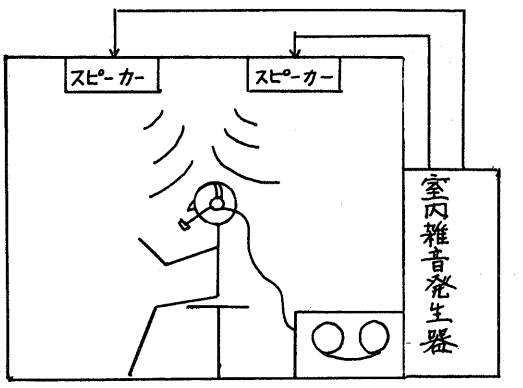


図-8 騒音下でのサンプル集録環境

て発声した各人各数字2仙使用して行った。結果を表-4 図-9に示す。( )内及び点線は、対応する騒音レベル時の数字を標準パタンとして実験した結果である。

発声者 \ dB(A)	65	75	85	95
A	0	0	3(0)	23(0)
B	0	0	2(0)	12(1)

表-4 騒音レベルと誤り数

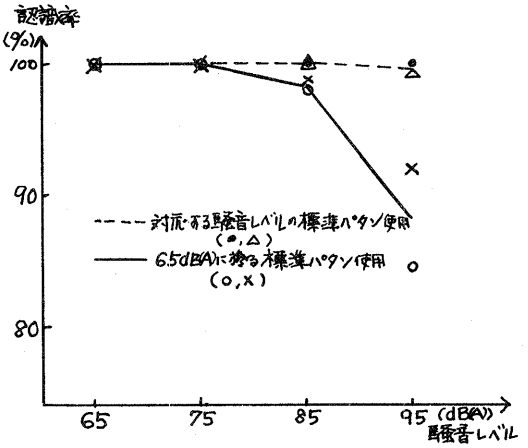


図-9 騒音レベルと認識率

5. 検討

発声速度に関する実験より、約1.9桁/秒程度の発声速度まで99%以上の認識率が得られることが判った。このこは、予備実験により3桁数字を発声した場合の速度が約1.9桁/秒位いであることから、本認識システムは、普通に発声した程度の連続3桁数字であれば、十分に安定に動作することが確認できる。発声速度を更に速めた場合(2.0桁/秒以上)には認識率の低下が見られるが、標準パタンとして、速く発声した時のパタンを使用することにより、低下を防げることが確認できた。これは発声のなまけ等により、標準パタンとのずれが生じていることが原因と考えられ、今後の問題として、標準パタンの選定法を更に工夫する必要がある。対周囲雑音についての実験では、静かな状態(約65dB(A))で発声したものを標準パタンとして使用した場合、約85dB(A)までの騒音レベルに対して、認識率の低下は、ほとんどないことが確認できた。また95dB(A)程度になると認識率は格段に悪くなるが、この場合でも、標準パタンとして、同じ騒音下で発声されたものを使用すれば、認識率の低下を防げることが確認できた。しかし、本実験において使用した騒音は定常的なものであり、実際には、瞬時的に非常に高いレベルの騒音も存在する訳で、この様な騒音に対する評価をする必要がある。

6. むすび

以上、2段DPマッチング法に基づいた連続単語認識システムの概要とこのシステムを用いた、発声速度及び周囲雑音に対する評価実験の結果を述べた。発声速度に関しては、標準パタンを適当に選定することにより、相当速く発声した場合でも99.3%の認識率が得られた。更に、周囲雑音に対しても、定常的な騒音であれば95dB(A)程度の騒音レベルまで安定に認識動作することが確認された。

今後、使い易さ、あるいは疲労度といった人間工学的な評価を進めていきたいと考えている。

謝辞

終りに、日頃御指導をいただく当研究所通信研究部金子部長、並びに本研究を進めるに当り、適切なる御指導を賜った加藤部長代理、騒音の実験に関し御助言をいただいた落合主任に深謝致します。

参考文献

(1) 加藤 ; 信学誌、昭49-45〔技術展望-3〕vol. 57. No.3 P.315(1974)

(2) 迫江、千葉 ; 音学誌 Vol.27. No.9 P.483 (1971)

(3) 千葉 ; 昭和48連大予稿 325 (1973)

(4) 鶴田、迫江、千葉 ; 昭49. 信学全大予稿 1596(1974-8)

(5) 鶴田、迫江、千葉 ; 音声研究会資料 S74-30 (1974-12)

(6) 迫江 ; 音学大会予稿 2-2-13 (1975-5)

(7) 迫江 ; 音声研究会資料 S75-28 (1975-11)

(8) R. B. NEELY, D.R. REDDY ; 7th ICA, 23-C-14, P177 (1971)