

音声による日本語テキスト入力

白井克彦 深沢良彰 松浦 博 久保田淳市 小林哲則
(早稲田大学 理工学部 電気工学科)

I. はじめに

音声理解システムは、1970年代になってから、米国を中心に各地で活発な研究が進められている。これらは音声の内容を理解しようとするものであり、そのため種々の知識や技法の有効性が確かめられてきている。しかし、これら多くはタスクに依存したシステムであり、実用性に関しては多くの問題点を残している。我々は補助入力装置を使用することにより、タスクに依存しない、音声による日本語テキスト入力システムを試作している。

本システムの構成としては、その処理時間や必要とする記憶域の量を考え、音韻識別部と単語識別部とからなる。しかもこれらは互いに他方の特徴を生かし、欠点を補うように設計されている。

音韻認識に関しては、種々の研究機関から数多くの有意義な研究結果が報告されている。特に単語音声の認識においては極めて高い認識率が得られており、製品化も進んできている。しかし、連続音声の認識においては、特微量の個人差、調音結合、発話ごとのばらつき等に多くの問題が残されている。

我々は調音結合の処理が実際の音声発生過程と最も明らかに対応し、話者適応も他の方法に比べて容易である点に注目し、調音モデルに基づく特微量抽出法を用いて音韻識別を行なっている。即ち発声のための調音運動を音声波から推定し、その軌跡を用いて音韻識別を行なっている。しかし音声信号の解析においては、本質的に曖昧さが存在するので、音韻識別の結果は音韻ラティスとして、単語識別部へ渡される。

単語識別部では種々の先駆的な情報を用いて、音韻ラティスから最も適した音韻列を選び、適切な仮名-漢字変換を行なうことが必要である。本システムでは音韻識別部の特性を考慮し、ダイナミック・プログラミングと音韻遷移マトリックスによる類似度計算と、母音列によって索引付けされた単語辞書の検索を中心としている。また、補助的に助詞・助動詞の性質、活用等も利用している。この際、処理をより正確で、容易なものとするために簡単なキー・ボードを用いている。

本発表では、まず調音モデルを用いた特微量抽出法について述べる。そのために連続音声のセグメンテーションと、母音と子音に関する識別方法を示す。第Ⅲ章では系統的に作られた辞書と各種の知識を使った単語識別に関する1方法について述べる。第Ⅳ章では現在までに得られた結果について簡単な議論を行なう結論とする。

II. 音韻識別部

2.1. セグメンテーション

連続音声の認識においては、まず音声波から音声区間を抽出すること、有声・無声の基本的特徴を見出すことが有効である。このために(1)信号の全パワー、(2)信号の低域パワーと全パワーとの比、(3)信号の1フレームの零交差数、(4)正規化した自己相関係数、の4つを特微量として用いている。これらの特微量

は各フレーム毎に計算される。

識別は次の2次識別関数を使って行なっている。

$$d_i(x) = (x - \mu_i)^T C_i^{-1} (x - \mu_i) + \ln |C_i| \quad (2.1)$$

この後、過渡的な個所での誤りや雑音部分を修正するためのスムージング処理を行なう。この結果の1例をFig.2.1に示す。

2.2. 調音モデル

音声波から調音運動が十分に推定されれば、これを音声のすぐれた特徴量として用いることができ、認識の際にも有用である。この節では、この方法の基礎とよっている調音モデルについて述べる。このモデルは実データの統計的な解析に基づいており、少數のパラメータで、各調音器官の位置・形状を正確に表現することができます。さらに、生理的な制約や音声学的な制約も自動的にモデルの中に含まれている。

このモデルの全体の構成図をFig.2.2に示す。あごは固定的下を中心とした一定半径 R_f の回転運動で近似し、下あごの位置 J は、あごの開き角度 X_J で与えられる。唇の形状は、あごの開きに依存した高さ L_R 、横幅 L_w 、突き出し量 L_p で表現される。舌面形状は下あごに固定された半極座標系で指定される。これらのパラメータはX線写真によるデータの統計的分析から得られる。

主成分分析により、母音の舌面形状 \bar{V}_f は次のように線形化式で表わされる。

$$\bar{V}_f = \sum_{i=1}^p X_{pi} V_i + \bar{V}_f \quad (2.2)$$

ここで、 X_{pi} ($i = 1, 2, \dots, p$) は舌の調音パラメータであり、 \bar{V}_f はほぼ舌面の中立形状に対応するベクトルである。

また、固有ベクトル V_i は次式で求められる。

$$C_f \cdot V_i = \lambda_i \cdot V_i \quad (i=1, 2, \dots, p) \quad (2.3)$$

但し、

$$C_f = \frac{1}{N} \sum_{k=1}^N \{(r_{v_k} - \bar{r}_f)(r_{v_k} - \bar{r}_f)^T\} \quad (2.4)$$

であり、 λ_i は次の固有方程式を満たす固有値である。

$$|C_f - \lambda I| = 0 \quad (2.5)$$

上記の統計技法を唇の形状について用いると

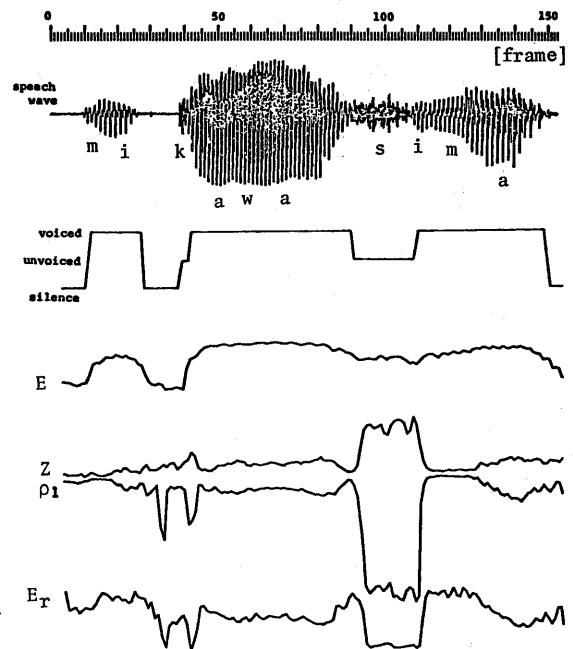


Fig. 2-1 Example of analysis /mikawasima/

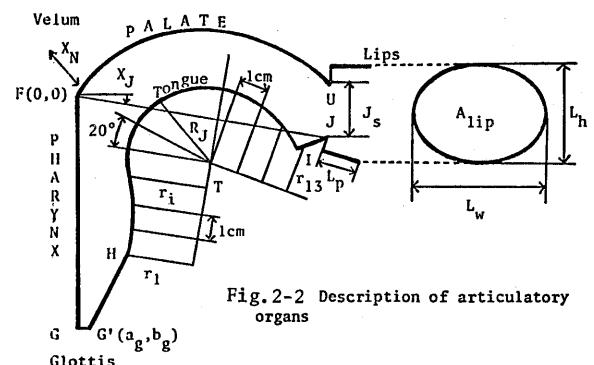


Fig. 2-2 Description of articulatory organs

$$\begin{bmatrix} L_R \\ L_W \\ L_P \end{bmatrix} = \begin{bmatrix} K_1 J_S + \bar{L}_R \\ K_2 J_S^{k_3} + \bar{L}_W \\ \bar{L}_P \end{bmatrix} + \begin{bmatrix} l_R \\ l_W \\ l_P \end{bmatrix} X_L \quad (2.6)$$

を得る。但し、 J_S は下あごの開き幅である。唇の能動的な変形成分は第2項で表わされる。 X_L は唇のパラメータである。

鼻腔を含めた全体のモデルをFig. 2.3に示し、調音パラメータの性質をTable 2.1に要約する。詳細は参考文献に記されている。

2.3. 調音運動の推定

この節では、音声波からの調音パラメータの推定を音声生成モデルの構成において考える。

音声生成モデルにおいて、音響パラメータ \hat{x} は、調音パラメータ x の非線形関数 $\hat{x}(x)$ として表わされる。ここで、音声波から測定された音響パラメータを \hat{x} とすると、調音パラメータの推定値 \hat{x} は次の評価関数を最小化するものとして得られる。

$$J(\hat{x}) = \|Y_s - \hat{y}(\hat{x})\|_Q^2 + \|X_k - \hat{X}\|_R^2 + \|X_{k-1} - \hat{X}\|_P^2 \quad (2.7)$$

ここで、 $\|X\|_R^2$ などは2次形式 $X^T R X$ を表わし、 P, Q, R は重み行列、 k はフレーム番号、 \hat{X}_{k-1} は前フレームの推定値である。この評価関数において、第1項はモデルの音響パラメータと測定値との間の重みづけられた誤差を表わす。第2項は調音パラメータの相補的な効果によって推定値が一定の領域外の値をとるのを避けるため、第3項は連続音声における調音パラメータの連続性に関する項である。

この方法を実音声に適用する場合、話者の調音構造とモデルとの間の違いを調整する必要がある。ここでは音声波より声道伸縮比を推定し、調音モデルの調整を行なった。次ページのFig. 2.4は/a i u e o/に対する推定結果である。

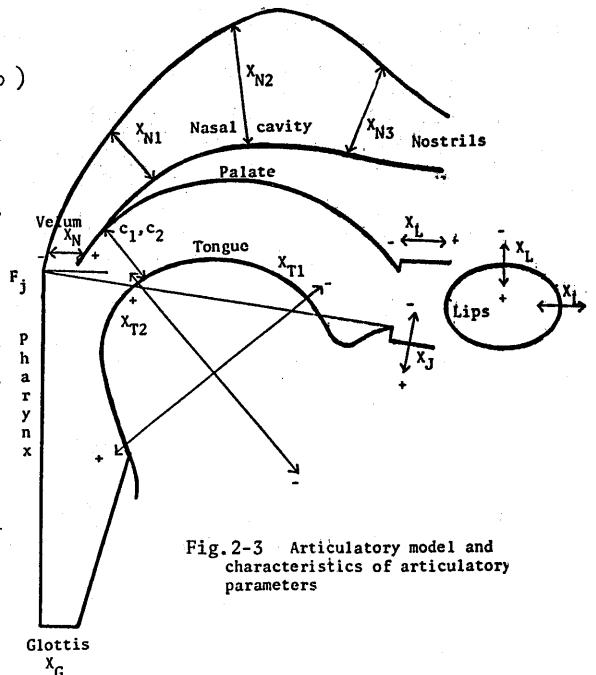


Fig. 2-3 Articulatory model and characteristics of articulatory parameters

Table 2-1 Articulatory parameters

Articulatory parameters	Tongue	Tongue	Jaw	Lip	Glottis	Velum
	X_{T1}	X_{T2}	X_J	X_L	X_G	X_N
+	back	high	open	round	open	open
-	front	low	close	spread	close	close

2. 調音パラメータによる連続母音の識別

ます、例として5連母音 /aieuo/ に関して推定されに調音パラメータの軌跡を示す。(Fig.2.5)

(2. 1)式の2次識別関数を用いて、各フレームごとに母音を識別する。この結果は過渡的な部分を含んでいるので、これらを除去するために次の⑤(8)を用いる。

$$S(k) = |X_{T_1}(k+1) - X_{T_1}(k)| + |X_{T_2}(k+1) - X_{T_2}(k)| + |X_J(k+1) - X_J(k)| + |X_L(k+1) - X_L(k)| \quad (2.8)$$

これにより、 $S(k)$ が谷となるところを定常部、ある閾値より高い山の部分を過渡部とみなすことができる。しかし、一定閾値では定常部と過渡部を十分に区別できない部分があるので、実際には、極大値を考慮しながら定常部を決定する。

音声データとして、成人男性2名が
5連母音のなかで2種を各々2回ず
つ発話し下ものを用いて識別実験を行

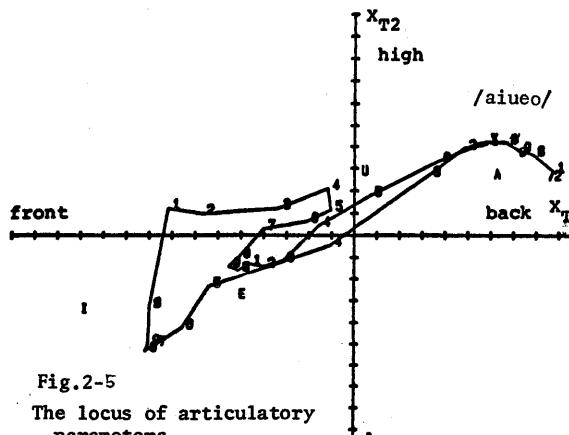


Fig. 2-5

The locus of articulatory parameters

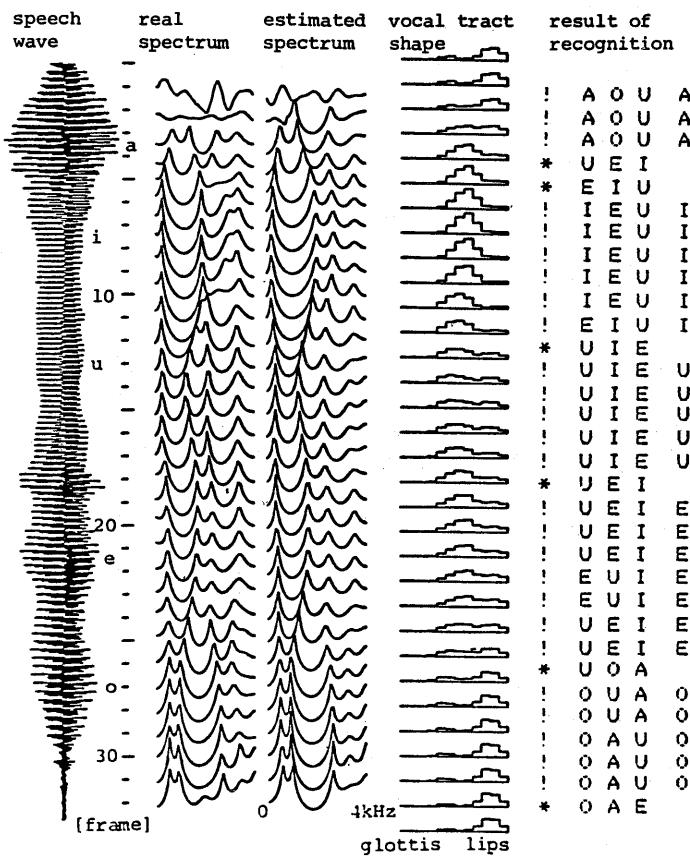


Fig.2.4 Estimated vocal tract shapes and results of recognition

Table 2-2 Vowel discrimination errors

speaker	S.M.		T.K.	
	No. 1	No. 2	No. 3	No. 4
1 /auiueo/				
2 /aeiou/				
3 /aeuoui/				
4 /iaoiu/		5iou--ouu	16 ao--o	26 ao--a
5 /iuoea/		6uo--u		
6 /ieauu/				7ou--auu
7 /uaieo/				27
8 /uaoei/	1ei--ii	7ao--o	17 uao--uoo	28
9 /ueaio/		8ei--ii	18 ei--i	uao--uou
10 /uoaoe/			19 eai--eei	29 eai--eei
11 /euiae/		9a--o		
12 /euicoa/	2e--i		20 uiuo--ueo	
13 /eoaui/				
14 /eoueui/		10 ei--ii	21 ei--i	
15 /oiaue/				
16 /oieua/		1ie--ii		
17 /ouiae/		12 ou--eo	22 iae--iee	
18 /auoua/			23 auo--ao	
19 /uieia/		13 uaui--ouu	24 uie--ue	
20 /eauau/		14 e--i		30 eau--eeu
21 /eoioe/	3ioe--iue			ioe--iue
22 /oaiia/	4iai--iei			31
23 /oaeue/		5ueu--uiuu	25 iai--iei	

なった。この例として用いた5連母音は、前後の異なる3連母音60種をすべて含んでいる。この際生じた誤りをTable 2-2に示した。

2. 調音パラメータによる子音の識別

子音に関する調音パラメータによる子音の識別では、子音と母音との間の渡りの区間のデータを用いて識別している。即ち渡りの区間にについて推定した調音パラメータに関するDPマッチングの手法と簡単な音響的特徴量を用いている。

子音から母音へは調音器官が比較的速く動くため、ハミング窓12msという短時間の分析フレームを用いているが、この際に現われるピッチ構造の影響を抑えるために、ピッチ同期の手法を併用している。調音パラメータの推定は、安定な母音側で、区分線形推定法を用いて初期値を設定し、ここから時間的に逆方向へと、行なっていく。その後、標準パターンとの終端フリーのDPマッチングを行なう。この時、波形の全パワーと高周波パワーも識別のためによく用いている。但し、有声子音の識別については、調音パラメータのDPマッチングの結果を主に用いているが、無声子音については、音響パラメータの方を中心にして行なっている。

2名の成人男性話者によるV₁CV₂型の

データを用いた識別実験の結果をTable 2-3に示す。この実験では、V₁は/a/に固定し、V₂は5母音すべてを用い、Cは子音/g/, /z/, /d/, /b/, /h/, /s/, /p/, /t/, /k/について行なっている。

この方法による子音の識別については、調音パラメータの推定の精度と安定性に若干の問題があり、また個人差の影響も大きい。これらは今後の課題であるが音響的特徴量の選択と組み込み方法を考慮することによって改善が見込まれている。

III. 単語識別

3.1 日本語テキストの入力方法

任意に発音されたテキストが完全に認識され、適切な仮名-漢字変換が行なわれるのが理想である。しかし、完全な音韻識別は不可能であり、また、分かれ書きされていない日本語仮名文の仮名-漢字変換も単語の切り出し、同音異義語の解析等に多くの困難な問題を持っている。そこで本システムでは補助入力装置として簡単なキー・ボードを使い、また音声入力の簡易性を利用して冗長な入力をすることによって、これらの問題を解決しようとしている。

入力の際には、まず入力する文字の種類を示すキーを押す。たとえば、キー(H)

Table 2-3 Results of the recognition experiment

in	out	/ga/	/za/	/da/	/ba/
/ga/	20	0	0	0	0
/za/	2	16	2	0	0
/da/	0	0	20	0	0
/ba/	0	0	0	20	0

in	out	/ka/	/ta/	/da/	/pa/	/ta/
/ka/	20	0	0	0	0	0
/ta/	1	17	2	0	0	0
/da/	6	0	13	1	0	0
/pa/	3	0	1	15	1	0
/ta/	0	1	0	0	0	19

in	out	/gi/	/zi/	/di/
/gi/	18	1	1	1
/zi/	0	20	0	0
/di/	0	1	19	0

in	out	/ki/	/hi/	/chi/	/pi/	/ri/
/ki/	13	0	2	0	5	0
/hi/	0	20	0	0	0	0
/chi/	0	0	20	0	0	0
/pi/	0	5	0	20	0	0
/ri/	5	0	0	0	0	15

in	out	/gu/	/zu/	/bu/
/gu/	19	1	9	0
/zu/	2	18	0	0
/bu/	0	0	20	0

in	out	/au/	/eu/	/ou/	/pu/	/ru/
/au/	19	0	0	1	0	0
/eu/	0	18	1	1	0	0
/ou/	1	9	10	0	0	0
/pu/	1	0	0	18	1	0
/ru/	0	0	0	1	19	0

in	out	/ga/	/za/	/da/	/ba/
/ga/	20	0	0	0	0
/za/	0	18	2	0	0
/da/	0	1	19	0	0
/ba/	0	0	1	19	0

in	out	/m/	/n/	/em/	/en/	/em/
/m/	19	0	0	0	0	1
/n/	0	20	0	0	0	0
/em/	0	2	14	4	0	0
/en/	0	0	1	16	3	0
/em/	0	0	0	2	18	0

voiced : 91.4 %

unvoiced : 85.4 %

はひらがなを表わす。このためのキー・ボードは7ヶのキーを持つのみであり、片手で十分操作可能である。続いてマイクに向い入力したい語を以下の規則に従い発音していく。場合によては、1文字に対する2通り以上の冗長な発声をすることがあるが、音声入力は他の方法に比して、大変容易であるので、ユーザに対しては大きな負担にはならない。

漢字に関しては、音読み・訓読みのほかに、漢字2字熟語を単位として入力することができる。(現在では、これらを区別するためのキーも押さねばならない。) 日本語では(漢字)+(ひらがな)という形式をとるもののが非常に多いので、これらはそのまま入力することができます。このために送りがなや活用、助詞・助動詞を分けて入力する必要はない。それ以外のひらがな、カタカナ、および特殊記号については、キー・ボードのキーとともに、通常のように日本語テキストを発音していけばよい。

3.2 入力データ

単語識別部に対する入力データの構成要素をFig 3.1に示す。もし音韻識別部が音韻の候補を1つに定めることができない場合には、ラティスとして、各候補についての音韻とその信頼度をわざす。即ち入力データは音声入力に対応する日本語擬似音韻列である。Fig 3.2は入力データの例であり、各音韻は、実際にはFig 3.3のようなラティスとなっている。

制御部分は音韻データの中から適したものを見出し、効率良い仮名-漢字変換を行なうために、キー・ボードから入力されるものである。

3.3 単語識別アルゴリズム

本システムでは原則として字種キー・ボードから与えられる。そこで全体のアルゴリズムの概略はFig 3.4のように分割される。

漢字として出力されるものに関しては、熟語・音読み・訓読みなどの冗長なデータが与えられることがある。これらの各々に対応する漢字の候補をFig 3.5のアルゴリズムを用いて求めめる。このアルゴリズムは母音が子音よりも高い音韻識別

率を持つことのある音韻がどの音韻と誤認識されやすいか(C J G8 I8 Z8 I7 U8 P5T5 U8 T GIJUTSU)

という音韻識別部の特性が、(Y A9 Z9 A9 T WAZA G7 ISE5 T GI)(Y Z8 I8 C8 U5E2 T JUTSU)

あらかじめわかっていること(H Y N5M5 08 T NO)

に基づいている。

使用されている漢字辞書は

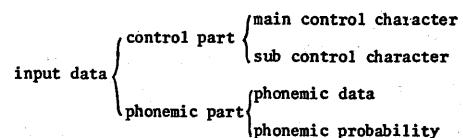


Fig 3.1 the component of input data to the word discrimination part

(原文)

こうした技術の進歩は働く……

H Y (koushita) T KOUSHITA
 CH Y (gijutsuno) T GIJUTSU-NO
 J (gijutsu)
 Y (gi)* D (waza)
 Y (jutsu)
 CH Y (shinpowa) T SHINPO-WA
 J (shinpo)
 Y (shin) D (susumu)
 Y (ho) D (aruku)
 CH Y (hataraku) T HATARA-KU
 Y (hataraku) D (dou)

Fig 3.2 入力例

Fig 3.3 音韻ラティス

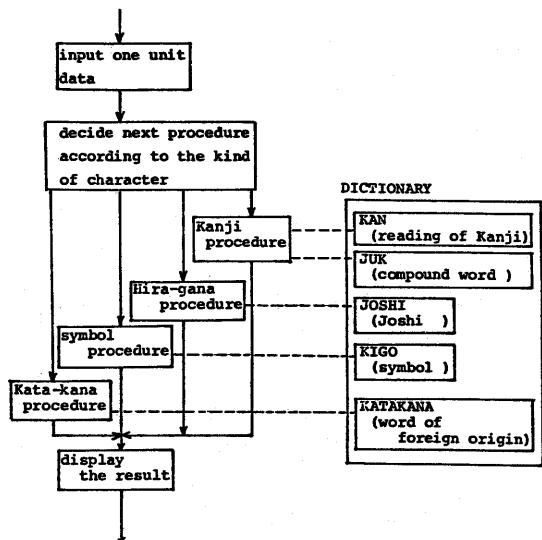


Fig 3.4 単語識別アルゴリズムの概略

どの辞書の現在の大きさを Fig 3.6 に示す。このなかで漢字辞書はその漢字に含まれる母音の数によってまず分けられ、さらに母音のならびによって索引付けされている。この一部を Fig 3.7 に示す。

識別された音韻列と漢字辞書または熟語辞書中に含まれる単語との類似度は遷移マトリックスを用いながら、ダイナミックプログラミングによって計算される。この遷移マトリックスは音韻識別実験の結果を用いて作られ、音韻識別部の特性を十分反映している。このマトリックスには、他の音韻との誤認識率、その音韻の脱落率、付加率が記入されている。特に付加ドットについては音韻識別部の性質から次の音韻の影響を強く受けることがわかつていて、後続の音韻との関係において記されている。

ダイナミック・プログラミングにおいては、長さ I の音韻列 D と長さ J の音韻列 W との間の類似度 $S(D, W)$ を次

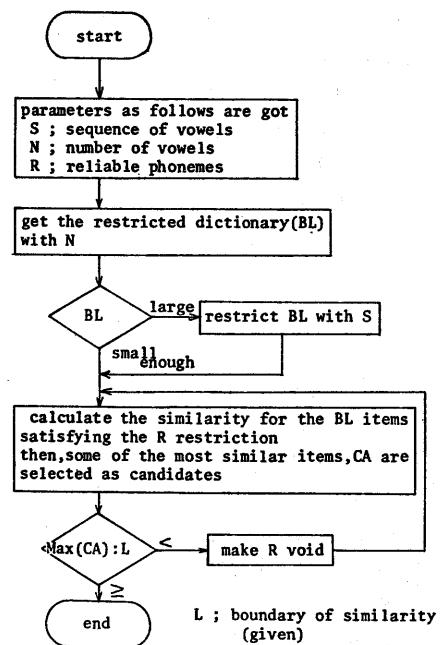


Fig 3-5 Algorithm to decide some candidates

Fig 3.6 Number of dictionary items

dictionary number of vowels	JUK	KANJI	KATAKANA	KIGO
1	1	114	2	7
2	104	389	86	38
3	270	254	164	3
4	141	119	117	1
5	-	14	37	-
6	-	2	14	-
7	-	-	3	-
8	-	-	1	-
total	516	882	424	49

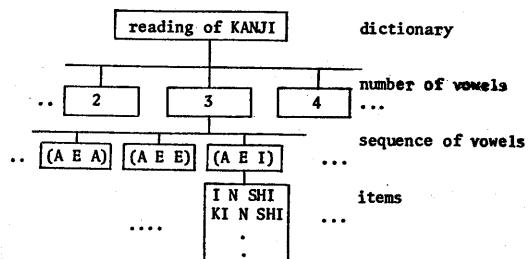


Fig 3-7 Example of a dictionary

のように計算している。

$$\begin{aligned}
 S(D, W) &= g(1, 1) / I \\
 g(i, j) &= \max \{ L(i, j), La(i, j), Lo(i, j) \} \\
 L(i, j) &= sim(i, j) + g(i+1, j+1) \\
 La(i, j) &= sim(i, j+1) + g(i+1, j+2) + adp(i, j) \\
 Lo(i, j) &= sim(i+1, j) + g(i+2, j+1) + omp(j) \\
 g(I, J) &= \sup(J - j) \\
 g(i, J) &= \sup(I - i)
 \end{aligned} \tag{3.1}$$

ここで $sim(i, j)$, $adp(i, j)$ および $omp(j)$ は遷移マトリックスによって与えられ、各々音韻 i と音韻 j との間の類似度、音韻 j の前に音韻 i が付加する確率、音韻 i が脱落する確率を表わす。また $\sup(i)$ は長さが一致していない時のペナルティの値である。

本アルゴリズムによって高いマッチング率を持つ漢字をいくつか選べ出し、冗長なデータが与えられている場合にはそれらも勘案し、最も適当な漢字を選択し、これを出力する。

日本語では（漢字）+（ひらがな）という組み合わせが非常に多い。これらは
 (i) (漢字) + (活用語尾)
 (ii) (漢字) + (送りがな)
 (iii) (漢字) + (助詞または助動詞)
 の3種に大別される。

(i) に関しては用言の語幹として使われる漢字に対しては、活用の種類を漢字辞書または熟語辞書に記入しておくことによって解決できる。(ii) に関しても、体言中に使われるものに対しては送りがなも辞書に登録しておき、漢字を決定した後に送りがなを分離している。(iii) については助詞・助動詞をその接続法とともに辞書として登録しておき、これらと優先的に比較することによって識別率の向上をはかっている。

一般にひらがなについては入力として1通りの読みしか与えられない。特に本システムの音韻識別部は、長い音韻列よりも短い音韻列の方が識別率が低い。これを補うためにも上記の組み合わせの処理は有効である。

カタカナおよび特殊記号に関しては、比較的数が少ないので、別に用意してある外来語の辞書と記号の読み方の辞書を用いて、最も類似した項目を選んでいる。

Fig 3.2 の例を処理した結果の一部を途中結果とともに Fig 3.8 に示す。

DIC-ITEM=YAWARAGU	SIMILARITY=5
DIC-ITEM=HATARAKU	SIMILARITY=7
DIC-ITEM=TATAKAU	SIMILARITY=6
DIC-ITEM=TAGAYASU	SIMILARITY=5
DIC-ITEM=SAKARAU	SIMILARITY=6
DIC-ITEM=KANARAZU	SIMILARITY=5
DIC-ITEM=ARAWASU	SIMILARITY=5
DIC-ITEM=AYAMARU	SIMILARITY=5
DIC-ITEM=WAKARERU	SIMILARITY=5
DIC-ITEM=MAKASERU	SIMILARITY=5
DIC-ITEM=NAGARERU	SIMILARITY=5
DIC-ITEM=SADAMERU	SIMILARITY=5
DIC-ITEM=SAKAERU	SIMILARITY=6
DIC-ITEM=KAXGAERU	SIMILARITY=5
DIC-ITEM=KAMAERU	SIMILARITY=5
DIC-ITEM=KATAMERU	SIMILARITY=5
DIC-ITEM=KASANERU	SIMILARITY=5
DIC-ITEM=ABARERU	SIMILARITY=6

SEIKAI=((HATARAKU))
 CODE=(3815)

Fig 3.8 処理結果

IV. おわりに

現システムは音韻データの抽出のために YHP 21 MX を用い、音韻識別部の他の部分と単語識別部は東大大型機センタの M 200 H 上に作られている。全システムを通してパフォーマンス、機能、ボトルネック等は、まだ実験中であるが各節で示したようなサブ・システムごとのいくつかの結果や問題点が明らかになってきている。

音韻識別部は不特定話者による連續音声を認識することを目的としている。そのため、調音次元における特徴抽出を行なっている。その結果として、安定した調音運動が満足すべき精度で推定されており、母音に関してはその有効性は明らかに確かめられている。また話者に対する適応も容易に行なえる。子音に関しては、調音運動の渡りの部分から調音パラメータを推定している。しかし、現在まだその精度と安定性および話者ごとのばらつきなどにいくつかの問題点が残されている。

単語識別部では、現在、冗長な音声入力とキー入力に対して、音韻ごとの遷移マトリックスを利用したダイナミック・プログラミングによる類似度計算を主体としている。日本語の性質としては、活用、助詞・助動詞など比較的利用しやすいものを使っていているのみである。今後、冗長な入力やキー入力が餘々に不要となってくるようには、種々の構文情報、意味情報などを組み込んでいく予定である。また、ある種の学習機能の必要性を感じている。しかし、本システムはオンラインで動くことを目標にしており、これらの拡張による機能と処理速度との間のトレード・オフは考慮すべき重要な点であろう。

参考文献

- (1) Lea, W. A. Ed. "Trends in Speech Recognition" Prentice-Hall (1980)
- (2) Newell, D. R. et al. "Speech Understanding System: Final Report of a Study Group" North-Holland (1973)
- (3) 白井, 誉田 '音声波からの調音パラメータの推定' 電子通信学会論文誌 Vol. 61-A No. 5 (1978)
- (4) Shirai, K. 'Feature Extraction and Sentence Recognition Algorithm in Speech Input System' 4th IJCAI (1975)
- (5) 白井, 誉田, 五味 '調音モデルの正規化と適応化' 音響学会誌 Vol. 33 No. 6 (1977)
- (6) 坂井, 中川 '不特定話者・連續音声向き単語音声の識別' 情報処理学会誌 Vol. 17 No. 7 (1976)
- (7) Shirai, K., Fukazawa, Y., Matsui, T., Matsuura, H. 'A Trial of Japanese Text Input System Using Speech Recognition' Proc. COLING 80 (1980)