T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	単音節入力による大語い単語音声認識
Title	
著者(和文)	古井貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会 音声研究会資料 S81-62, , , pp. 493-500
Journal/Book name	, , , pp. 493-500
発行日 / Issue date	1981, 12

Powered by T2R2 (Science Tokyo Research Repository)

単音節入力による大語い単語音声認識

古井 貞熙 （日本電信電話公社 武蔵野電気通信研究所）

昭和56年12月22日 於 武蔵野電気通信研究所

内容梗概 ケプストラムおよびパワー時系列の全音声区間における大局的特徴（フーリエ展開係数）と語頭部の局所的特徴（時系列の時間軸シフトマッチング）の組合せによる単音節認識方法を提案し，日本語68単音節および101単音節の認識実験により，この方法が，従来試みられた他の方法よりも少ない処理量で，高い性能を有することを確認した．さらに，この方法を用いれば，100音節／分程度の速度で音節単位に発声した数千～1万程度の大語いの単語音声に対して，高い単語認識率が得られることが確認された．発声速度と認識率の関係，単語中の音節数（モーラ数）の自動決定，促音（っ）の自動検出，標準パタンの適応化等に関しても検討を行い，音節認識率と単語認識率の関係については，理論値と実験結果がよく対応することを明らかにした．

A New Technique for Spoken Japanese Syllable Recognition and Its Application to Large Vocabulary Word Recognition

Sadaoki FURUI
(Musashino Electrical Communication Laboratory, N.T.T.)

December 22, 1981 at Musashino E.C.L.,N.T.T.

Abstract This paper describes a new technique for automatic recognition of spoken Japanese syllables. LPC-cepstrum coefficients and gain are extracted successively throughout an utterance to form time functions. The operation of the system is based on time functions of the initial transitional part of the utterance and Fourier expansion coefficients for time functions of the overall utterance. Results of experiments indicate that syllable recognition accuracy of 93.6% and word recognition accuracy of 96.6% can be obtained, respectively, for 68 Japanese CV-syllables and 12,813 four-syllable words uttered by isolated syllables. Relation between speaking speed and recognition accuracy, automatic decision for the number of syllables in the word utterance, automatic detection of /Q/-moras, and an automatic adaptation technique for the syllable references are also investigated.

Ⅰ まえがき

単音節音声の認識はかなり古くから検討されているが,[1] 単音節は日本語のカナに対応しているので,近年特にワードプロセッサへの入力手段や,名前等の大語いの認識のための方法として注目されている。単音節に区切った発声は,発声者への負担が大きいが,カナタイプライタや,ペンタッチ等による漢字の入力も,一般の人にとって決して易しいことではないので,高い性能の単音節認識系が実現されれば,かなり使われる可能性がある。

これまでに検討された単音節認識方法は,音声波を帯域通過フィルタ群によって分析し,定常部のスペクトルによって母音を認識し,語頭の過渡部のスペクトルの時間伸縮マッチングによって子音を認識するというものが多い。[2][3]他この際,語頭の切り出しの不安定性にどう対処するか,過渡区間に対してどの程度の伸縮を認めるのが適切かというような課題がある。また子音情報は調音結合のために母音部にまでおよんでいるので,語頭部だけで子音を認識するのは必ずしも得策ではない。

筆者は,このような問題に対処する方法として,全音声区間におけるスペクトルの大局的特徴と,語頭部の局所的特徴の組合せによる単音節認識方法を提案し,単音節発声による大語いの認識実験によってこの方法の有効性を確認した。この方法は日本語10数字の認識にも適用でき,[4]英語のアルファベットの認識にも拡張できると考えられる。

Ⅱ 単音節認識アルゴリズム

2.1. 基本原理と音声分析

ここで検討した単音節認識方法の原理を図1に示す。この方法では,母音情報は主としてスペクトルの比較的ゆっくりした変化を表現する大局的特徴として,子音情報は大局的特徴と語頭部の過渡的な特徴の両者からとり出し,これらの特徴を組合せて音節の判定を行う。

音声波はまず,相関法によるLPC分析によって,10ms毎の1～10次のLPCケプストラムと(対数)パワーの時系列に変換される。各フレーム毎のパワーを,あらかじめ平均雑音レベルをもとに決めてあるしきい値と比較して,雑音区間と真の音声区間の分離を行う。このしきい値は2種類設定されており,これらを越える区間長を用いた論理操作により,語頭部の検出誤りを減らす工夫がされている。

2.2. フーリエ展開係数による距離

音声区間と決定された区間の各ケプストラムおよびパワーの時系列は,大局的特徴をとり出すために,次式に従ってフーリエ級数展開される。

$$F_{ij}=\left\{\sum_{t=1}^{L} x_{it}\cos\frac{\pi(t-1)j}{L}\right\}/L \qquad (1)$$

ここで,x_{it}は時刻tのi番目のスペクトルパラメータ(ケプストラムおよびパワー),Lは音声区間のフレーム数,F_{ij}はi番目のパラメータのj次の展開係数である。高次の展開係数は認識への寄与が少ないので,ここではj0

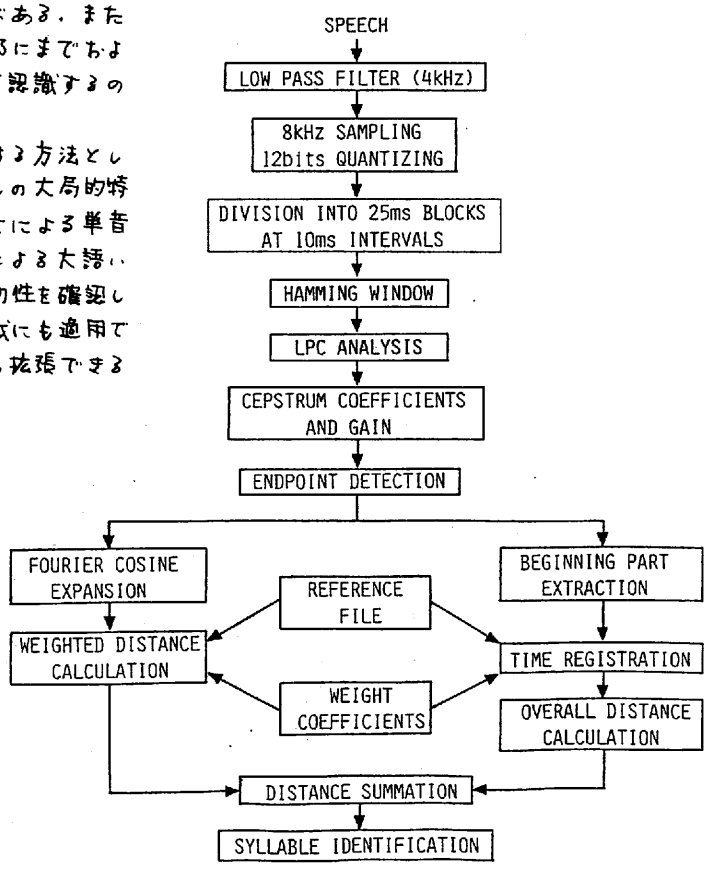

Fig.1-Principal operation of the Japanese syllable recognition system.

次から第6次までの展開のみを行った。スペクトルパラメータは全部で11個あるので，展開係数は計77成分抽出されるが，このうち認識への寄与が比較的大きい41成分だけをとり出し，これを認識に用いる。どの41成分を選択するかは，濁音・半濁音を除く45単音節を対象とした予備実験[5]における，男性5名の音声を用いた分散分析の結果にもとづいて決めた。フーリエ展開係数に関する入力音声と標準パタンの距離は，次の式に従って計算する。

$$d_F(r)=\sum_{i,j\in S}\{F_{ij}-R_{ij}(r)\}^2 w_{ij}^2 \qquad (2)$$

ここで，rは音節の番号，F_{ij}, $R_{ij}(r)$はそれぞれ入力音声と標準パタンのフーリエ展開係数，Sは認識に用いる展開係数（41成分）のインデックスのセットである。w_{ij}は重みで，各展開係数の標準偏差の逆数を用いる。この値も45単音節の分散分析の過程で得られたものを用いた。

2.3. 語頭部のマッチングによる距離

局所的特徴として，語頭の過渡区間の時系列をとり出す。とり出す区間長lは次のような簡単な式によって定めた。

$$l=\mathrm{Min}\{L\times h,\ \mathrm{LMAX}\} \qquad (3)$$

ここで，h，LMAXの値は，実験的にそれぞれ0.33，15（フレーム）とした。

語頭時系列に関する入力音声と標準パタンとの距離としては，次のように時間軸方向に±3フレームまでのシフトを許して，両者の時系列が最もよくマッチングできたときの，両時系列のフレーム当りの平均距離を計算する。

$$d_B(r)=\frac{1}{l}\mathrm{Min}_{0\le k\le 3}\left\{\sum_{t=1}^{l}\sum_{i=1}^{11}(x_{it}-y_{i(t+k)}(r))^2 w_{i0}^2,\ \sum_{t=1}^{l}\sum_{i=1}^{11}(x_{i(t+k)}-y_{it}(r))^2 w_{i0}^2\right\} \qquad (4)$$

ここで，x_{it}, $y_{it}(r)$はそれぞれ入力音声と標準パタンの語頭時系列である。各フレームどうしの距離には，フーリエ展開係数の距離の計算と同様の重みつき距離を用いる。重みには，便宜的に各パラメータの第0次の展開係数の重みを用いた。

2つの時系列の対応づけの方法としては，DP（動的計画法）を用いた時間軸の非線形伸縮による方法[2]が知られているが，ここでは次のような理由でシフトのみによるマッチングを行っている。

i) マッチング処理の簡単化をはかる。

ii) 時間的に短い過渡区間においては，定常区間に比べて発声毎の時間伸縮の度合いが少なく[6]，過大な伸縮を許すと，かえって誤認識が増加する可能性がある。

iii) DPを用いたマッチングにおいては，終端を端点フリーにすることはできるが，始端を完全な端点フリーにすることができず，語頭における厳密なマッチングが必要な単音節認識においては，これでは不十分と考えられる。このために時間軸を逆にたどるDPマッチングを用いる方法も提案されているが，シフトのみであれば，完全な端点フリーマッチングが実現できる。

そこで，このような問題点を調べるため，ここではDPマッチングを行う場合についても検討を行った。

2.4. 総合的距離による認識

上述のようにして，スペクトル時系列の大局的特徴と局所的特徴のそれぞれについて計算された入力音声と標準パタンの距離は，最後に次の式に従って重みつき平均され，この値が最も小さい標準パタンの音節が，認識結果として出力される。この式の中の重み係数αの値は，実験的に最適値に定めた。

$$d_T(r)=\{\alpha\times d_F(r)+d_B(r)\}/(\alpha+1) \qquad (5)$$

III 単音節認識実験

3.1. 実験方法

成人男性10名が計算機端末室において68単音節（/kja/等の拗音節を除く）を5回発声した音声を用いて実験を行った。端末室の平均騒音レベルは約50dB(A)，音声の平均S/N比は20～25dBである。認識実験は各話者毎に行い，各音節について，5回発声したうちの初めの3回を学習用として残りの2回をテスト用とする実験と，後の3回を学習用として初めの2回をテスト用とする実験を行った。従って，各話者の未知サンプルは各音節について4サンプル，全話者の総未知サンプル数は2720サンプルである。

各話者の各音節の標準パタンは，予備実験の結果[4]にもとづき，次のような方法で作成した。すなわち，フーリエ展開係数については全学習サンプルの平均値を標準パタンとし，語頭時系列については学習サンプルのすべてをそのまま

標準パタンとして用いて，Nearest Neighbor 法により距離を計算した。

3.2. 実験結果

フーリエ展開係数または語頭時系列のみの場合，および両者の特徴を組合せた場合（重み係数 $\alpha = 1$）の平均誤認識率を図2に示す。語頭時系列に関しては，シフトのみによるマッチングの場合と，DPマッチングの場合の2通りの結果を示した。ここで用いたDPマッチングの方法は，文献(7)で報告した方法で，DPパスは非対称形であり，1/2と2の間に傾斜制限され，終端と入力音声側の始端は端点フリーとした。

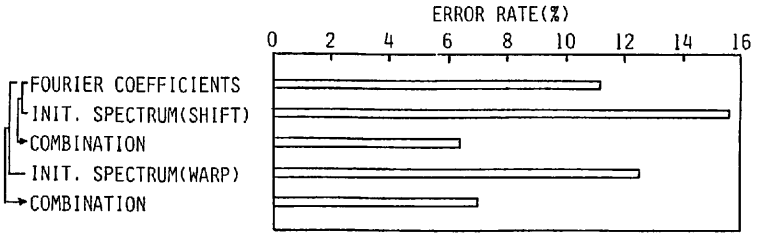

Fig.2-Average recognition error rate for the 68 syllable recognition as a function of the feature parameter set.

この結果から，スペクトル時系列の大局的・局所的両特徴には大きな組合せ効果があることがわかる。語頭時系列の2通りのマッチング方法の結果を比較すると，フーリエ展開係数を組合せない場合は，DPマッチングによる方が誤認識率が小さいが，フーリエ展開係数を組合せた場合は，逆にシフトのみによるマッチングの方がよい。このときの平均誤認識率は6.4％で，この値は一方の特徴のみによる場合の平均誤認識率の約1/2である。

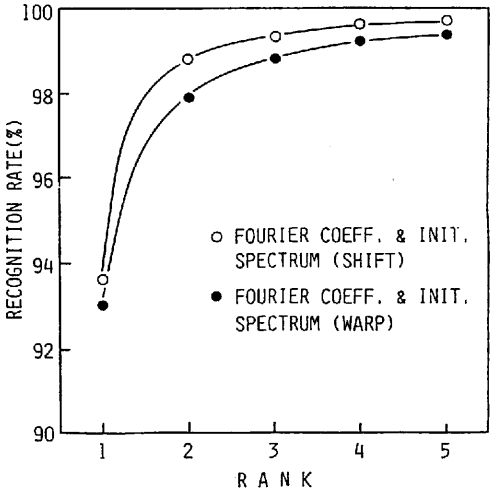

Fig.3-Accumulated recognition rate for the 68 syllable recognition.

図3には，大局的・局所的両特徴の組合せの場合に，正しい音節が第1位～第5位に入る割合いを示した。この結果からも，語頭時系列のマッチングは，DPよりもシフトによるマッチングを行う方が良いことがわかる。後者の方法を用いて第2位の候補まで含めれば，約99％の精度で正しい音節が出力される。このとき第5位までに正しい音節が入る割合いは99.7％である。

表1は，フーリエ展開係数とシフトマッチングの組合せの場合の誤認識行列を示したものである。ここでは/ʃ/は/s/に，/tʃ/,/ts/は/t/に含めている。母音部の誤認識は全くなかったので，すべての誤認識は，この子音別に整理した表に含まれている。サ行とタ行の音節間の誤認識，特に/su/ ↔ /tsu/ の誤認識が比較的多い。

Table 1-Confusion matrix for the 68 syllable recognition.

out \ in	(v)	k	s	t	n	h	m	y	r	w	N	g	z	d	b	p	rate
(v)	191					7										2	95.5%
k	4	183		1		8										4	91.5
s			173	25									1			1	86.5
t	1		24	171		2										2	85.5
n					198		1		1								99.0
h	8	9				181						1				1	90.5
m					2		198										99.0
y								120									100
r			1	1	1				188				2	3	2	2	94.0
w										40							100
N											40						100
g	1	2		1		1						188		1	3	3	94.0
z			2	1					1				192	3		1	96.0
d			1	3					3				2	111			92.5
b	1		1						4			2		2	184	6	92.0
p	3	1		3	4											189	94.5
																average	93.6

図4は，母音部の種類による誤認識率の相違を示したものである。一般に顎の開きの大きい母音を持つ音節ほど，誤認識率が小さい。

ここで，語頭時系列のみによる認識の場合の誤認識の分析を行ったところ，シフトのみによるマッチングの場合は，子音が正しく母音が誤って認識されることがあるが，DPマッチングの場合は母音の誤りはほとんどなく，かわりに子音の誤りが若干ではあるが増加することがわかった。このうち母音の誤りはフーリエ展開係数との組合せによって完全に除去されるので，その結果として，シフトマッチングとフーリエ展開係数の組合せの方が，DPマッチングと後者との組合せよりも良い結果を示す。

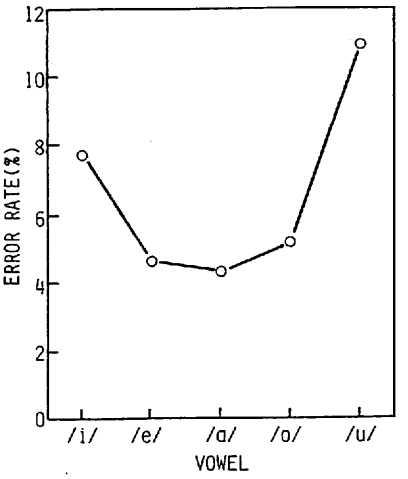

Fig. 4-Recognition error rate as a function of the vowel part of the syllables.

3.3. 発声速度の影響

男性2名が比較的騒音レベルの高い計算機室（70 dB(A)）で，68単音節を種々の速度で発声した音声を用いて，発声速度と認識率との関係を調べた。各音節の学習には，種々の速度で発声したサンプルを用いた。式(5)の重み係数は，大局的・局所的各特徴による68音節標準パタンとの距離の最小値が等しくなるように，自動的に設定した。実験結果は図5に示す通りで，発声速度が毎分100音節程度以下ではほぼ認識率が一定しているが，それより速くなると認識率が急激に低下する。

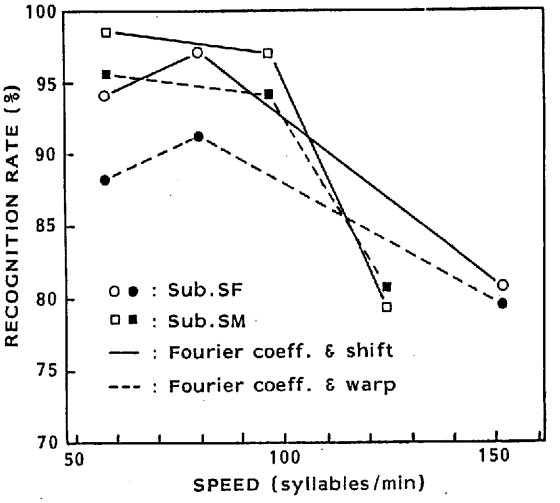

Fig. 5-Recognition rate as a function of the uttering speed.

この結果からも，語頭部のマッチングは，DP*よりもシフトによる方が良いことがわかる。以下の実験では，語頭部に関してはシフトによるマッチングのみを行う。

3.4. 音節候補の予備選択

処理量の低減をはかるため，まずフーリエ展開係数を用いて未知サンプルと各音節との距離を計算し，この値にもとづいて上位の5位または10位までの候補のみについて語頭部の（シフト）マッチングを行う方法について検討した。この方法のブロック図は図6に示す通りで，有効性の評価実験には3.1～3.2節の実験と同じ音声サンプルを用いた。

フーリエ展開係数のみを用いたときに，正しい音節が第1位～第5位に入る割合い（累積認識率）と，10位までで予備選択して語頭部のマッチングを行い，両者の距離を平均化したときの累積認識率を図7に示す。10位までで予備選択したときに正しい音節がその中に入る割合いは99.8%で，語頭マッチング後の累積認識率は，第1位から第5位まで，予備選択を行わない場合（図3）と変わらない。また，5位まで予備選択しても，最終的な第1位の認識率は変わらないことがわかった。このことにより，全体としての計算量を大幅に低下できることが確認された。

3.5. 物音節を含む認識

上記の実験に用いた話者のうちの5名について，物音を含む全101音節の認識実験を行った。実験方法は3.1～3.2節と同じで，実験の結果，物音節と物音節以外との誤認識は全くなく，5

＊ この実験では 端点フリーの対称形DP（文献(11)）を用いた。

名の平均認識率は表2に示す通りであった。この結果からわかるように，本方式を用いれば，対象を101音節に拡張しても，認識率はほとんど低下しない。物音節間の誤認識行列を表3に示す。/hju/↔/kju/，/ʃa/↔/tʃa/，/rjo/↔/gjo/の誤りが比較的多い。

Table 2-Mean recognition rates for the 68 and 101 syllable recognition.

/CV/ SYLLABLES (68)	92.6 %
/CV/ & /CjV/ SYLLABLES (101)	92.5 %

Table 3-Confusion matrix for the /CjV/ syllables.

out / in	y	ky	sy	ty	ny	hy	my	ry	gy	zy	by	py	RATE(%)
y	59					1							98.3
ky		56				4							93.3
sy			57	3									95.0
ty			3	57									95.0
ny					55		4		1				91.7
hy		9				50						1	83.3
my					1		59						98.3
ry								54	4		2		90.0
gy	1							4	53		1	1	88.3
zy				1				1	2	56			93.3
by						1			1	1	56	1	93.3
py						2						58	96.7
												AVERAGE	93.1

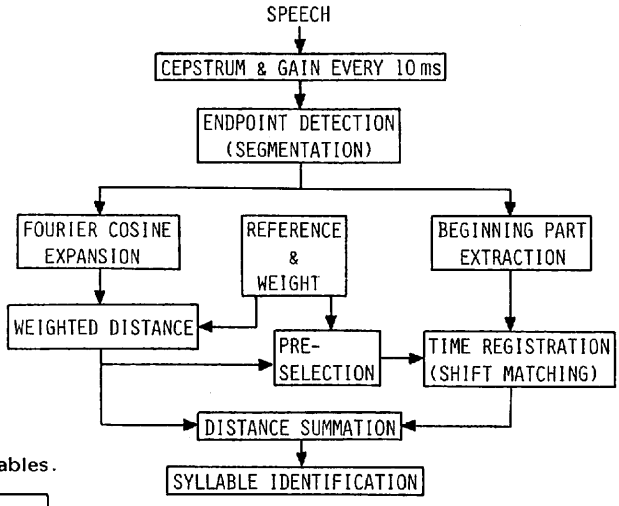

Fig. 6-Block diagram of the improved syllable recognition method.

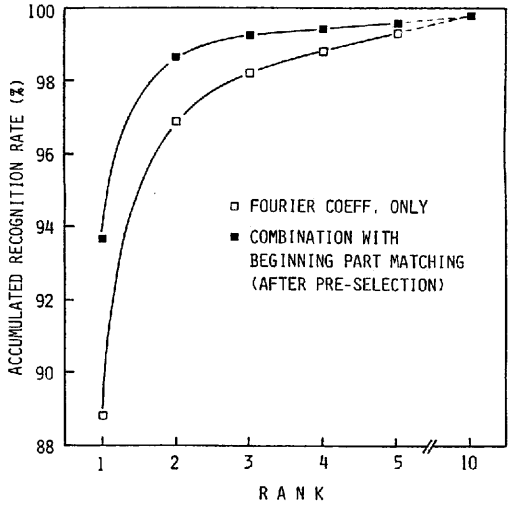

Fig. 7-Accumulated recognition rate for the 68 syllable recognition.

Ⅳ 大語い単語音声認識

4.1. 単語辞書

音節毎に区切って発声された大語いの単語音声認識に関する，理論的・実験的検討を行った。

まず，ある国語辞典に含まれている単語の数をモーラ数別に調べたところ，4モーラの単語が最も多いことがわかったので，4モーラ4文字（小文字を除く）の全名詞から同音異義語を除いた12,813語を認識対象とすることにした。さらに比較のために，ここから4単語毎に選んだ3,203語のセットを作成した。この3,203語中における各音素の出現頻度は，文献(1)に報告されている，対談・小説等の話し言葉に近い5万単語に関する統計と大きな差はない。

4.2. 大語いシミュレーション実験の方法

ここでは，単音節認識率と単語認識率との正確な対応関係を得ることが今後の検討のための基礎資料として重要な意味をもつと考えられるので，新たに単語を発声することはせず，単音節認識実験に用いた音声（68音節）を，各話者毎に各単語の形に並べたものを実験に用いた。各話者の各単音節の未知サンプルは4サンプルずつあるので，これらが公平に用いられるように各単語に割当てた。

音節毎に区切って発声された場合でも，音節間の調音結合がないとは言えないが，ある程度ていねいに区切って発声した場合には，その影響は小さいと考えられるので，このようなやり方をとっても大きな問題はないであろう。具体

的な認識方法としては，入力単語音声の各音節毎に，単語辞書に表記されている音節の標準パタンとの距離を計算し，この総和が最も小さくなる単語を認識結果とした。

ところで，単音節の認識率と単音節のつながりとしての単語の認識率との関係については，すでに阿部ら[8]によって文字認識を対象として行われている理論的検討の結果が，近似的に適用できると考えられる。ここではこのモデルを適用したときの理論的結果と，実際の実験結果との比較を行う。

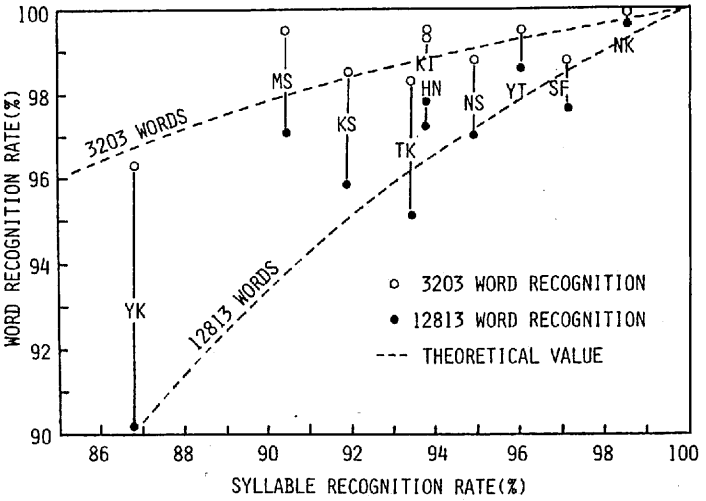

Fig. 8-Relation between syllable recognition rate and word recognition rate.

4.3. シミュレーション実験の結果

10名の各話者の単音節認識率と単語認識率の関係，および理論値を図8に示す。[9] 実験値には話者によるばらつきがあるが，理論値とおおむねよく対応している。

図9は，正しい単語が第1位～第5位に入る割合いの，10名の話者に関する平均値を示したものである。第1位の平均認識率は，3,203語に対し98.9%，12,813語に対し96.6%であり，第2位の候補までとれば，それぞれ99.6%と98.9%の認識率を得ることができる。

単音節認識では/su/↔/tsu/の誤りが多かったが，この区別が唯一の手がかりとなって区別される単語対は日本語の辞書中には少ないため，これはほとんど単語の誤認識の原因とはなっていない。

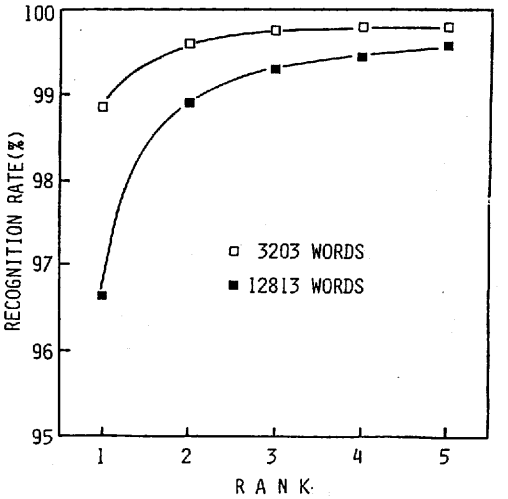

Fig. 9-Accumulated word recognition rate for the large vocabulary word recognition.

4.4. 音節数の自動決定と促音の検出

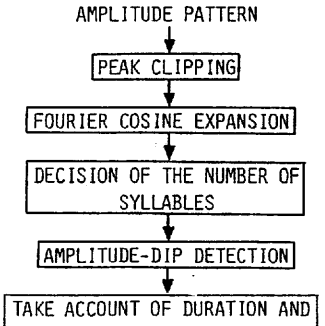

Fig. 10-Block diagram for syllable number decision and /Q/-mora detection.

実際に音節単位に発声された単語音声中に含まれる音節（モーラ）数を自動的かつ正確に決定できれば，単語の候補がしぼれるので非常に有効な手段となる。また促音も特別なキー操作等を行わずに入力できることが望ましい。ここでは図10に示すような，単語音声区間全体の振幅パタンのフーリエコサイン展開係数（式(1)と同様）を用いた方法により，これらの機能の実現をはかった。[10]

図11に，/koQkaʃikeN/（国家試験）と発声したときの振幅パタンを，図12に，/waraibanaʃi/（笑い話）を含む2つの場合の展開係数の抽出例

を示す．この値が急激に低下する位置から音節数（モーラ数，促音を含む）を決定し，振幅値ヒディップによって音節のセグメンテーションと促音の検出を行う．男性2名が発声した，促音を含む4～6音節（モーラ）の210単語音声を用いて，促音の検出実験を行ったところ，各音節数に対し検出率はそれぞれ91%，92%，97.5%であった．

4.5. 標準パタンの適応化と単語音声認識

前述の3,203語を対象として，そこから無作為に抽出した321単語を男性1名（Sub.SF）が平均125音節／分の速度で発声した音声を用いて，認識実験を行った．各音節の標準パタンは3.1および4.2節の実験で用いたものを初期値とし（未知音声との時期差は6か月），未知音声を発声する前に，68単音節を各2回と，学習用単語として定めた10単語を1回ずつ発声して，適応化（フーリエ展開係数は平均化，時系列はなぎかえ）を行った．

実験の結果，音節数の決定とセグメンテーションの誤りは全くなく，音節毎および単語としての認識率は図13に示す通りであった．第1位の音節認識率は90.5%，単語認識率は99.1%で，これらの結果は図5（発声速度と音節認識率の関係）および図8（音節認識率と単語認識率の関係）の結果とほぼ対応している．誤認識単語の内訳は，/oNbou/→/oNtou/，/zaigaku/→/zaitaku/，/seiraN/→/heitaN/ である．音節認識において第10位までに正しい音節が入らなかったのは，1,284音節中/ba/，/ga/（2回），/ro/，/tsu/の5音節（0.4%）であった．

Ⅴ むすび

スペクトル時系列の大局的・局所的両特徴の組合せによる単音節認識方法を提案し，日本語68単音節および101単音節の認識実験により，この方法が，従来試みられた他の方法よりも少ない処理量で，高い性能を有することを確認した．さらに，この方法を用いれば，100音節／分 程度の速度で音節単位に発声した数千～1万程度の大語いの単語音声に対して，高い認識率が得られることがわかった．ここで検討した方法では，音節数（モーラ数）を自動的に決定することができるので，音節数が異なれば，語い数がいくら増加しても認識率は低下しない．

促音の自動検出や，実際に単語を発声したときの訓練効果，標準パタンを逐次適応化してゆく効果等については，発声者をさらにふやして，実験と改良を進めて行きたい．

[謝辞] 日頃御指導頂く当所鮫岡基礎部長，橋本統括役，板倉第4研究室長に感謝する．実験の一部を担当された松永社員，ならびに討論頂いた研究室の諸氏に感謝する．

[文献] (1) 坂井他：信学誌 p.1696 (昭38). (2) 吉田他：音学講論 3-2-16 (昭54-06). (3) 新田他：信学総全大 1350 (昭55). (4) 古井：音学講論 1-1-18 (昭55-10). (5) 古井：信学総全大 1351 (昭56). (6) 古井他：音学講論 4-1-25 (昭53-05). (7) S. Furui：IEEE Trans. ASSP. p.254 (1981). (8) 阿部他：信学論 52-C. p.305 (昭44). (9) 古井：音学講論 3-1-12 (昭56-05). (10) 古井：音学講論 3-1-2 (昭56-10). (11) 鹿野：音声研資 S80-60 (昭55).

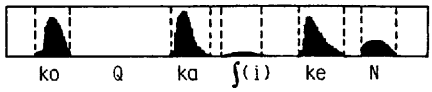

Fig. 11-An example of the amplitude pattern of a spoken word.

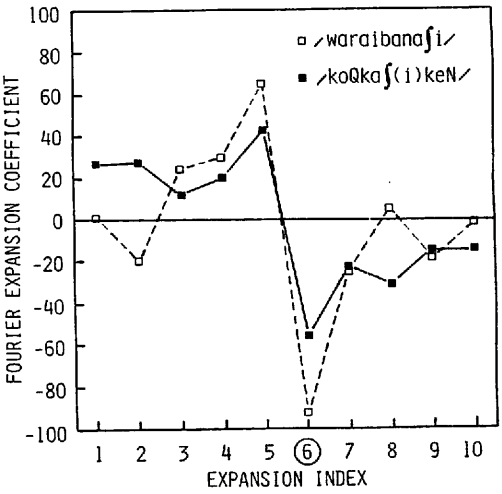

Fig. 12-Examples of the Fourier cosine expansion coefficients.

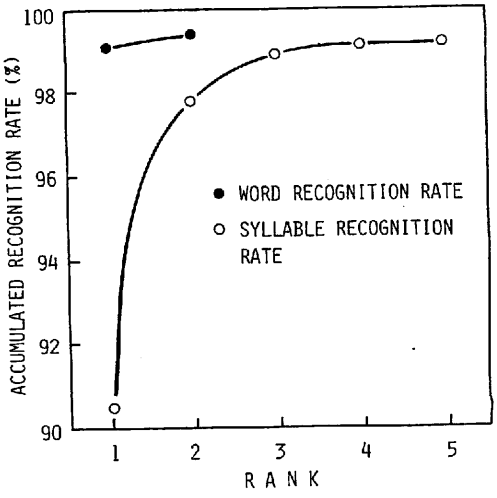

Fig. 13-Accumulated syllable and word recognition rates.