論文 / 著書情報
Article / Book Information

論題(和文)	SPLIT法における標準パタン数の減少効果
Title	
著者(和文)	菅村 昇, 古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会 音声研究会資料 S81-71, , , pp. 565-572
Journal/Book name	, , , pp. 565-572
発行日 / Issue date	1982, 1

SPLIT 法における標準パタン数の減少効果

管村　昇　　古井 貞熙
（日本電信電話公社　武蔵野電気通信研究所）

昭和57年1月26日　　於 大阪大学

内容梗概　単語音声認識の方法として提案したSPLIT法において，認識に用いる擬音韻標準パタン数と認識率との関係について報告する。擬音韻標準パタン作成時のクラスタリングの閾値を可変にし，標準パタン数を16～256個選択した場合の認識結果について述べる。この結果，標準パタン数が32個の場合の641都市名単語の認識率は，94.5%（4話者平均），64個では95.3%となり，256個の標準パタンを用いた場合の認識率に比べ，それぞれ1.9%，1.1% の低下におさまることが明らかになった。つぎにこれらの少数パタンの分析を行い，パタンと音韻性との対応関係を求め，いくつかの音韻とよく対応がとれていることを示す。この結果，SPLIT法が単に情報圧縮ということにとどまらず，音韻との対応を利用した処理にも利用できる可能性が得られた。

Relationship Between Number of Pseudo-Phoneme Templates and Recognition Accuracy in SPLIT Method

Noboru SUGAMURA and Sadaoki FURUI
(Musashino E.C.L.,N.T.T.)

Jan. 26,1982　　at Osaka University

Abstract　This paper describes a relationship between the number of pseudo-phoneme templates and recognition accuracy in the SPLIT word recognition system. Pseudo-phoneme templates are automatically extracted from training speech data using a clustering technique which we proposed previously. The number of pseudo-phoneme templates is relevant to memory saving for word dictionary and reduction of calculation amount for D.P. matching. Experimental results show that the recognition rate decreases only 1.9 % by reducing the number of pseudo-phoneme templates from 256 to 32. Correspondence between the pseudo-phoneme and the conventionally defined phonemes is also investigated using a multi-dimensional analysis technique.

1. まえがき

単語音声認識の方法としては，これまでいくつかの方法が提案され，一部のものは実用化されている。我々は，先に新しい単語音声の認識方法として，スペクトルパタンの自動的なクラスタ化によって得られる擬音韻標準パタンを用いる方法（ＳＰＬＩＴ法とよぶ）を提案し，いくつかの実験を通じて，その有効性を確認してきた。[1][2] この結果，特定話者の大語い単語音声認識や，マルチテンプレート法による不特定話者単語音声認識[3]などに適用できる見通しを得た。これらの検討結果を踏まえ，１０００単語を実時間で認識できる装置を試作し[4]，シミュレーション実験と同等の性能を確認した。本稿では，ＳＰＬＩＴ法において，擬音韻標準パタン数を減少させた場合の認識率の実験結果について述べ，擬音韻標準パタン数を固定した場合の，最適なクラスタリングの方法について述べる。また擬音韻標準パタンと音韻性との関係を分析した結果についても報告する。

2. SPLIT法による単語音声認識

単語音声認識に用いる単語辞書として，フレーム単位に，分析されたスペクトルパラメータをそのまま蓄積しておかずに，スペクトルパラメータに一種の量子化を施した離散的な標準パタンの系列で表現することが考えられる。このようにすることによって，ＤＰのためのスペクトル距離計算量，単語辞書に対するメモリ量を減少させることができる。このような考え方をもとに，新しい認識方法を提案したが以下に簡単にその方法について説明する。

2-1 単語音声認識システム

新しい認識方法を，ここではクラスタリングにパラメータ分布空間を分割（Splitting）していくことと，このようにして得られる擬音韻標準パタンの系列で単語辞書を表現することを兼ねてＳＰＬＩＴ（Strings of Phoneme-LIke Templates）法とよぶ。

認識システムを図１に示す。本システムでは未知入力音声の認識に先立って，つぎの２つの準備が必要である。

（１）擬音韻標準パタンの作成

あらかじめ発声された音声データをもとに，簡単なクラスタリング[5]によって，擬音韻標準パタンを作成する。この場合，各クラスタの広がりを指定するクラスタリングの閾値は，単語認識システムにおいて，どれくらいの標準パタンを用いるかということから決定する必要がある。

（２）単語辞書の作成

擬音韻標準パタンを用いて，認識対象単語をこれらの系列として表わす単語辞書を作成する。すなわち認識対象単語をあらかじめ発声し，スペクトル分析したパラメータと擬音韻標準パタンとのスペクトル距離計算をフレームごとに行い，スペクトル距離が最小となる擬音韻標準パタンを見出す。このようにして認識対象単語を擬音韻標準パタンの系列で表現し，図１-（ｄ）に蓄積しておく。

以上のような準備をした後，未知入力音声の認識を行う。入力音声は，図１-（ａ）でスペク

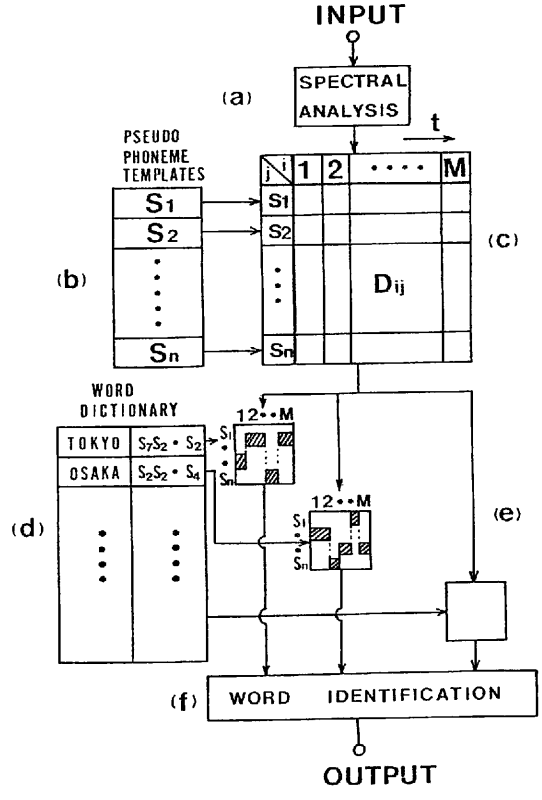

図1 SPLIT法による単語音声認識システム

トル分析され，フレームごとに擬音韻標準パタンとのスペクトル距離計算が行われ，距離行列が生成される。この距離行列と単語辞書を用いて，音声情報の時間伸縮を吸収するスペクトルマッチングをDPによって行い，マッチング距離が最小となる単語を認識結果として出力する。なおこの段階において，誤認識をさけるためにマッチングの距離の値に閾値を設けて，リジェクトすることも可能である。DPマッチングを行う場合の距離計算は，1つの未知入力単語に関し，各擬音韻標準パタンに対して1回だけ計算しておけば，認識対象単語のすべてに利用できる。このためDPのためのスペクトル距離計算量が，単語辞書として分析パラメータをそのまま蓄積しておく方法に比べ，大幅に減少できる。

2-2 SPLIT法の特徴

単語認識方法として，単語辞書に分析パラメータをそのまま蓄積しておき認識する方法[6]（方法Aとする）と音韻を単位とする標準パタンを作成し，このパタンで表現した単語辞書を用いる認識方法[7]（方法Bとする）をとりあげ，SPLIT法の特徴について，これらの方法と比較して説明する。

(1)標準パタンの作成に関して，方法Bでは，音韻に対応する標準パタンの決定が，ややむずかしいが，SPLIT法では，標準パタンの作成に，自動的なクラスタリング手法を用いているため，つぎのような特徴をもっている。

(i)パラメータ分布空間に距離を導入することだけで標準パタンを作成することができ，人手が介入する過程がない。このことは大量のデータを処理するのに適しており，データベースが同じであれば，どこでも同じ標準パタンを作成することができ，あいまいさがない。

(ii)国語に依存しない標準パタンを構成できるので認識対象の言語に対する適応性がある。たとえば認識対象言語が英語などであっても同様の処理で行える。

(iii)認識対象単語の種類にかかわりなく標準パタンを構成することができ，単語を構成している音韻を意識する必要がない。

(2)SPLIT法では，DPマッチングのためのスペクトル距離計算量および標準パタン，単語辞書用のメモリ量が，方法Aに比べて大幅に減少できる。方法Aと比べ，DPのための距離計算量，メモリ量の比較を行うとつぎのようになる。

【距離計算量の比較】

認識対象単語数をL，未知単語のフレーム数をM_ℓ，DPの窓幅をN_w，標準パタン数をNとすると，1単語を認識するためのDPに必要なスペクトル距離計算回数は，方法Aでは，約$M_\ell \times N_w \times L$であるのに対し，SPLIT法では，$M_\ell \times N$である。したがって，SPLIT法の方法Aに対する1フレームあたりの距離計算量の比は，およそ

$$N/(N_w \times L) \qquad (1)$$

となる。

【メモリ量の比較】

標準パタン数をn_bビット($\log_2 N$)，1フレームのスペクトルパラメータの個数をn個，1つのパラメータ精度をN_aビットとすると，方法Aでは

$$\left(\sum_{\ell=1}^{L} M_\ell\right) \times n \times N_a \ \text{ビット} \qquad (2)$$

であるのに対し，SPLIT法では

$$N \times n \times N_a + \left(\sum_{\ell=1}^{L} M_\ell\right) \times n_b \ \text{ビット} \qquad (3)$$

である。

1単語のフレーム数を50フレーム，Nを128個，したがってn_b=7ビット，N_aを16ビット，nを16パラメータ，N_wを15フレームとしたときの，SPLIT法の方法Aに対するDPのための距離計算量の比，メモリ量の比を，認識対象単語数の関数として図2に示す。この結果から，メモリ量，DPのための距離計算量の点からは，SPLIT法は，大語いになればなるほど方法Aに比べて有利であることがわかる。

以上，SPLIT法の長所を方法A，Bと比較して説明したが，問題点としては，つぎのようなことがあげられる。

(1)SPLIT法は，方法Aにおいて単語辞書のスペクトルパラメータを量子化して表現したものを利用する方法と考えることができる。このことから量子化が粗い場合には，量子化誤差が大きくなり，スペクトルパラメータの時系列を精度よく表現できず，方法Aに比べ，認識率

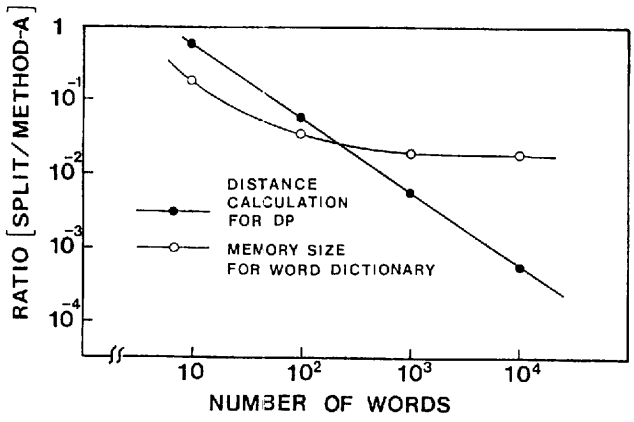

図2　メモリ量比，距離計算量比と認識対象単語数との関係

が低下することが予想される。これは似かよった単語が多くなる大語い単語認識の場合，特に注意を要するが，これまで641国内都市名単語の認識実験の結果，擬音韻標準パタン数を128～256個選べば，認識率は，方法Aに比べ，ほとんど低下しないことが確かめられている。[2] しかも，これらの擬音韻標準パタンは，話者が変ってもほぼ共通に利用できることも実験的に確かめられている。[8] また学習サンプルを用いてあらかじめ擬音韻標準パタンを作成する過程が必要であり，このためのメモリや演算が必要であることなどが方法Aに比べての問題点である。

(2) SPLIT法や方法Aでは，スペクトルパタンと音韻との関係は，単語認識においては必ずしも必要としないという立場であり，認識対象語いについては，利用者があらかじめ発声して登録しておくことを前提にしている。このため話者が変った場合や単語辞書を変更，追加する場合に利用者に大きな負担となる。これに対し，方法Bでは，標準パタンと音韻の間で対応がついているため，単語辞書の自動作成ルールができれば，認識対象単語を発声しなくても単語辞書を作成することができる。[9] また話者が変った場合には，学習によってその話者の音韻標準パタンが作成できれば，話者への適応もきわめて容易に行える。[10]

以下では，擬音韻標準パタン数を，256個よりも少くした場合の認識率の低下を明らかにするとともに，前述した方法Bに対する欠点を検討するために，少数の擬音韻標準パタンと音韻性との対応を明らかにしていく。

3. 擬音韻標準パタンの減少効果

前述したように，メモリ量，距離計算量は，認識対象語い数だけでなく擬音韻標準パタン数の関数でもある。Lを1000語とした場合に，擬音韻標準パタン数を1024～16（10ビット～4ビット）に変化させた場合のメモリ量，距離計算量の比は，図3のようになる。

擬音韻標準パタン数を減少させた場合，認識率が低下しないなら，パタン数が少ないほど望ましいのは言うまでもないが，一般にはそれだけ量子化誤差が増加し，認識率の低下が予想される。一方単語のもつ冗長性から，それぞれのフレームは必ずしも精確に表現する必要はないことが，これまでの実験結果から確かめられている。ここでは，このような観点から，標準パタン数を256～16まで変化させた実験結果について述べ，標準パタン数と認識率との関係を明らかにする。

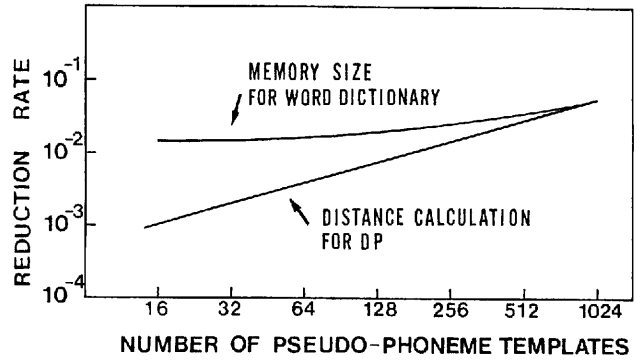

図3　SPLIT法の方法Aに対する距離計算量メモリ量の比較

3-1 クラスタリングの閾値固定の場合

あらかじめ発声された男性話者1名の音声データをもとに，先に提案したクラスタリング手法[5]により，擬音韻標準パタン$S_1 \sim S_N$が得られたとする。ここでNは，クラスタリング時におけるスペクトル距離θの関数である。Nが256以上になるようにθを設定してクラスタリングを行い，作成した標準パタンの中から，同一クラスタに属するフレーム数が多い順に，16，32，64，128，256個選び認識実験を行った。スペクトルパラメータの抽出条件認識条件は，表1に示している。$\theta=0.05$とした場合の実験結果を図4の折線Aに示す。この結果からわかるように，θを小さい値に固定して標準パタン数を減少させると，標準パタン数が64個の場合，256個の場合に比べ，認識率が1.5％低下し，32個では，約6％低下する。

3-2 クラスタリングの閾値可変の効果

前節では，θを固定したクラスタリングによって，同一クラスタに属するパタン数が多い順に，標準パタンを選択して認識実験を行った。しかし，1章でも述べたように，最終的に用いる標準パタン数が決定されている場合，それに対して適切なクラスタリングの閾値を設定することが望ましい。ここでは，クラスタリングにおけるスペクトル距離の閾値θを変化させ，同じ標準パタン数で認識率がどのように変化するかを実験的に明らかにする。$\theta=0.10$，0.15，0.20，0.25，0.30とさらに5種類のθを選びクラスタリングを行い，得られた標準パタンを用いて認識実験を行った。$\theta=0.20 \sim 0.30$の場合，標準パタン数は，それぞれ256，128個未満となった。実験結果を，同じ図4のB，C，D，E，Fに示す。この結果，標準パタン数を32個にしても，θを適切に選べば，96％程度の認識率が得られること，また16個にしても93％程度の認識率が得られることが明らかになった。

$\theta=0.30$における32個の標準パタンは，認識率だけの観点からは，$\theta=0.05$における256個の標準パタンと同じ効果をもつものと考えられる。$\theta=0.05$の場合に，パタン数を32個以下にすると，認識率が急に低下するのは，クラスタリングのスペクトル距離の閾値が小さいため，出現頻度の高い音韻のスペクトルのみが標準パタンとして数多く選択され，単語辞書が精度よく表現できないためである。

表1 認識実験条件

実験対象単語	都市名 641単語
発声者	男性1名，2回発声
分析条件	8KHz サンプリング
	32msec ハミング窓
	フレーム周期 16msec
分析パラメータ	自己相関係数
	LPCケプストラム
	それぞれ16次
スペクトル距離尺度	WLR[11]
DP手法	両端点フリーDP[12]

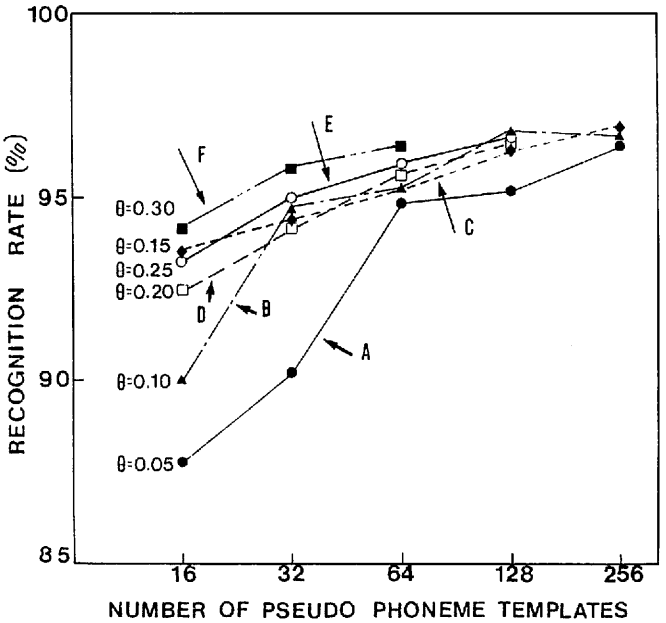

図4 クラスタリングの閾値θを可変にした場合の擬音韻標準パタン数と認識率との関係

つぎに擬音韻標準パタンの作成時に，もとの音声サンプルの情報が，どの程度反映されているかを定量的にみるために，つぎのような評価値を定義する。

いまM個のフレームデータをもとにクラスタリングを行いS1～SNの標準パタンが得られたとし，それぞれのパタンが同一クラスタに含むフレーム数をm1, m2, ……, mNとし，m1≧m2≧…≧mNの関係があるとする。Nはθを固定してクラスタリングを行った場合に得られる標準パタン数であり，$\sum_{i=1}^{N} m_i = M$の関係を満たす。N個の標準パタンのうちn個を選んだとき，このn個には，もとのフレームのうち，どれだけが含まれているか，すなわちパタン作成に寄与しているかを次式で表わす。

$$R = \sum_{i=1}^{n} m_i \ / M \qquad (4)$$

このRを包含率（Covering Rate）とよび，それぞれのθにおける標準パタン数とRとの関係を図5に示す。この結果，前述したθ=0.05 0.30の標準パタン数32個のときのRの値はそれぞれ23%，92%であり，θ=0.30の場合の方がより多くの音声サンプルの情報が，標準パタンに反映されていることがわかる。

3-3 単語辞書作成時におけるスペクトル誤差と認識との関係

クラスタリングにおけるθを変化させ，各標準パタン数における最適なθを，そのつど認識実験を通じて決定しようとすると膨大な時間がかかる。一方擬音韻標準パタンの作成と，単語辞書の作成は，認識実験に比較して短時間で行える。またDPそのものがスペクトルマッチングを基本にしていることから，単語をどの程度精確に表現できるかが，認識率の1つの目安となることが推定される。そこでθと標準パタン数を変えたときの，単語辞書作成時におけるスペクトル誤差を評価にとり，認識との対応関係について検討する。ここでスペクトル誤差は，次式で定義する。

$$SD = \sum_{i=1}^{M} \sqrt{D_i} \ / M \qquad (5)$$

ここで　D_i はiフレームのWLR距離
　　M　はフレーム総数

この式を用いて，単語辞書作成時のスペクトル誤差を，前述したすべての場合について計算した結果を図6に示す。図において，認識に用いる標準パタン数を固定した場合に最も認識率が高くなるポイントを矢印で示している。この結果，各標準パタン数において，スペクトル誤差最小のところが，ほぼ最高認識率を示しており，単語辞書作成のスペクトル誤差を認識率の

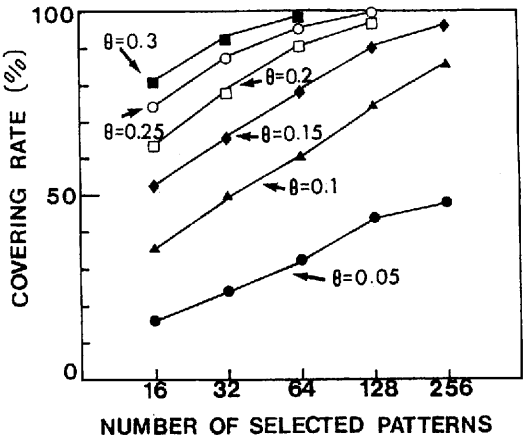

図5　選択する標準パタン数とRとの関係

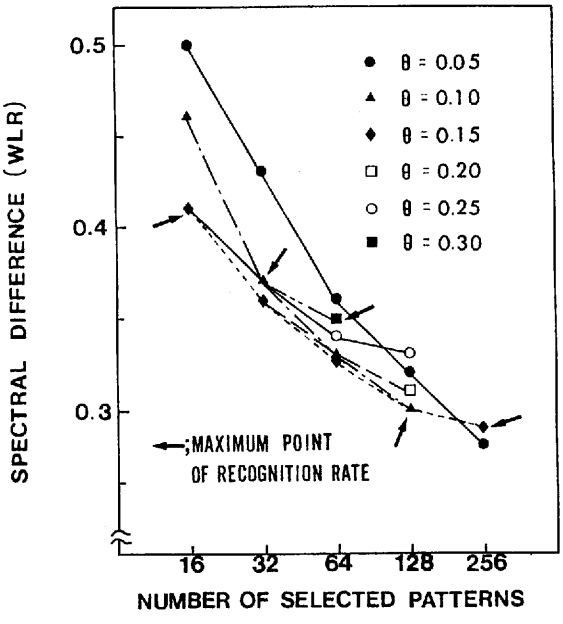

図6　選択する標準パタン数と単語辞書作成時におけるスペクトル誤差との関係

目安を得る1つの評価関数として用いることができることが明らかになった。

このような実験結果を踏まえ，標準パタン数を減少させた場合の認識率の信頼性を増すために，さらに男性話者3名の追加実験を行った。この3名に対しては，ここで述べたように，単語辞書作成時におけるスペクトル誤差最小のポイントにおける認識実験のみを行った。

話者4名の認識率とその平均値を，それぞれの標準パタン数ごとに整理して表2に示す。またその場合のクラスタリングの閾値も示している。さらに比較のために，方法Aによる認識率も同表に示している。

表2 スペクトル誤差最小点における認識率

N \ SPEAKER / R.R.		NS.	KA	KI	KS	AVE.
16	R.R.	92.5	94.5	93.0	91.6	92.9
	θ	0.20	0.25	0.25	0.20	
32	R.R.	94.2	96.4	94.2	93.0	94.5
	θ	0.20	0.20	0.15	0.15	
64	R.R.	95.2	96.3	95.5	94.2	95.3
	θ	0.15	0.15	0.10	0.10	
128	R.R.	96.7	96.1	96.7	95.6	96.3
	θ	0.10	0.10	0.10	0.10	
256	R.R.	96.6	96.6	97.0	95.2	96.4
	θ	0.10	0.10	0.10	0.10	
METHOD-A		97.0	97.0	97.7	96.4	97.0

この結果，標準パタン数が減少するにしたがって認識率は低下していくが256の場合に比べ，パタン数16で3.5%，32で1.9%，64で1.1%程度であることが明らかになった。なお256の場合と方法Aとの差は0.6%である。

4. 擬音韻標準パタンと音韻性との関連

2章で述べたようにSPLIT法では音韻と標準パタンとの対応は，必ずしも必要ではないという立場から出発している。しかし3章で述べたように擬音韻標準パタン数を32個程度にしてもθを適切に選べば，93～96%程度の認識率が得られることから，これらのパタンを解析し，音韻性との関係を明らかにしておくことは，つぎの点において意味あることである。

(1) SPLIT法が，単に方法Aの情報圧縮ということにとどまらず，方法Bに対する欠点を補える可能性がある。

(2) 32個程度の標準パタン数は，人間にとって扱いやすく，認識の種々の処理に利用できる可能性がある。

擬音韻標準パタンの一例として，話者KIのθ=0.15における上位(同一クラスタに属するフレーム数が多い順に)16個のスペクトル包絡を図7に示す。またこれらを含む上位32個の標準パタンを，多次元尺度構成法[13]により，2次元平面上にプロットしたものを図8に示す。あらかじめ人間が音声データの各フレームごとに音韻のラベリングをしたものと，この上位32個の標準パタンとの対応をフレームごとにとり，各標準パタンに最も頻度が高く対応づけられた音韻の種類を同図に示す。これらの結果，5母音，/P,t,k/，/A/，/m,n/などへの対応が比較的よく得られていることがわかる。また5母音は，1つのクラスタに3～7の標準パタンが存在していることがわかる。

5. あとがき

SPLIT法について概説し，SPLIT法の特徴を，これまでの代表的な二つの方法と比較して，その長短所を説明した。つぎにSPLIT法における擬音韻標準パタン数の減少効果について述べ，標準パタン数と認識率との関係について報告した。この結果，クラスタリングにおけるスペクトル距離の閾値を，単語辞書作成のスペクトル誤差を評価関数として適切に設定することにより，標準パタン数32個で95%，16個で93%程度の認識率が得られることが明らかになった。さらに32個のパタンの音韻性との対応関係について検討し，いくつかの音韻とはよく対応が取れていることを示した。これらの結果，SPLIT法が，単に情報圧縮や距離計算量の低減という特徴だけではなく，方法Bに対する欠点を補える可能性をもっていることが明らかになった。

謝辞 日頃御指導頂く畔柳基礎研究部長，橋本統括役，板倉第四研究室長に深謝します。また認識実験に際し，御協力頂い

た第四研究室鹿野調査員，杉山研究主任に感謝します。またスペクトル誤差評価は鹿野調査員の提案によるものであり，ここに付記します。

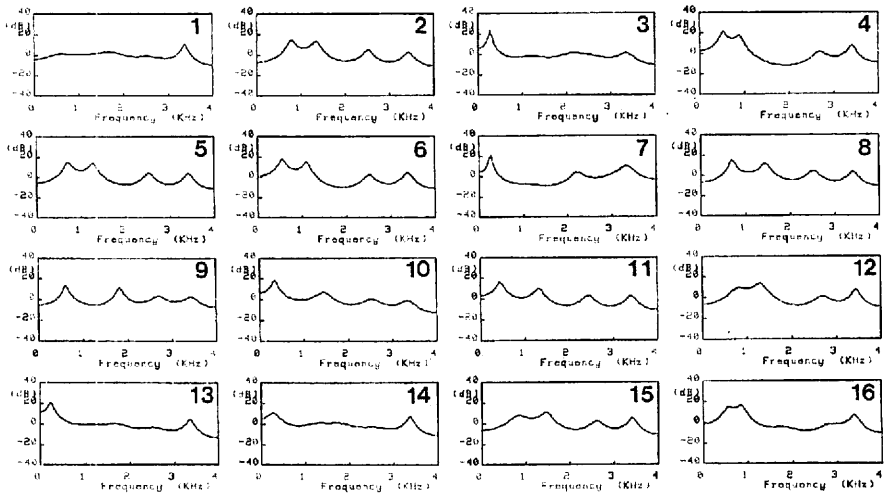

図7　16個の擬音韻標準パタンのスペクトル包絡（話者 KI，θ=0.15）

参考文献

(1) 管村,古井,箱田：擬音韻標準パタンによる大語い単語音声認識　信学総全大(1981)
(2) 管村,古井,箱田：クラスタ化による音韻標準パタンを用いた単語音声認識　音声研究会資料 S80-61 (1980-12)
(3) 古井,管村：SPLIT法による不特定話者単語音声認識　信学総全大発表予定（1982－3）
(4) 古井,管村,笹島：SPLIT法による大語い単語音声認識装置　音響学会講論集（昭56－10）
(5) 管村,板倉：パタンマッチング符号化による音声情報圧縮　音声研究会資料 S79－08（1979－5）
(6) 迫江,千葉：動的計画法を利用した時間正規化に基づく連続音声認識　音学誌 27，9（1971）
(7) 好田,橋本,斎藤：数字音声の機械認識系　信学論，55D－3（1972）
(8) 鹿野,管村：単語音声認識における標準パタンの効果，音響学会講論集（昭56－10）
(9) 箱田：音韻を単位とする単語音声認識における辞書の作成法　音響学会講論集（昭56－10）
(10) 古井：単語音声認識における学習の能率化　音声研究会資料　S77-43（1977-12）
(11) 杉山,鹿野：ピークに重みをおいたLPCスペクトルマッチング尺度の提案，音声研究会資料　音声研究会資料　S80-03（1980）
(12) 鹿野：大語い単語音声認識におけるLPCスペクトルマッチングの提案　音声研究会資料 S80－60（1980）
(13) ニャナデシカン著，丘本，磯貝訳：統計的多変量解析　日科技連出版（昭54）など

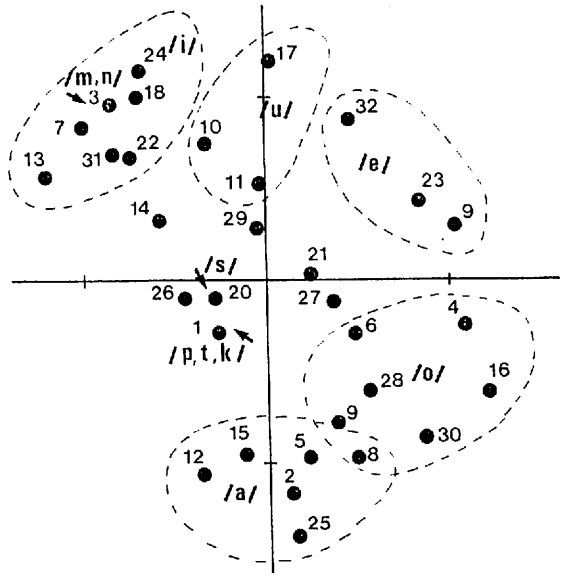

図8　32個の擬音韻標準パタンの2次元での位置関係（多次元尺度構成法による）