論文 / 著書情報
Article / Book Information

論題(和文)	単語音声認識における時間正規化方法の検討
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子通信学会, PRL76-25, Vol. 76, No. 25, pp. 59-68
Citation(English)	, Vol. 76, No. 25, pp. 59-68
発行日 / Pub. date	1976, 7
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1976 Institute of Electronics, Information and Communication Engineers.

ＰＲＬ７６－２５

単語音声認識における時間正規化方法の検討

古　井　貞　熙

（武蔵野通研）

1976年7月20日

社団法人 電子通信学会

PRL76-25

単語音声認識における時間正規化方法の検討

TIME-NORMALIZATION METHODS FOR MACHINE RECOGNITION OF SPOKEN WORDS

古井 貞熙（日本電信電話公社 武蔵野電気通信研究所）
Sadaoki FURUI (Musashino ECL, NTT)

Abstract: Time-normalization methods to optimally register the reference pattern onto the input one are investigated from the viewpoint of the effectiveness to the isolated word recognition. Four kinds of conditions are applied to the time-warping function, and are evaluated. It has been made clear by the recognition experiment of the spoken digits, that the time-normalization method which uses the length of the utterance is most effective. Reference pattern of each word is compressed into a phoneme string or preserved uncompressed, and the performance of the two cases are compared.

1. まえがき

音声の機械認識の研究には，単語から文章を対象とするものまで種々のレベルのものがあるが，単語音声の認識が現在最も実用化の可能性が大きいと考えられる．区切って発声した単語音声の認識に関して重要な課題の一つに，時間伸縮の正規化の問題があり，この対策としては，入力サンプルと標準パターンの時系列の間で，伸縮の度合いに一定の制約を持たせながら，最もよく整合するようにDP（動的計画法）の手法を用いて非線形に対応づける方法が広く用いられている．[1][2][3]

この伸縮の度合いに関する制約の与え方としては，種々の方法が試みられているが，これらの有効性に関する定量的な評価や総合的な比較はまだ行われていないようである．ここでは認識実験により，これまでにとりあげられていない新しい方法を含めた種々の方法の評価を行う．

個人差の能率的な学習や，認識時間の短縮，装置の簡単化などのためには，標準パターンをどのような単位で用意するかという問題も重要である．ここでは単語単位に用意する方法と音韻単位に用意する方法の比較を，上記の検討とあわせて行う．

2. 音声スペクトルと類似度の表現

単語音声波を3.2kHzの低域通過フィルタに通したのち，8kHz，11ビットでサンプリングおよび量子化を行い，15msの区分（フレーム）毎に10次までの相関係数を抽出する．さらに線形予測分析法（最尤スペクトル推定法）により，予測係数を求める．

各区分における入力サンプル（x）と標準パターン（予測係数aで表現する）の距離（類似度）を，板倉[3]の最小予測残差原理（Minimum Prediction Residual Principle）にもとづいて，次のように正規化対数予測残差によって表現する．

$$d(x|a) = \log(a\mathbf{V}a'/\hat{a}\mathbf{V}\hat{a}') \quad (1)$$

$$= \log(aa) + \log[(br)/(\hat{a}r)] \quad (2)$$

ここで，(yz)は内積をあらわし，$\mathbf{V}$は相関行列，rは相関係数，$\hat{a}$は入力サンプルから推定した予測係数で，$b=(1, b_1, b_2, \cdots, b_p)$の各要素は次の通りである．

$$b_i = 2\sum_{j=0}^{p-i} a_j a_{j+i} / (aa) \quad , p=10 \quad (3)$$

正規化対数予測残差は，入力信号（入力サンプル）をモデル（標準パターン）の予測係数aで処理したときの予測残差パワーと，入力信号に関する予

剰残差パワーの最小値（入力信号から推定した $\hat{a}$ を用いたときの予測残差パワー）の比になっている．a で定義されるモデルが，x を生成した実際のプロセスに近ければ，最尤推定値 $\hat{a}$ は a に近くなって，$d(x|a)$ の値は 0 に近づき，そうでなければ $d(x|a)$ は大きな値になる．$d(x|a)$ には尤度比としての意味もあり，また $d(x|a)\cdot N$ は，x が a で表現されるモデルに従うとき，漸近的に自由度 p の χ^2 分布に従う．（N はデータ数）

〔注〕従来通研において好田らによって用いられて来た対数尤度[2]を $\ell(x)$ であらわすと，

$$\ell(x) = -\log(br) - \log(xa) \quad (4)$$

$$= -d(x|a) - \log(\hat{a}r) \quad (5)$$

と書け，$\ell(x)$ は正規化対数予測残差の符号を変えて，入力サンプル a のみに依存する項を除去したものになっている．これは，従来通研で用いられて来た認識方法では，ここで後に用いる方法のうちのいくつかと異なって，入力サンプルのみに依存する項は認識の判定に影響を与えないためである．

3. 非線形伸縮マッチング

3.1. 時間軸の正規化　入力サンプルと k 番目の標準パターンとの総合的な距離（類似度）を次のように定義する．

$$D(k) = \min_{C}\left[\left\{\sum_{\ell=1}^{L} d(c_\ell; k)\cdot\Delta_\ell\right\} \Big/ \sum_{\ell=1}^{L}\Delta_\ell\right] \quad (6)$$

ただし，

$$C = \{c_\ell\} = \{(m_\ell, n_\ell)\} : \ell = 1, 2, \cdots, L$$

$$c_1 = (1, 1),\ c_L = (M, N(k)) \quad : \text{境界条件} \quad (7)$$

c_ℓ は標準パターンの時点（区分）と入力サンプルの時点（区分）の対応づけ（組合わせ）を示し，C はその集合（系列）である．$d(c_\ell; k)$ は，c_ℓ で対応づけられた時点における入力サンプルと k 番目の標準パターンとの距離で，(4) 式で示した正規化対数予測残差によって表現する．

この操作は，図 1 に示すように，入力サンプル（長さ M）と標準パターン（長さ N(k)）の時間軸を対応づける伸縮関数：

$$n = w(m), \quad (c_\ell = (m_\ell, n_\ell) = (m_\ell, w(m_\ell))) \quad (8)$$

を考え，両者の全体としての距離が最小となるものを選ぶことに相当する．上記の c_1, c_L の境界条件は，入力サンプルと標準パターンの始端と終端をそれぞれあわせることに対応する．

Δ_ℓ としては，c_ℓ と $c_{\ell-1}$ の絶対値距離（city block distance）を用い，$\Delta_\ell = |m_\ell - m_{\ell-1}| + |n_\ell - n_{\ell-1}|$ とする．こうすると (6) 式の分母は，

$$\sum_{\ell=1}^{L}\Delta_\ell = M + N(k) - 1 \quad (9)$$

となり，一定値となる．

3.2. 伸縮関数に関する制約　伸縮関数 $w(m)$ のとりうる範囲とそのステップに，次のような4通りの制約を加える．

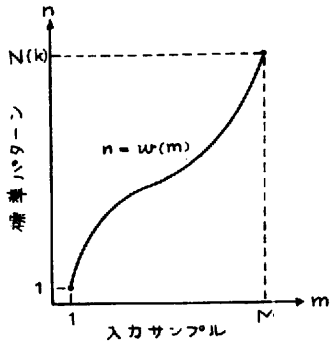


図 1. 入力サンプルと標準パターンの時間軸の対応づけ

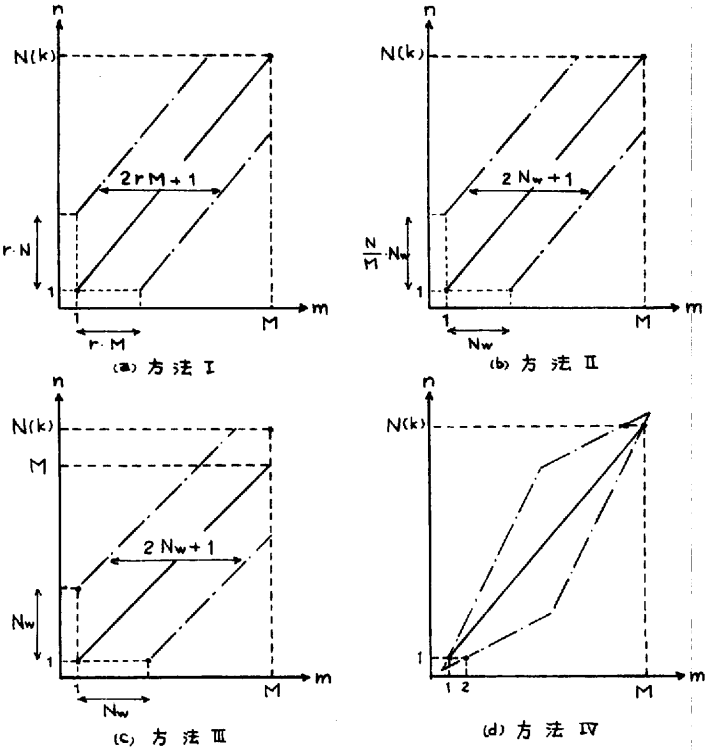


図 2. 伸縮関数に対する種々の制限（整合窓）

[方法Ⅰ][6]: $|w(m)-m\cdot N(k)/M| \le r\cdot N(k)$ (10)

[方法Ⅱ][4]: $|w(m)-m\cdot N(k)/M| \le Nw$ (11)

[方法Ⅲ][1]: $|w(m)-m| \le Nw$ (12)

ただし，$0 \le r \le 1$，$0 \le Nw$ とする．これらの三つの方法は，m－n平面上に図2(a)～(c)に示すような領域を設けて，伸縮関数をその内側に制限することに相当する．この領域を整合窓と呼ぶことにする．方法Ⅲが，(1,1)と(Min(M, N(k)), Min(M, N(k)))を結ぶ傾き1の直線の両側に対称に，$2Nw+1$ の幅の窓を設けるのに対し，方法ⅠとⅡでは，(1,1)と(M, N(k))を結ぶ対角線の両側に窓を設け，方法Ⅱの窓の幅は $2Nw+1$ で固定であるが，方法Ⅰの窓の幅は $2rM+1$ で，音声サンプル長に比例して変わるようになっている．

方法Ⅲで $|M-N(k)| > Nw$ となる場合は，(M, N(k))の点が整合窓の中に入らないので，このときは対応づけの終端(境界条件)を，窓の中の(M, N(k))に最も近い点に変更する．(他式も変わる)

以上の方法Ⅰ～Ⅲにおいては，入力サンプルと標準パターンの対応づけの系列に関して，C_ℓ と $C_{\ell-1}$ の関係を次のように制限する．

$$C_\ell = \begin{cases} C_{\ell-1}+(1,0), & \text{このとき } \Delta_\ell = 1 \\ C_{\ell-1}+(0,1), & 〃 \quad \Delta_\ell = 1 \\ C_{\ell-1}+(1,1), & 〃 \quad \Delta_\ell = 2 \end{cases} \quad (13)$$

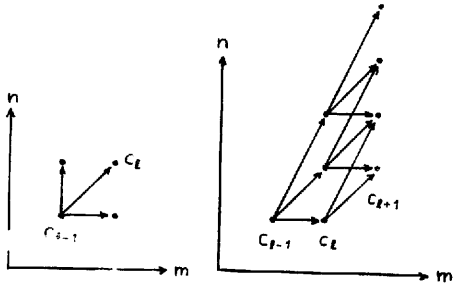
m－n平面で考えると，C_ℓ と $C_{\ell-1}$ の間に，図3(a)のような関係のみを許したことになり，伸縮関数が一般に満たすべき条件である単調性と連続性が確保される．

[方法Ⅳ][3]：上記の方法と多少異なる方法として，次のような方法を検討する．＊で示した C_ℓ と $C_{\ell-1}$ の関係は，$C_{\ell-1}-C_{\ell-2} \neq (1,0)$ の場合のみ許す．

$$C_\ell = \begin{cases} C_{\ell-1}+(1,0), & \text{このとき } \Delta_\ell = 1 \ * \\ C_{\ell-1}+(1,1), & 〃 \quad \Delta_\ell = 2 \\ C_{\ell-1}+(1,2), & 〃 \quad \Delta_\ell = 3 \end{cases} \quad (14)$$

m－n平面で考えると，この方法は，$C_{\ell-1}$，C_ℓ，および $C_{\ell+1}$ の間に，図3(b)のような関係のみを許したことになり，これは入力サンプルと標準パターンの各時点の相対速度を1/2と2の間に制限したことに相当する．この方法は，追江[5]による"傾斜制限DPマッチング"にほぼ等しい．この方法における伸縮関数 $w(m)$ のとりうる範囲は，図2(d)に示す平行4辺形の内側になる．

以上の方法Ⅰ～Ⅳのいずれの方法の場合も，DPの手法を用いて，最適な伸縮関数とそのときの距離 $D(k)$ を能率よく求めることができる．



(a)方法Ⅰ～Ⅲ　(b)方法Ⅳ

図3．許される $C_{\ell-1}$，C_ℓ および $C_{\ell+1}$ の関係

4. 認識実験

4.1. 実験サンプル　男性10名が区切って発声した10数字音声を用いて，認識実験を行った．各話者について，各数字をはじめに1回ずつ発声したサンプルから標準パターンを作成し，その後に各数字を10個ずつ含むランダムな順序の100個の数字を発声したものを，認識実験の未知入力サンプルとして用いた．標準パターンと無声化傾向が異なる入力サンプル(全体の3.5%)は実験の対象から除いたので，実験に用いた入力サンプルは計965サンプルである．

4.2. 標準パターンの作成法　各数字音声の標準パターンを次の2通りの方法で作成し，比較を行った．

[単語単位標準パターン]　学習サンプルを2章で述べたような方法で15msの区分毎に線形予測分析し，単語音声区間の全区分の予測係数 a (実際に用いるのは b) の系列を標準パターンとする．

[音韻単位標準パターン]　学習サンプルを線形予測分析したのち，次のような12種類の音韻の区間にセグメンテーションする．(学習機能[2]による)

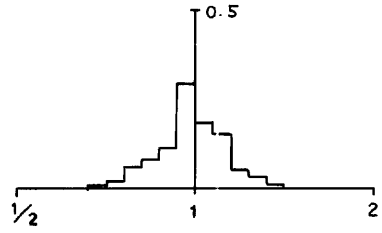
/a/ (2種)，/i/ (2種)，/u/ (2種)，/e/，/o/，/n/，/N/，/g/，/r/

10数字の学習サンプルから共通の音韻区間を抽出し，それらの区間の相関係数の平均値から予測係数を求めて，各音韻の標準パターンを作る．次に学習サンプルの各区分に，該当する音韻の標準パターンをあてはめて，各数字の標準パターンを作成する．単語内の，パワーがある一定レベルよりも低い区分は，雑音，無音または無声子音区間と見なして，そのような区分について平均した相関係数から求めた値をあてはめる．

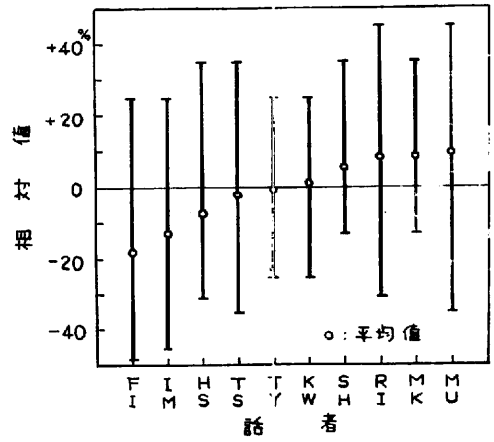
4.3. 休止区間の検出　単語内にパワーが一定レベルよりも低い区分が連続してある数以上含まれる場合は，休止区間を含む単語であると見なし，

休止区間を含む入力サンプルと含まない標準パターン，および休止区間を含まない入力サンプルと含む標準パターンの距離は，$D(k)=\infty$ とおくという条件を考える．この条件を付加した場合と，付加しない場合の認識結果の比較を行う．

4.4. 音声サンプルの長さの分布　認識に用いる全入力サンプルについて，入力サンプルと同じ話者の同じ単語の標準パターンの長さを1としたときの入力サンプルの長さの分布を調べた．図4(a)に10名の話者について累積した分布を，図4(b)に各話者の平均の長さと分布範囲を示す．全話者を通じて，入力サンプルの長さは，標準パターンの長さの0.5から1.5倍までの間に分布しており，話者別に見ると，最も広がりの小さい話者はMKとSH，最も広がりの大きい話者はFIである．FIとIMは，入力サンプルの短くなる傾向が最も著しい．入力サンプルが平均的に標準パターンよりも長くなる話者と短くなる話者は半々で，はっきりした傾向はないが，平均的には短くなる傾向にある．これは，標準パターン作成時には10単語のみを発声するのに対し，入力サンプルは100単語を次々に発声するためであろう．



(a) 10名の累積分布



(b) 各話者の分布

図4. 標準パターンに対する入力サンプルの分布

3.2節で述べたように，方法IIIにおいては，$|M-N(k)| > N_w$ のときには，対応づけが十分になされないので，入力サンプルの伸縮がはげしい場合は，整合窓の幅をかなり大きくしなければならないことになる．方法I～IIでは，m－n平面の対角線の両側に整合窓を設けるので，一たん全体の長さに関する線形な正規化を行ってから，整合窓の範囲内で非線形な正規化を行うことに相当し，上述のような条件は生じない．

4.5. 対応づけるサンプルの長さの制限　上述の検討により，入力サンプルの長さMは，

$$N(k)/2 < M < 2 \cdot N(k)$$

の範囲内に収まることがわかったので，ここでは方法I～IIIにおいて，認識実験の際に，全体としての伸縮率が1/2以下あるいは2以上となる場合は $D(k)=\infty$ とおくという条件を付加することにした．

方法IVの場合は，その定義からもわかるように，長さの比が1/2以下または2倍以上のサンプルの対応づけは行われない．

5. 実験結果

5.1. 単語単位標準パターンによる実験結果

5.1.1. 総合結果　整合窓の幅に関するパラメータ N_w および r を変えたときの方法I～IIIの結果，および方法IVの結果を図5に示す．休止区間の検出は行っていない．横軸の N_w と r は，標準パターンと入力サンプルの長さの，10名の話者の全サンプルに関

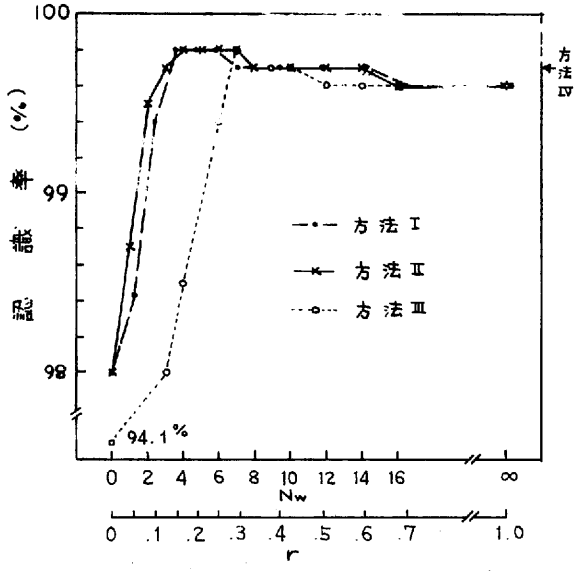


図5. 単語単位標準パターンによる認識結果(休止区間検出なし)

する平均値（23.6区分）によって換算して示してある．この結果から次のことがわかる．

(1) 認識率のピーク値は99.8％で，方法I～IIIで共通である．

(2) 整合窓を用いることにより，用いない場合（$N_w=\infty$, $r=1$）よりも誤認識が1/2に減少する．

(3) 方法IIIで$N_w=0$の場合は，時間軸の正規化を全く行わずに，パターンの先頭から順に対応づけることに相当し，方法Iで$r=0$の場合は，パターンの全長による線形の時間軸正規化のみを行うことに相当する．従って両者の結果の比較から，線形正規化によって正規化前よりも認識率が4％向上し，誤認識数がほぼ1/3に減少することがわかる．

(4) 方法IとIIは類似した結果を示し，方法IIIよりも方法IあるいはIIの方が安定した結果を示す．

(5) 方法IIIでは，ピーク値をとるNwは7区分であるが，方法IおよびIIでは4区分程度でよい．これは前述のように，方法IIIではパターン全体としての伸縮を，整合窓の幅にもたせなければならないためである．

(6) 方法IVによる認識率は99.7％で，他の方法のピーク値よりも0.1％低いが，ほぼ同程度の結果が得られると見なせる．

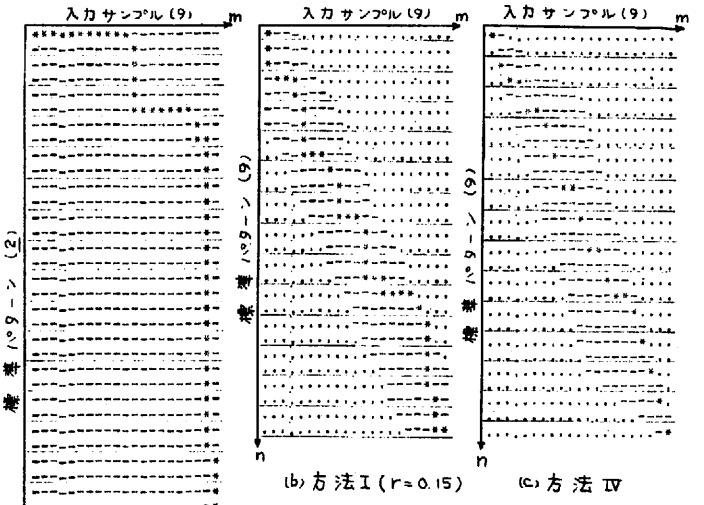


図6. 対応づけの結果の例
（話者FIによる"9"）

5.1.2. 整合窓の効果の背景　整合窓の設定法によって認識率が変化する背景を調べるため，整合窓を設定しないときに誤認識を生じ，適切な窓を設定したときに正しく認識される2サンプル（話者FIが"9"と発声したもの）について，最も小さな距離を示す標準パターンとの対応づけの結果を調べてみた．図6は，整合窓を用いない場合（"2"に誤認識）と，方法Iで$r=0.15$とした場合（正しく認識），および方法IVの場合（正しく認識）の結果の例を示したものである．表示の都合上，これまでの図（図1など）とは標準パターンの時間軸の方向が逆になっている．図中，－は整合窓の範囲を示し，＊は対応づけの系列を示す．

この図の結果から，整合窓を設けない場合には，標準パターンの1区分が入力サンプルの極めて多数の区分と対応づけられたり，逆に入力サンプルの1区分が標準パターンの極めて多数の区分と対応づけられたりし，これが同じ単語の標準パターンとの距離よりも小さな距離を生ずる原因となっていることがわかる．単語音声には，短い単語でも過渡的に極めて多様なスペクトルを含むため，このような極端な対応づけを生ずるものと考えられる．

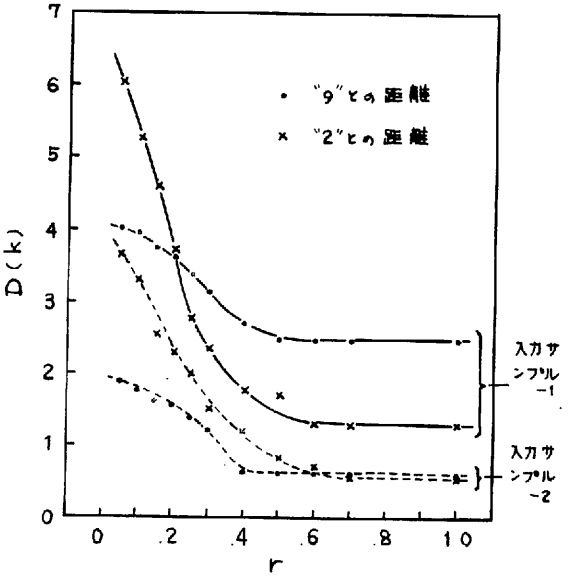


図7. 入力サンプル("9")と標準パターン("2","9")との距離

図7には，上で調べた2種の"9"の入力サンプルについて，方法Iを用いてrの値を変えたときの，"9"および"2"の標準パターンとの距離の変化を示したものである．入力サンプルと同じ"9"の標準パターンに対しても，異なる"2"の標準パターンに対しても，整合窓の拡大とともに単調に距離が減少する

が，入力サンプルと同じ"9"に対する場合よりも，異なる"2"に対する場合の対応距離の値の変化が著しい．そして，rの値が大きくなって，"2"に対する値が"9"に対する値よりも小さくなったときに，誤認識を生じている．rが0.7よりも大きい範囲では，距離の値が一定になっている．これは，r=0.7程度でm-n平面の90%以上が整合窓の範囲に入ってしまうためである．

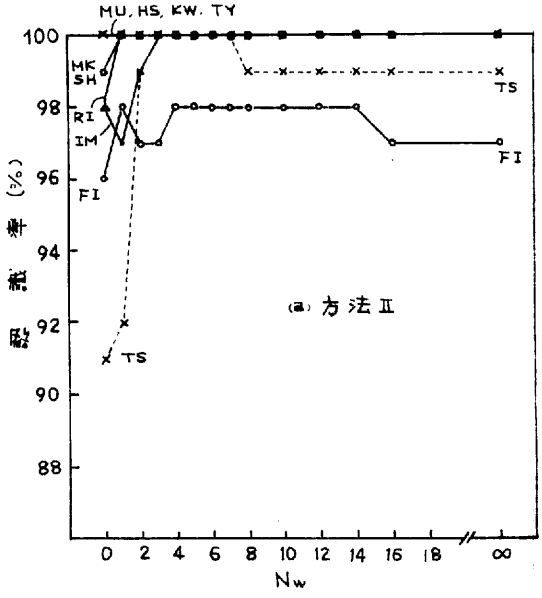
図6および図7の結果は，整合窓が大きすぎると異なる単語の標準パターンとの極端な対応づけを生じて，これが誤認識の原因となることを示している．

5.1.3 話者別・単語別結果　図8に，方法IIとIIIを用いたときの話者別の認識率を示す．方法Iは，方法IIと類似した結果を示す．これらの結果から次のことがわかる．

(1) $N_w=0$ および $r=0$ 付近の結果を2つの方法について比較すると，方法IIIのときに誤認識の多い話者は，多い方から順にFI, IM, TS, SH, MK, KW, …であるが，このうちTS以外は，方法IIを用いることにより，大幅に誤認識が減少している．TSの結果は方法IIとIIIでほとんど変わらない．方法IIを用いたとき，すなわち線形伸縮を施したときに認識率の改善の著しい話者はFIとIMであるが，図4(b)から，この2人は，入力サンプルと標準パターンの長さの比とばらつきが最も大きい話者に相当していることがわかる．

(2) 方法IIIにおいて，話者FIの認識率がピーク値に達するのは $7 \leq N_w \leq 10$，話者IMの認識率がピーク値に達するのは $4 \leq N_w$ の範囲であり，この下限の値は，それぞれの話者のサンプルの平均の長さに，未知サンプルと標準パターンの長さの変化の割合い(平均値)を乗じた値に近い．このことは，方法IIIでは線形伸縮の影響を整合窓でカバーしなければならないことを示している．

(3) 方法IIで認識率に対する整合窓の幅の効果が最も大きい話者TSは，時間軸の非線形伸縮の最も著しい話者と考えることができる．



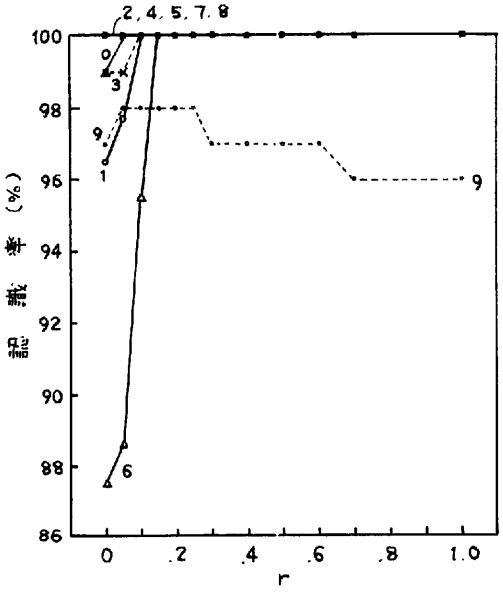


図9. 単語別認識結果(方法I)

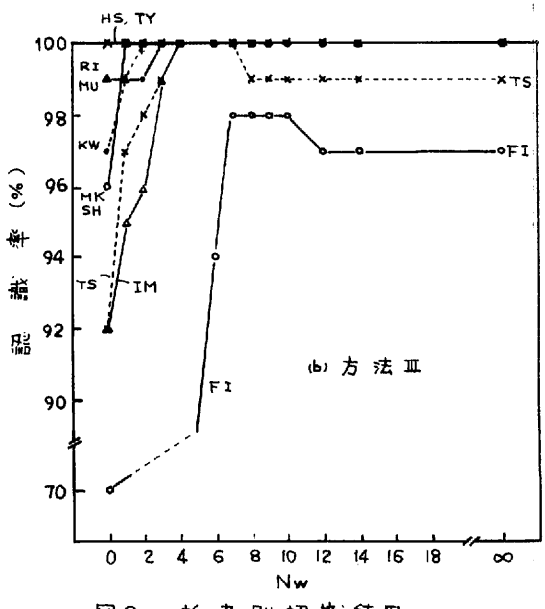


図8. 話者別認識結果

図9に，方法Ⅰを用いたときの単語別の認識率を示す．方法Ⅰ～Ⅲにおいて，整合窓の幅を最適にしたときの誤認識数はいずれも2サンプルで，どの方法の場合も"9"を"4"に誤っている．方法ⅠおよびⅡで，整合窓の幅が狭いときに誤りの多い単語は"6"と"1"で，無声化しかけたサンプルが関係している．整合窓が広すぎるときに認識率が低下するのは"9"で，この単語は過渡的に種種のスペクトルを含むため，窓の幅が広いと，他の単語と対応づけられやすいものと考えられる．方法Ⅳにおける誤認識は，いずれも無声化していない"6"を"4"と誤ったものである．

5.1.4. 休止区間検出の効果　方法ⅡとⅣについて，休止区間検出機能を付加した場合と，付加しない場合の認識率の比較を行った．図10に結果を示す．方法Ⅱでは，休止区間検出機能を付加することにより，Nw=0 付近の認識率は大きく改善され，必要な整合窓の幅もやや小さくてすて安定性が増すが，認識率のピーク値は変わらない．

方法Ⅳでは，休止区間検出を行うことにより，認識率が99.7%から100%に改善された．方法Ⅰと Ⅲの場合は，方法Ⅱと同様に，休止区間検出を行っても認識率のピーク値は変わらないことが誤認識の結果からわかるので，以上を総合すると，休止区間検出機能を付加した時は，方法Ⅳが最も良い認識率を示すことになる．

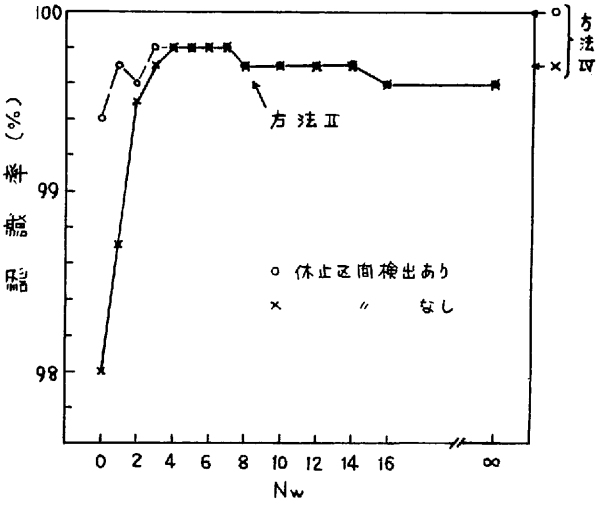


図10. 休止区間検出の効果

整合窓		方法Ⅱ		方法Ⅳ	
休止区間検出		あり	なし	あり	なし
標準パターン	音韻単位	#1	#2	#5	#6
	単語単位	#3	#4	#7	#8

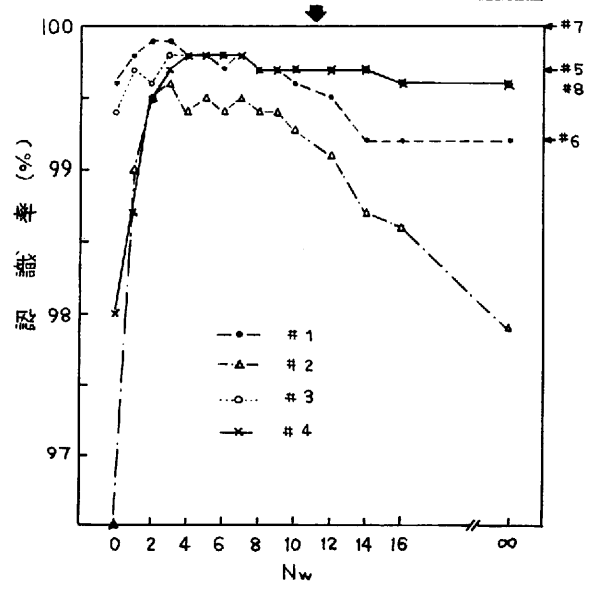


図11. 各条件の認識結果

5.2. 音韻単位標準パターンによる実験結果

5.2.1. 総合結果　前節までの検討により，方法Ⅲに比して方法ⅠあるいはⅡの方が安定した良い結果を示し，しかも方法ⅠとⅡは極めて類似した結果を示すことがわかった．また休止区間検出機能を付加した場合は，方法Ⅳが最も良い認識率を示すことがわかった．

このためここでは，方法Ⅱと方法Ⅳを用いて，音韻単位標準パターンを用いる方法により認識実験を行った．休止区間検出を行う場合と行わない場合の2通りの検討を行った．前述の単語単位標準パターンを用いる場合の結果とあわせて，図11に実験結果を示す．

この結果から，方法Ⅱに関して次のことがわかる．

(1) 音韻単位標準パターンを用いる場合は，単語単位標準パターンを用いる場合に比べて，休止区間検出の効果が著しく，休止区間検出を行わない場合に比べて，誤りがほぼ1/2に低下する．

(2) 休止区間検出を行わない場合は，音韻単位標準パターンを用いるときよりも，単語単位標準パターンを用いるときの方が認識率が良いが，休止区間検出を行う場合については，

$4 \leq N_w \leq 9$ の範囲では両者の差はなく，$N_w \leq 3$ の範囲では音韻単位標準パターンを用いる場合の方が良い認識率を示す．音韻単位標準パターンを用いて，$2 \leq N_w \leq 3$ としたときの認識率は，99.9％である．

(3) 単語単位標準パターンの場合は，N_w を大きくしたときの認識率の低下よりも，N_w を小さくしたときの認識率の低下の方が大きいが，音韻単位標準パターンの場合は，N_w を大きくしたときの認識率の低下が著しく，休止区間検出を行うときは，N_w を小さくしたときの認識率の低下よりも，大きくしたときの低下の方が大きい．すなわち，単語単位標準パターンの場合は，整合窓の幅を大きくとりすぎてもその影響は比較的少ないが，音韻単位標準パターンの場合は，認識率の大きな低下をもたらす．

(4) 音韻単位標準パターンの場合と単語単位標準パターンの場合の，認識率のピーク付近での安定性の比較を行うと，休止区間検出機能を付加したときに 99.8％以上の認識率をもたらす N_w の範囲は，前者は $1 \leq N_w \leq 5$，後者は $3 \leq N_w \leq 7$ で，その幅は変わらず，99.7％以上の認識率をもたらす N_w の範囲は，前者は $1 \leq N_w \leq 9$，後者は $3 \leq N_w \leq 14$ で，前者の方がやや狭いが，その差は小さい．

次に方法IVに関して次のことがわかる．音韻単位標準パターンを用いる場合の認識率は，休止区間検出を行うとき 99.7％，行わないとき99.2％で，単語単位標準パターンを用いる場合に比べて，休止区間検出を行うとき 0.3％，行わないとき 0.5％認識率が低下する．

以上を総合すると，音韻単位標準パターンを用いる場合は，方法IIで休止区間検出機能を付加し，$2 \leq N_w \leq 3$ とするのが最もよい．休止区間検出機能を付加し，最適な値付近の N_w を用いるときの認識率は，単語単位と音韻単位の2種の標準パターン作成方法のどちらを用いても，ピーク値，安定性共にほとんど変わらない．

5.2.2. 話者別・単語別結果　休止区間検出機能を付加した方法IIを用いたときの，話者別および単語別認識結果を調べた．その結果，N_w が小さいときに誤認識の多い話者は RI，N_w が大きいときに誤認識の多い話者は MK で，単語単位標準パターンを用いる場合の結果（図8(a)）とは，かなり傾向が異なっていることがわかった．また，単語別では，N_w が小さいときは"6"，N_w が大きいときは"9"の誤認識が多く，単語単位標準パターンを用いる場合の結果と類似していることがわかった．

6. 音韻単位標準パターンによる簡易的方法

6.1. 方法　前章までの音韻単位標準パターンによる方法では，標準パターンとして，同じ音韻が続く区間には同じスペクトルパラメータを繰返し並べたものを用いていた．従って入力サンプルの各区分との距離を並べるm-n平面にも，同じ距離の値が連続して並ぶことになり，不経済な印象を受ける．これまでに検討した方法I～IVを正確に適用するためには，これまでのような広いm-n平面を用いる方法が必要であるが，適切な簡略化を行えば，標準パターンとして同じ音韻のパラメータを繰返さずに，図12に示すような平面で対応づけを行うことが可能であろうと考えられる．

好田ら[2]はこのような平面において，各音韻の継続時間の範囲に制約を設けるとともに，入力サンプルのスペクトルの変化の比較的大きい時点から，音韻の変化時点をしぼり，高い認識率を得ている．

迫江[8]は主として文字の方向指数列を対象とした Rubber String Matching 法を提案し，音声認識への適用可能性を述べている．この方法は各音韻の継続時間の範囲と終端の範囲のほか，各音韻の重要度に対応する重みを指定するものであり，好田らによる方法に類似している．この重みを事前に最適に設定するのは，かなり難しいと思われる．

ここでは極めて簡略化された認識方法について検討する．これを［方法V］と呼ぶことにする．まずあらかじめ標準パターンとして，各音韻のスペクトルパラメータに，音韻の変化時点の情報を付け加えたものを用意しておく．つぎに標準パターンの各音韻と入力サンプルの各区分との距離を計算し，図12のように蓄える．前章までの方法では縦軸の長さは標準パターンの区分数であったが，この方法では音韻の数（$N_p(k)$ であらわす）になっており，音韻の数は10数字で平均して4程度であるから，これまでの方法に比べてほぼ1/6程度に小さくなっている．

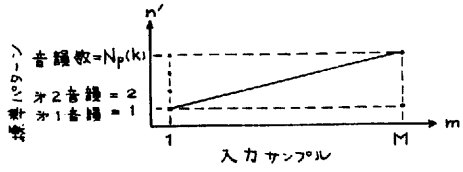


図12. 方法Vにおける対応づけの平面

標準パターンの各音韻の変化時点（区分番号）を，入力サンプルと標準パターンの長さの比で線形補正した値を求め，入力サンプルの音韻の変化時点を，その補正した時点から N_w 区分以内の点に制限する．この制限を加えても，図12の m-n′平面上の始点 (1, 1) から終点 (M, $N_p(k)$) に至る最適な伸縮関数をこれまでと同じ方法で決定して，単語全体としての距離を求めたのでは，良い認識率が得られない．これは，入力サンプルの1時点が標準パターンの極めて長い区間と対応づけられても，それが単語全体の距離の計算に直接的に含まれないためである．

そこで，標準パターンにおける同じ音韻の繰返しを擬似的に実現して，m-n′平面の縦軸を，元の標準パターンの長さと結果的に等しくなるようにするため，伸縮関数を求めるためのDPのプロセスを次のように変形する．標準パターンと入力サンプルの対応づけが進む方向は，図13に示すような2方向とする．

$$C_\ell = \begin{cases} C_{\ell-1} + (1, 0) \\ C_{\ell-1} + (1, 1) \end{cases} \qquad (15)$$

m-n′平面上で斜めに進む時は，音韻が変化することに対応するので，このときには，その1つ前の音韻に対応づけられた入力サンプルの区分のうちの中央付近で，最も距離の小さい区分における距離を，標準パターンにおいてその音韻が繰返された数だけ加えることにする．この操作は図14の例に示すように，元のm-n平面において，整合窓の範囲内に縦方向（標準パターン方向）への対応づけの進む道をつけ加えることに相当する．

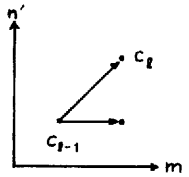


図13. 方法Vにおいて許される $C_{\ell-1}$ と C_ℓ の関係

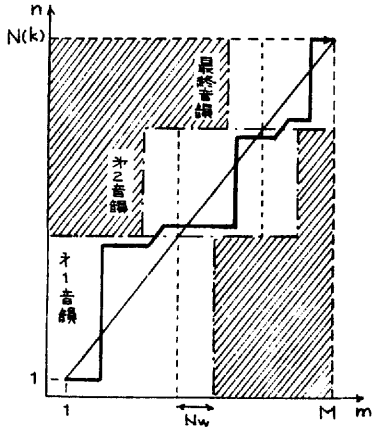


図14. 方法Vで伸縮関数の許される範囲とその例（m-n平面）

この縦方向の対応づけの位置を，前の音韻に対応づけられた入力サンプルの区分の中央付近に限定することにより，入力サンプルの不安定な過渡部が標準パターンの多数の区分に対応づけられるのを防ぐとともに，実質的な整合窓を元のm-n平面の対角線のまわりに比較的狭い幅で設定したことと同様の効果があると考えられる．

6.2. 認識実験の結果　休止区間検出機能を付加して，10数字音声の認識実験を行った結果を図15に示す．$2 \leq N_w \leq 3$ のときの認識率は99.7%で，$N_w = 0$ すなわち線形伸縮のみの場合に比較して0.1%向上している．簡易化せずに方法II（休止区間検出あり）で認識を行う場合と比較すると，m-n′平面の縦軸（標準パターン軸）が短くなった分にほぼ比例して演算回数が減少するが，認識率も0.1%低下している．さらに比較のために好田らによる認識方法を，ここで用いた入力サンプルに適用してみたところ，無声化したサンプルも含めた場合に99.8%，除いた場合に100%の認識率が得られた．

ここで検討した方法Vの有効性を確かめるためには，多数の語彙を対象とした認識系において，経済性を含む種々の観点から定量的な検討を進める必要があり，今後更に適切な簡易化方法などについても検討してゆく必要があろう．

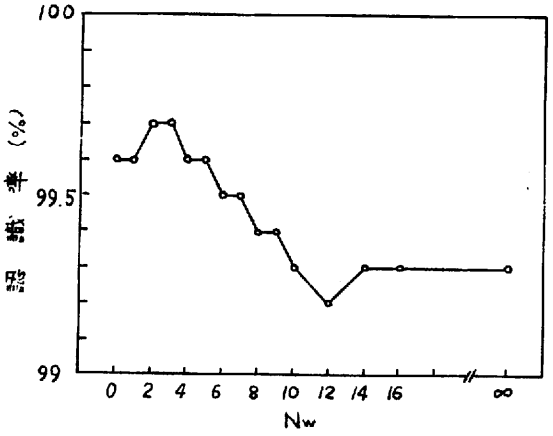


図15. 方法Vによる認識結果

7. むすび

標準パターンの作成方法も含め，単語音声認識における時間軸正規化の問題に関連した種々の検討を行った．具体的には，次のような点について検

討した.

(1) 種々の時間軸正規化マッチング法の比較
(方法Ⅰ, 方法Ⅱ, 方法Ⅲ, 方法Ⅳ)

(2) 単語単位と音韻単位標準パターンの比較

(3) 整合窓の有効性の評価とその背景の分析

(4) 休止区間検出機能の有効性の評価

(5) 音韻単位標準パターンによる簡易的方法の検討
(方法Ⅴ)

男性10名が発声した10数字音声の認識実験を行った結果, 次のようなことが明らかになった.

(1) 入力サンプルと標準パターンの長さ情報を用いて, 両者の作る平面の対角線のまわりに整合窓を設定する方法Ⅰおよび Ⅱの方が, 長さ情報を用いずに傾き1の線のまわりに整合窓を設定する方法Ⅲよりも安定で, しかも窓の幅を小さくできる. 方法Ⅰおよび方法Ⅱは一たん時間軸の線形な正規化を行ってから, 整合窓の範囲内で非線形な正規化を行うことに相当する.

(2) 休止区間検出機能は全般的に有効で, 単語単位標準パターンを用いた方法Ⅰ～Ⅲでは, 認識率のピーク値は変わらないが, 整合窓を狭くしたときの認識率が改善され, 認識の安定性が増す. 音韻単位標準パターンを用いた方法Ⅱでは, 休止区間検出機能により, 整合窓の大きさに拘らず誤認識が半分に低下する. 方法Ⅳでは, 標準パターンの作り方に無関係に, 休止区間検出機能により認識率が 0.3～0.5 % 向上する.

(3) 単語単位標準パターンを用いる場合は, 休止区間検出機能を付加した方法Ⅳ(各時点の標準パターンと入力サンプルの相対速度を 1/2 と 2 の間に制限する方法)が最も有効で, 100% の認識率が得られる. 単語単位標準パターンを用いた方法Ⅰ～Ⅲの認識率のピーク値は 99.8 % であった.

(4) 音韻単位標準パターンを用いて, 休止区間検出機能を付加した場合は, 方法Ⅱと方法Ⅳを比較すると方法Ⅱの方が有効で, 認識率のピーク値は 99.9 % であった. 方法Ⅳの認識率は 99.7% であった.

(5) 休止区間検出機能を付加した方法Ⅱでは, 標準パターンとして単語単位のものを用いても, 音韻単位のものを用いても, 認識率はほとんど変わらない. 多数語彙を対象としたときに, 本報告の10数字音声を対象とした結果がそのまま成り立つことは必ずしも期待できないが, 語彙数が多くなったときには, 個人差の学習や記憶容量の点で, 音韻単位の方が有利であるから,(7) この結果は好ましいと言える.

(6) 音韻単位標準パターンを用いる簡易的方法の試みでは, ピーク値で 99.7% の認識率が得られた. この方法は極めて高速に処理できる利点があるが, 認識結果がやや不安定で, ピーク値も簡易化しない方法の場合よりやや低く, さらに検討を進める必要がある.

(7) 整合窓の幅に関しては, どの認識方法の場合にも最適値があり, 幅が広すぎたり狭すぎたりすると, 認識方法, 標準パターン作成法, 話者, 発声内容等が関連した複雑な要因により, 認識率が低下する.

以上の検討で得られた結果は, 10数字音声を対象とした小規模な実験によるもので, 認識率の 0.1 % 程度の小さな差は必ずしも有意とは言えないが, 単語音声認識における時間軸正規化の効果や標準パターンの作成法の影響などに関して, 種々のことが明らかになった. 今後は, 認識対象語彙数やサンプル数をさらにふやして, 実験や検討を行う予定である.

[謝辞] 日頃御指導御鞭撻を頂く斎藤特別研究室長, ならびに熱心に討論して頂いた好田調査員をはじめとする研究室の諸氏に感謝の意を表する.

文献

(1) 迫江・千葉: "動的計画法を利用した音声の時間軸正規化に基づく連続単語認識", 音学誌, 27, 9, P.483 (1971)

(2) 好田・橋本・斎藤: "数字音声の機械認識系", 研実報, 21, 7, P.1361 (1972)

(3) F.Itakura: "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. ASSP-23, 1, p.67 (1975)

(4) 長島・好田: "音韻の標準パターンを用いた単語音声の認識", 春季音学講論, 4-2-7 (1976)

(5) 迫江: "傾斜制限DPマッチングによる音声認識", 春季音学講論, 1-2-15 (1974)

(6) 古井: "単語音声認識における時間軸正規化方法の検討", 春季音学講論, 3-2-16. (1976)

(7) 斎藤: "限定語彙の音声認識における諸問題", 春季音学講論, 4-2-11 (1976)

(8) 迫江: "Rubber String Matching 法による手書き文字認識", 信学研資, PRL74-20 (1974)