論　文

不特定話者・連続音声向き単語音声の識別*

坂井利之** 中川聖一**

Abstract

As a sub-system of the speech understanding system, we developed a classification method of spoken words in continuous speech for many speakers.

Speech wave is converted into a time series of short time spectra by 20-channel filter bank and is segmented into four groups: silence, unvoiced-nonfricative, unvoiced-nonplosive, and voiced group. The unvoiced groups are classified into a unit of phoneme by heuristic algorithms and voiced group by Bayes rule.

To normalize the variation of reference patterns among speakers, vowel patterns are learned by the non-supervised learning method.

The optimum matching between a just recognized phoneme string and a phoneme string of a given word in the word dictionary is performed by utilizing the phoneme similarity matrix and Dynamic Programming.

According to the results tested upon 1,500 samples of isolated digits, spoken by 20 male speakers, about 97% were correctly recognized and, in case of the system adapting for each speaker, 98% correctly recognized.

1. まえがき

人間の発声器管によって発話された音声波を，離散的な言語符号列に機械で自動的に変換する，いわゆる音声タイプの実現が難しいとされている理由は，観測される音声波が話者の個人差や調音結合などの変動要因を受けているためである[1),2)].

一方，識別対象が限定された単語音声の識別では，音韻のレベルまで識別の単位を分割せず直接単語レベルで識別するか，自然言語が持つ冗長な情報を利用することにより調音結合の問題が回避できる．この場合，話者を限定することによって高い識別率を得る手法が開発されている[3)~5)]．また，不特定話者に対する単語音声の識別でも比較的高い識別率を得る手法が開発されているが[6),7)]，識別対象となる各単語の選び方でアルゴリズムが異なる方法を採用しており一般的ではない．

したがって不特定話者の連続音声を対象とし，かつ識別する語彙数の増大をねらうならば，話者に負担をかけずに何んらかの話者標準パターンを機械に学習させる機構の導入が必要である．また処理時間および記憶容量の両面から考えれば，システム構成は音韻識別と単語識別の二段階にするのが望ましいと思われる[8)]．

この手法を採る研究が種々試みられてはいるが[9),10)]数字音声の場合ですら不特定話者に対して 97%～98% 程度の識別率が得られる方法は，未だ開発されていなかった．

われわれは，音声理解システムの研究[11)]の一環として，不特定話者の連続音声向き限定語彙単語音声の比較的高精度な識別手法を開発した．限定話者を対象として現在までに開発された手法で高い識別率を得たものは，すべて算法としてダイナミックプログラミング (Dynamic Programming：以後 DP と略記) を採用

* A Classification Method of Spoken Words in Continuous Speech for Many Speakers by Toshiyuki SAKAI and Seiichi NAKAGAWA (Faculty of Engineering, Kyoto University)

** 京都大学工学部情報工学科

している[3)~5),8),12)]．これは時系列の長さに差のある入力パターンと参照パターンの照合に対して，時間の向きに従って連続的に処理するものである．われわれも音韻系列間の照合の際にDPの手法を用いる新しい手法を提案する．

この報告で，われわれはシステムの概要を述べ，次に音声認識のための特徴パラメータで最も妥当と言われている[13)]音声スペクトル情報からの音韻識別法，さらに音韻系列からの単語識別方法について述べる．この方法に対して，比較的多量の音声試料で性能を評価した実験結果も述べてある．

2. システムの構成

Fig. 1 にシステムの構成図を示す．音韻識別の処理対象は，折点周波数 1.6 kHz 6 dB/oct の高域強調フィルタによってプリエンファシスされた音声波を，フィルタバンクに通しサンプリングした後，A/D 変換器によりディジタル信号に変換して得られる短時間スペクトル（以後フレームと呼ぶ）である．フィルタバンクは 20 チャネル・1/4 オクターブであり，音声波の 200 Hz から 6,400 Hz を分析する．サンプリング間隔は 10 ms, A/D 変換器の精度は符号＋9ビットである．

まず初めに，音声波のパワーやスペクトルの偏り等の特徴により無音，無声非摩擦音，無声非破裂音および有声音の4グループに大分類する．無声子音グループに分類された音声区間（以後セグメントと呼ぶ）は，その継続時間長や前セグメントが無音性かどうか，スペクトルの偏りや時間的変化の状況等の物理的なパラメータにより次の5グループに細分類される．①無声破裂音 /p/（/p/, /t/, /k/ 間は分類せずまとめて /p/ と記す），②気音 /h/，③無声破擦音 /c/，④無声摩擦音 /s/，およびそのいずれでもない（棄却される：Fig. 1 の A）⑤有声音．

有声音と判定されたセグメントは，母音および有声子音識別プロセスで処理される．母音の識別では，話者による音韻パターンの変動を考慮する必要がある．ここでは，話者に特別の負担をかけないでシステムが自動的に話者ごとの各母音の標準パターンを学習していく，いわゆる教師なしの学習を行う（Fig. 1 の B）．母音の識別プロセスで棄却された部分（Fig. 1 の C）と長い過渡的な音声区間（Fig. 1 の D）およびこれと独立にパワーの時間的変化の検出（Fig. 1 の E）によって得られた有声子音候補区間は，有声子音識別プロセスで処理する．このとき用いる有声子音の標準パターンは，母音パターンの学習結果を基に推定される（Fig. 1 の F）．

以上のようにして得られる音韻列は，若干の音韻論的規則により修正が加えられて音韻レベルでの識別の最終出力が決まり，次の単語識別部への入力となる．単語識別部は，単語辞書で与えられる音韻列と識別された音韻列とを，音韻間類似度を使って DP マッチングを行い最適な照合を得る単語（辞書）を識別結果とする．

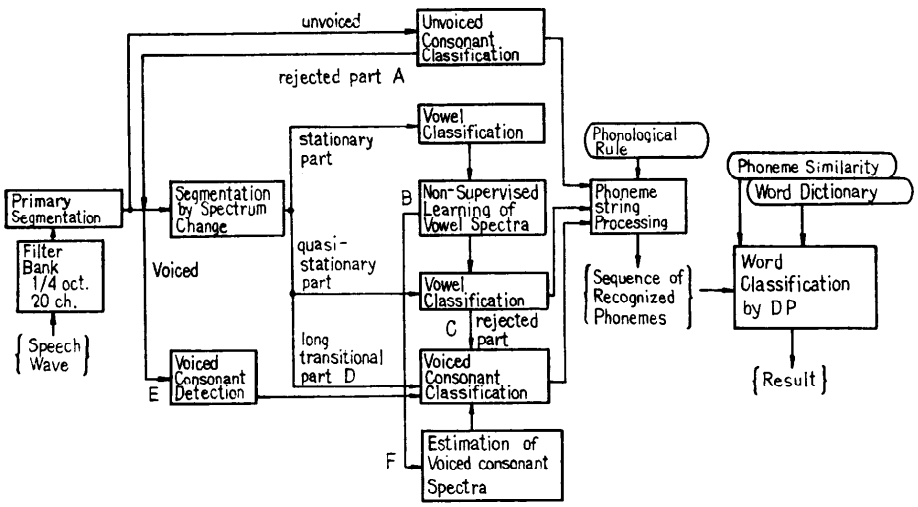

Fig. 1 Block diagram of spoken words recognition.

3. 音韻識別

3.1 セグメンテーション

3.1.1 第一次セグメンテーション

入力音声を無音，無声非摩擦音，無声非破裂音および有声音の4グループに分類することは既に述べた．時間的に連続して同じグループに分類されたフレームは接合し，若干の修正を加えてセグメント（音韻に相当する音声区間）系列を得る．ここで用いる第nフレームに関する特徴パラメータは，時間的変化に対して同一セグメント内では比較的安定な次の3つである．ただし，$\boldsymbol{d}(d_1, d_2, \cdots, d_{20})$はスペクトル（フィルタからの出力）の値，$\boldsymbol{x}(x_1, x_2, \cdots, x_{20})$は$\boldsymbol{d}$を1に正規化（$x_i = d_i/(d_1^2+d_2^2+\cdots+d_{20}^2)^{1/2}$）したものとする．

AMP_n＝第nフレームのパワー：
すなわち $(d_1^2+d_2^2+\cdots+d_{20}^2)^{1/2}$

$CNL_n = \sum_{i=1}^{j} x_i > 0.5$ なる最小のチャネル番号 j：
スペクトルの低域への偏りの目安

$MOM_n = x_{18}+2\cdot x_{19}+3\cdot x_{20}$：
高域スペクトルのモーメント

また，1つのセグメント内でのそれぞれの平均をAMP, CNL, MOMと記し次節で用いる．これらを用いたセグメンテーションの手順を次に示す．

1. もし$(AMP_n<15)\wedge(CNL_n\leqq 7)\vee(AMP_n)<10$なら無音．
2. もし$(AMP_n\geq 15)\wedge(CNL_n\leq 5)\vee(MOM_n)<1.0$なら有声音．
3. その他なら無声子音グループ．
4. 隣接フレームが同じグループなら接合する．
5. もし無声子音グループのセグメント内で，$CNL_i\geq 15$なるフレームiが存在すれば，無声非破裂音グループ，存在しなければ無声非摩擦音グループ．

上の手順1～5を繰り返えすことによってセグメントの系列が得られる．

3.1.2 有声音グループ内のセグメンテーション（第二次セグメンテーション）

有声音部として分類されたセグメントは，定常部と過渡音部に分けられる．前後のフレームとの間でスペクトルの時間的変化量を計算し，フレームごとに定常性の判定がなされ，それより定常部が取り出される．定常かまたは非定常かの判定は，統計的判定法を用いる[14]．

* $\min(a, b, \cdots, c)$は$a, b, \cdots, c$のうち最小値を，maxは最大値を示す（以後同様）．

スペクトルの変化量を求めるために，定常部の検出ではスペクトル$\boldsymbol{d}$を対数変換したもの（学習用に用いる音声サンプルの抽出区間は，音声パワーのレベル変化も小さいところが望ましいので，スペクトルを正規化しないで時間的変化量を求める．以後この区間を特に学習用定常部と呼ぶ．）と正規化されたスペクトル$\boldsymbol{x}$（母音と撥音の検出用）とを用いる．有声音部で定常部以外はすべて過渡音部とみなす．

3.2 無声子音の識別

ここで用いる新たなパラメータ$DUR, PRAMP, SL, AMPMAX, AMPMIN$を定義しておく．

DUR＝識別対象となっているセグメント内のフレーム数．

$PRAMP$＝識別対象となっているセグメントの直前の3フレームのパワーの平均値．

$SL = \min(1.0, 1.3-PRAMP/50)$*

音声のパワー値による無音性の尤度（SLが大きいほど無音性大）．

$AMPMAX$＝識別対象となっているセグメント内の最大パワー値．

$AMPMIN = AMPMAX$を持つフレームからセグメントの終りまでの最小パワー値．

無声子音の識別には，これらのパラメータを用いて各々の音韻カテゴリ（音韻グループ）の尤度を1.0を基準にして求め，最大の尤度を持つカテゴリを識別結果とする．ただし有声音の尤度が最大になった場合は，無声子音の識別は棄却され有声音処理プロセスに進む．以下各カテゴリの尤度を求める手法を述べる（これらは，多量の音声試料の分析によりヒューリスティックに求めたものであり，音声の物理的性質との関連は，文献22)に発表予定）．〔 〕の演算は，if条件が満たされた時に限り実行されることを示す．

（1） 無声非摩擦音セグメントの識別

有声音の尤度

$$=\frac{AMP}{200}\times\frac{1}{MOM}\times\left[\frac{DUR}{12};\ \text{if } DUR\geq 15\right]$$
$$\times\{1.2;\ \text{if } (AMP_1=AMPMAX)\vee(AMP_{DUR}=AMPMAX)\}\times\left[\frac{8}{CNL};\ \text{if } CNL\geq 8\right].$$

無声破裂音 /p/ の尤度

$$=SL\times\{1.5;\ \text{if } DUR\leq 4\}$$
$$\times\left[\left(\frac{AMPMAX}{AMP_1}+\frac{AMPMAX}{AMPMIN}+\frac{AMP_{DUR+1}}{AMPMIN}\right)\times\frac{1}{4.5};\ \text{if } DUR\geq 4\right]$$

$$\times\left[\frac{8}{DUR};\ \text{if } 4<DUR\leq 7\right].$$

気音 /h/ の尤度

$$=1.0\times\left[\frac{DUR}{8};\ \text{if } (DUR>8)\wedge(AMP\leq 150)\right]$$
$$\times\left[\frac{DUR}{4};\ \text{if } DUR<4\right]$$
$$\times\left[\frac{CNL}{9};\ \text{if } CNL>9\right]$$
$$\times\left[\frac{50}{AMP+10};\ \text{if } (AMP\leq 80)\right.$$
$$\left.\wedge(AMPMAX<AMP\times 1.3)\right].$$

（2） 無声非破裂音セグメントの識別

気音 /h/ の尤度

$$=(3.7-MOM)\times\left[\frac{60}{AMP+10};\ \text{if } AMP\leq 40\right].$$

無声破擦音 /c/ の尤度

$$=\min(SL\times 1.2, 1.0)$$
$$\times\{1.2;\ \text{if } 2.5<MOM<3.4\}$$
$$\times\{1.2;\ \text{if } DUR\geq 7\}$$
$$\times\left[\left(\frac{AMPMAX}{AMP_{DUR/2-2}}+\frac{AMPMAX}{AMP_{DUR/2+2}}\right)\right.$$
$$\left.\times\frac{1}{2.4};\ \text{if } DUR\geq 8\right].$$

無声摩擦音 /s/ の尤度

$$=(MOM-2.3)\times\left[\frac{DUR}{8};\ \text{if } DUR<8\right]$$
$$\times\left[\frac{DUR}{12};\ \text{if } DUR>12\right].$$

この手法は，各カテゴリが持つ時間的変化も考慮に入れた特徴をそれぞれ独立に評価し，それらを相乗していくもので，いつでも第1，第2等の候補が存在しているので，トリー状に分類していく手法のような固さがなく処理の柔軟性が高いと思われる．

3.3 母音および撥音の標準パターンの学習と識別

特徴パラメータ空間上に分布する各カテゴリを分類する方法は種々知られている[15]．これらの手法が音韻識別においては，どの程度識別率に差があるかを検討するために，10 名の男性話者の音声試料 2,450 単語中の母音 4,612 サンプルの識別実験を行った．

識別手法として，市街距離 (Chebychev norm)．ユークリッド距離，線形判別関数（重み係数を判別分析より求める）およびベイズの識別規則に基づく2次判別関数の4通りで識別実験をした．結果を **Table 1** に示す．ただし話者別2次判別関数の場合は，平均ベクトルだけ個人ごとに準備した．この結果より2次判別関数による結果が一番良いことがわかる．そこで，われわれは多少計算時間が他の識別法よりかかるが2次判別関数を用いることにした．

Table 1 Results of vowel recognition by various classification methods

method	cityblock distance	Euclidian distance	linear discriminant function	quadratic discriminant function
speaker-independent (common)	86.3%	87.4%	91.8%	93.0%
speaker-dependent (individual)	91.8%	92.2%	95.0%	95.0%

Table 2 Results of vowel recognition by quadratic discriminant functions in parts used for standard samples extraction

method	standard pattern		error ratio					
	mean	covariance	*a*	*i*	*u*	*e*	*o*	total
(a)	common	common	0.0%	9.1%	6.8%	7.6%	5.0%	4.7%
(b)	individual	common	0.4	1.5	4.0	0.0	2.0	1.5
(c)	individual	individual	0.4	0.8	1.1	0.0	0.0	0.5

この場合，標準パターンとしてスペクトルの平均ベクトルとその共分散行列が必要となる．予備実験として，5名の男性話者が通常の速度で発声した算術文80文章の音声試料から視察によって抽出した定常母音区間を次の方法で識別した．(a)全話者共通の平均ベクトルと共分散行列を用いる場合．(b)話者別の平均ベクトルと全話者共通の共分散行列を用いる場合．(c)話者別の平均ベクトルと共分散行列を用いる場合．**Table 2** がその結果であり，誤り率は(a)と(b)でかなり差がある．(b)と(c)にも差が見られるが，一般に共分散行列の学習には多数のサンプルが必要であり簡単ではない．そこで，ここでは全話者共通の平均ベクトルを初期値として用い，入力された音声から個人用の平均ベクトルを自動的に学習させて(Table 2の(b)に近づけて）識別率を徐々に向上させていく教師なしの学習を行う．学習用定常区間は，コンテキストの影響が比較的少なく，全話者共通の平均ベクトルと共分散行列 ($\boldsymbol{U}_i, \boldsymbol{\Sigma}_i$) を用いても，ほぼ確実に識別できる．

学習用定常区間として検出され，ある母音（以下の説明では撥音も母音と同じ扱いにする）と判定された入力サンプルは，その母音の標準パターンの更新データとして用いられる．例として話者 j の母音 i（平均ベクトルを $\boldsymbol{U}_i^j$ とする）の学習法を説明する．なお母音と有声子音の識別では，スペクトル $\boldsymbol{x}$ を対数変

換して得られる $\boldsymbol{U}(U_i=\log x_i)$ を用いる．

いま母音 i と判定されたサンプル $\boldsymbol{U}_1,\cdots,\boldsymbol{U}_N$ (N は学習サンプル数) が与えられると，学習後の平均ベクトル $\boldsymbol{U}_{i,N}{}^j$ を正規パターンの平均学習法[16]により次式で計算する．

$$\boldsymbol{U}_{i,N}{}^j=\frac{1}{\alpha+N}(\alpha\cdot\boldsymbol{U}_i+N\langle\boldsymbol{u}\rangle)$$

$$=\frac{1}{\alpha+N}\{(\alpha+N-1)\boldsymbol{U}_{i,N-1}{}^j+\boldsymbol{U}_N\}$$

$$\langle\boldsymbol{u}\rangle=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{U}_k$$

ここで $\boldsymbol{U}_i=\boldsymbol{U}_{i,0}{}^j$ は，初期値を表わし全話者共通の平均ベクトルを用いる．本実験では $\alpha=10$ とした．実際に母音 /e/ のスペクトルパターンを学習した結果を **Fig. 2** に示す．図からも判るように基本周波数の影響の大きい低周波数領域では，学習させると 3, 5 チャネルにピークが出て学習効果が目立つ．この際，学習誤りを極力避けるために，信頼度が一定値以上でないものは学習サンプルに使用しないことはもちろんである．

次に学習後の標準パターンを用いて，定常区間で母音の識別をする．この時一つの定常区間に学習用定常区間が 2 ケ所以上あり，しかもその識別結果が異なる場合は，その中間点に音韻の境界を設定しセグメンテーションの誤りを防ぐ（認識結果のセグメンテーションへのフィードバック）．以後同一話者の発声の終了まで，この学習と識別の操作を並行して繰り返す．

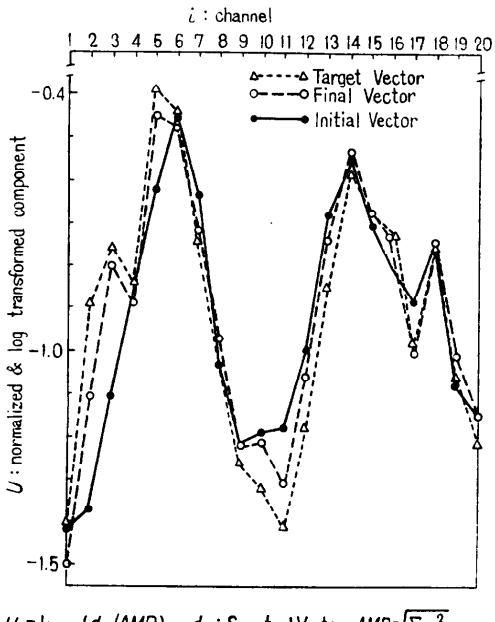

Fig. 2 Learning of mean vector of vowel /e/.

3.4 有声子音の検出と識別

音声中の有声子音候補部分は，次の 3 通りで検出される．(a)セグメンテーションの結果，過渡音部が 40 ms 以上続く部分．(b)母音の識別プロセスで棄却された部分．(c)これらとは別に，エネルギーレベルに顕著な谷がある部分．

有声子音 (/m/, /n/, /η/, /b/, /d/, /g/, /γ/, /z/) の識別は，母音の識別と同じく 2 次判別関数でおこなっている．特に有声子音では話者によるパターンの変動が大きい[17]ために，なんらかの方法でこの変動を取り除く必要がある．また母音と有声子音の交互作用も無視できない[17]．われわれは，有声子音 l の標準パターン $\boldsymbol{U}_i{}^j$ を，学習された母音パターン $\boldsymbol{U}_{i_l,N}{}^j$ (有声子音 l によって決まる母音 i_l) を使って推定する方法を採っている[21]が，詳細は別の機会にゆずる．

3.5 音韻書き替え規則による識別結果の修正

以上のようにして得られる音韻識別の結果は，各セグメントについて，決定音韻 (第一候補) だけでなく，第二候補の音韻および第一候補の信頼度 (第一候補が確実なら 1，第一候補と第二候補の信頼度が同じなら 0) およびセグメントの継続時間長からなっている．これらの音韻列に対して，いくつかの音韻書き替え規則によって識別誤りを修正する．このとき半母音 /y/ と /w/ は，書き替え規則によって識別される．無音区間は，その継続長の短い順より /・/ (無声破裂音の閉止区間に相当)，/-/ (促音に相当) および /// (句の境界に相当) の 3 種類に分類される．こうして音韻識別部の最終出力が得られ，単語識別部への入力とな

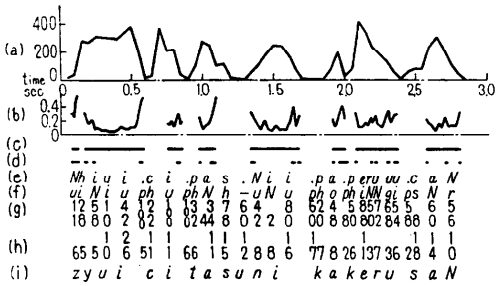

Fig. 3 An example of phoneme recognition result. (a): power of input speech (b): degree of spectra change (c): voiced part (d): transitional part (e): first candidate (f): second candidate (g): confident degree of first candidate (×100) (h): segment duration (×10 ms) (i): input speech (arithmetic expression: 11+2×3)

る．

Fig. 3（前頁参照）は算術文 11+2×3（ジュウ イチ タス ニ カケル サン と発声）の音韻識別結果を示している．（c）欄のスペクトル変化量は，**3.1.2** で述べた x の変化量を示す．タス (/tasu/) の /u/ は無声化されて識別されていない．5名の男性話者が普通の速度で発声した算術文 80 文章（560 単語）の音声試料の音韻識別率は，母音と撥音，半母音，有声子音，無声子音の順に平均としてそれぞれ 86%，30%，34%，81% であった．

4. 音韻列からの単語識別

4.1 音韻間類似度

音韻識別は主に 20 次元スペクトルの統計量（平均と共分散）に基づいて行っている．したがって，識別における誤りはこの統計量によって表わされる統計的分布が似かよっているために生じると考えられる．これらの誤りは次の3つに分類される．（a）音韻のラベル付けの誤り．（b）余分な音韻の挿入誤り．（c）音韻の脱落誤り．これらの誤りに対処するため，任意の2つの音韻間の類似度をこれらの2つの音韻のスペクトルの統計的分布の距離から求めておき，これによって音韻列間の一致または相異の数値的評価を行う．

音韻 i と j のスペクトルの平均と共分散をそれぞれ $(\boldsymbol{m}_i, \boldsymbol{\Sigma}_i), (\boldsymbol{m}_j, \boldsymbol{\Sigma}_j)$ とすると2つの分布間の距離を，次式で与えられる Bhattacharyya の距離 $B(i, j)$ で定義する．

$$B(i,j)=\frac{1}{8}(\boldsymbol{m}_i-\boldsymbol{m}_j)^t\left(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_j)+\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j|}{2|\boldsymbol{\Sigma}_i|^{1/2}\cdot|\boldsymbol{\Sigma}_j|^{1/2}}\right)$$

この距離が大きくなる程ベイズの識別規則による i と j の間の識別誤りは小さくなる関係がある[18]．この $B(i, j)$ を無声子音を除く音韻のすべてのペアに対して求める．ただし音韻 i が母音で，j が有声子音の場合は，この距離を若干修正する（われわれの音韻識別の方法では，母音が有声子音に識別されることはその逆より少ないための補正）．この $B(i, j)$ から音韻 i と j の類似度 $S(i, j)$ を0から100の数値の間になるように線型変換して求める．また音韻 i または j が無声子音か無音の場合は，識別の際にスペクトルの平均や共分散を用いていないので，音韻識別結果を参考にして適当な類似度を設けた．

Table 3 に音韻間類似度の例を示す．"In" の欄が単語辞書（後述）の音韻に，"Out" の欄が識別された音韻に対応する．音韻表現 /*/ は，無声破裂音が無音と対応づけられるのを防ぐための擬似音韻である．

4.2 単語辞書

単語の識別は，識別された音韻列に最も良く適合する単語を見出すことに相当する．そのため各単語に対して，ほぼローマ字書きの音韻列からなる単語辞書を用意する．さて音韻間類似度は，音韻間同志の一般的な類似度を示すものであった．一方単語中の一部の音韻で，そのコンテキストによって特有の変形を受け（調音結合）脱落や他の音韻への誤りの起こり易いものがある．これに対処するために辞書の音韻表現に "$I/k(c)$" という形式で表現される I（主音韻と呼ぶ）の副音韻 k とその重み係数 $c(0\leq c\leq 1)$ を導入する．音韻識別部で識別された第 j セグメント（第一候補音韻を J，第二候補音韻を l，第一候補音韻の信頼度を p とする）と，ある単語辞書の i 番目の音韻（主音韻を I，副音韻を k，副音韻の重み係数を c とする）と

Table 3 Phoneme similarity matrix

In \ Out	*a*	*i*	*u*	*e*	*o*	*m*	*n*	*η*	*p*	*t*	*k*	•	—	/	…
a	100	36	51	69	75	51	55	58	5	5	5	5	5	5	…
i	36	100	83	73	56	72	74	85	50	50	50	5	5	5	…
u	51	83	100	74	80	81	83	89	5	5	5	5	5	5	…
e	69	73	74	100	69	59	69	74	5	5	5	5	5	5	…
o	75	56	80	69	100	64	65	75	5	5	5	5	5	5	…
m	61	82	91	69	74	100	92	86	5	5	5	5	5	5	…
n	65	84	93	79	75	92	100	88	5	5	5	5	5	5	…
η	68	95	99	84	85	86	88	100	85	85	85	30	5	5	…
*	5	5	5	5	5	5	5	5	100	100	100	5	5	5	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮⋮⋮

Table 4 Example of entries in the Word Dictionary

Word	Symbol	Phoneme Representation	Maximum Duration	Minimum Duration
ichi	1	i ·/c(1.0) c i^b/c(1.0)	350ms	100ms
ni	2	n i	300	100
san	3	s a N	550	200
yon	4	y/g(0.9) o N	450	150
go	5	g o	300	100
roku	6	r/p(0.85) o ·/k(0.95) k u^b/*(1.0)	450	100
nana	7	n a/N(0.85) n/a(0.85) a/N(0.85)	550	200
hachi	8	h a/N(0.85) ·/c(1.0) c i^b/c(1.0)	500	150
kyu	9	/c(0.95) k/c(0.95) y/u(0.95) u	500	200
rei	0	r/p(0.85) e i^b/e(0.95)	400	100

が対応した場合は，類似度を次式で定義する．

$$S(I,k,c:J,l,p)=\max\begin{cases}S(I,J)\\ c\times S(k,J)\\ p\times S(I,J)+(1-p)\times S(I,l)\\ p\times c\times S(k,J)+(1-p)\times c\times S(k,l)\end{cases}$$

以後 $S(I,k,c:J,l,p)$ を簡単に $S_0(i,j)$ と書くことにする．**Table 4** に 10 数字の単語辞書の例を示す（○と△印は後述，副音韻のないものは $c=0$ に相当する）．ある単語に対応づけられた入力音韻列の継続長が表の継続時間欄の最小～最大の間にない場合は，マッチング適合度（尤度）を減じる．なお副音韻とその重み係数は，原則として単語辞書作成規則により自動的に設定される[8]．

4.3 DP による音韻列間のマッチング

単語辞書中のある単語の音韻列と識別された音韻列に対してある対応づけが与えられたとき，その対応の良さ（尤度またはスコアと呼ぶ）を評価する必要がある．ここでは「辞書の単語中の一つの音韻と識別音韻列の一つの音韻（セグメント）の対応に対しては，**4.2** で述べた方法により類似度を求める．辞書の単語中の一つの音韻が，複数個の識別音韻列と対応づけられているときは，それらの類似度を算術平均する．一単語全体にわたるスコアは，辞書の単語の各音韻の類似度を単語全体にわたって算術平均したもの（**Fig. 5** 次頁参照）．」と定義する．あらゆる可能な対応づけに対して，最も良いスコアをその単語の尤度とする．この対応づけ操作に関して，音韻識別部の能力からみて妥当と考えられる次の制限を設定した．

1. 辞書中の母音および撥音は，識別音韻列中の連続する 3 音韻以内と対応づける．
2. 辞書中の子音は，識別音韻列中の連続する 2 音韻以内と対応づける．
3. 辞書中の連続する 3 音韻は，識別音韻列の 1 音韻とは対応づけない．
4. 識別音韻列の連続する 3 音韻の継続時間長の和が 250 ms を越える場合は，辞書中の長母音を除く母音は，この 3 音韻と対応づけない．
5. 識別音韻列のある音韻の継続時間長が，100 ms を越えない場合は，辞書中の長母音はこの音韻だけとは対応づけない．
6. 辞書の音韻列が単語継続時間長の範囲外で，識別音韻列と対応づけられる場合はスコアを減じる．

なおこれらの制限は，単語辞書の表現により自由に変更できる．たとえば Table 4 の単語辞書中で○印を持つ音韻は制限 3 を取り除き，△印は制限 1 をより強い制限にする．

識別音韻列と単語辞書中のある単語の音韻列との最良な対応づけは，DP の手法を使うことによって能率よく次のような漸化式で計算される．$L(i,j)$ を辞書中のこの単語の音韻列の i 番目までと識別音韻列の j 番目までの最良マッチングスコアとすると，辞書の i 番目の音韻が母音の場合については，$i\geq 2$ のとき，

$$L(i,j)=\max\{L_1(i,j),L_2(i,j),L_3(i,j),L_4(i,j),L_5(i,j),L_6(i,j)\}$$

で求められる．ここで

$$L_1(i,j)=L^*(i-1,j)+S_0(i,j)$$
$$L_2(i,j)=L(i-1,j-1)+S_0(i,j)$$
$$L_3(i,j)=L^*(i-1,j-1)+\{S_0(i,j-1)+S_0(i,j)\}/2$$
$$L_4(i,j)=L(i-1,j-2)+\{S_0(i,j-1)+S_0(i,j)\}/2$$
$$L_5(i,j)=L^*(i-1,j-2)+\{S_0(i,j-2)+S_0(i,j-1)+S_0(i,j)\}/3$$
$$L_6(i,j)=L(i-1,j-3)+\{S_0(i,j-2)+S_0(i,j-1)+S_0(i,j)\}/3$$

である．ただし $L^*(i,j)=\max\{L_2(i,j),L_3(i,j),L_4(i,j),L_5(i,j),L_6(i,j)\}$ で前述の制限 3 の条件に対応する．また $L(1,1)=S_0(1,1)$，$L(1,2)=\{S_0(1,1)+S_0(1,2)\}/2$，$L(1,3)=\{S_0(1,1)+S_0(1,2)+S_0(1,3)\}/3$ である．子音の場合は，$L(i,j)=\max\{L_1(i,j),L_2(i,j),L_3(i,j),L_4(i,j)\}$ で求められる．$L_1,L_2,\cdots,L_6$ は，対応づけの種類に対応するもので **Fig. 4**（次頁参照）における対応づけの道 $r_1,r_2,\cdots,r_6$ に対応する．

孤立単語音声の認識においては，マッチングされる単語辞書の音韻列長（音韻数）を i_0，識別音韻列長を

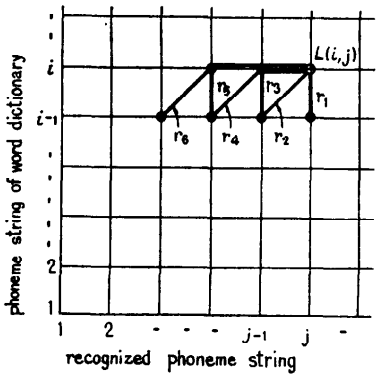

Fig. 4 Kinds of matching route

j_0 とすると，この単語に対する尤度は $L(i_0, j_0)/i_0$ で与えられる．識別対象であるすべての単語に対して尤度を求め，最高の尤度を持つ単語を識別結果とする．

Fig. 5 は単語「カケル」のマッチング例を示している．

連続単語音声の認識および連続音声中の任意の単語の検出は，この手法を拡張することによってなされる[19),20)]．

5. 数字音声の識別実験

連続音声の認識用に開発した以上の手法が，孤立単語音声の認識にどれだけ有効かを検討するため，数字音声（イチ，ニ，サン，ヨン，ゴ，ロク，ナナ，ハチ，キュウ，レイまたはレーと発声）の認識の実験をした．音韻の標準パターンと音韻間類似度は，10 名の男性が発声したあらゆる有声子音を含む有意味単語，計 2,450 単語から求めた．この 10 名が 10 数字を各 5 回発声（試料 1）と 6 ケ月後に各 5 回発声（試料 2）それに他の男性 10 名が各 5 回発声（試料 3）したもの，合計 1,500 サンプルで実験をおこなった．音声試料は一旦録音テープに録音した後，音声研究用小型計算機 (MELCOM-70) で管理しながら KUIPNET を通じて汎用計算機 (NEAC-2200/250) のファイルに編集した．認識実験はすべて NEAC-2200/250 でおこなった．各単語の辞書は，Table 4 と同じである．結果を **Table 5** に示す．表で話者適応あり (speaker-dependent) の欄は，各話者ごとの音韻標準パターンを用いた場合であり学習はおこなっていない．認識率は，発声の時期の差および話者にさほど関係がなく，不特定話者に対して 97%以上の認識率が得られた．音韻パターンの学習機能を追加すれば，98% 以上の認識率を得ることは容易である．

Table 5 Recognition rate of spoken numerals

method	test samples *No.* 1	test samples *No.* 2	test samples *No.* 3
speaker-independent	96.4%	97.1%	97.6%
speaker-dependent	97.4%	98.2%	

音韻識別結果の第二候補や第一候補の信頼度の導入は，音韻識別部の信頼性を高め単語の認識率の向上に大変有効であった[8)]．なお音韻識別のアルゴリズムを若干簡略化し，ほぼ実時間で限定語彙の単語音声を小型計算機で認識するシステムを別に開発している[21)]．

6. むすび

音声理解システムの一環として開発した不特定話者の連続音声向き単語音声の識別法について述べた．この手法は，処理時間の大部分が音韻識別にかかっており，単語辞書とのマッチングに要する時間は短い．また識別する語彙数が増加しても識別時間はあまり変わらず，単語辞書の記憶容量もそれ程必要としない特徴がある．

比較的多数の話者の多量の数字音声に対して，識別率が 97% を越えたことは，ここで述べた方法が孤立

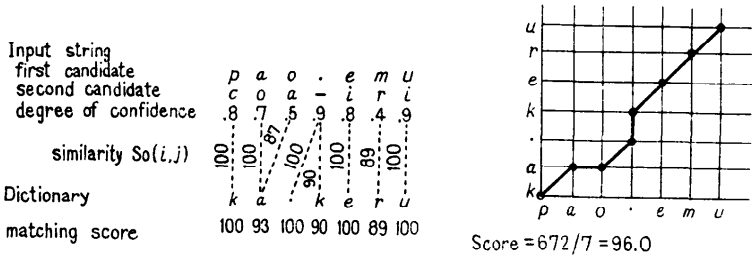

Fig. 5 Graphic representation of word matching and matching score

単語音声に対しても大変有効であることを示している．また話者別の標準パターンを採用することによって 98% 以上の識別率が得られることを示した．これらの実験結果は，今後の単語音声の認識研究に対して重要な示唆を与えるものと期待される．さらにこの方法は，連続単語音声の認識や連続音声中に存在するキーワードの検出に利用でき，音声理解システムには一層有用である[8),11)]．

本稿で述べた音韻列からの単語識別法は，他の時系列パターンの認識にも適用可能と思われる．より高精度な識別手法を確立するために，残された問題は次の点である．

1. 音韻識別能力の強化（p, t, k の識別等）．
2. 単母音の発声等による話者に負担をかけない迅速・確実な話者パターンの予備学習法[21)]．
3. 話者の個人差を考慮した単語辞書や音韻間類似度の構成法もしくは学習法[21)]．
4. 時系列の前後関係を考慮した文脈依存なマッチング法[21)]．
5. 韻律情報の利用法の確立．

これらについては，稿を改めて報告する予定である．

参 考 文 献

1) 坂井，堂下：会話音声識別装置，信学誌，Vol. 46. No. 11, pp. 1696～1702 (1963).
2) J. R. Pierce: Whither Speech Recognition?, JASA, Vol. 46, No. 4, pp. 1049～1051 (1969).
3) V. M. Velichko and N. G. Zagoruyko: Automatic Recognition of 200 Words, Int. J. Man-Machine Studies, Vol. 2, pp. 223～234 (1970).
4) 好田，橋本，斉藤：数学音声の機械認識系，信学論，Vol. 55-D, No. 3, pp. 186～193 (1972).
5) 迫江，千葉：動的計画法を利用した音声の時間正規化に基づく連続単語認識，音響誌，Vol. 27, No. 9, pp. 483～490 (1971).
6) 加藤，千葉，永田：数字音声認識装置，信学誌，Vol. 47, No. 9, pp. 1319～1325 (1964).
7) M. R. Sambur and L. R. Rabine: A Speaker-Independent Digit-Recognition System, BSTJ, Vol. 54, No. 1, pp. 81～102 (1975).
8) 坂井，中川：音声理解システム LITHAN のシステム評価，音響学会音声研資，S75-30 (1975).
9) 樽松，武田，井上：書き替え規則を用いて音声認識をおこなう場合の一構成法，信学論，Vol. 55-D, Vol. 2, pp. 91～98 (1972).
10) 白井，藤沢：音韻識別部をもつ数字音声認識，信学論，Vol. 57-D, No. 3, pp. 135～142 (1974).
11) 坂井，中川：タスク内の自然言語音声を理解するシステム-LITHAN，信学会パターン技報，PRL 75-54 (1975).
12) 中津，好田：VCV 音節を単位とした単語音声の認識，音響学会春季大会 (1973-5).
13) 中野，市川，中田：音声認識のための各種パラメータの評価，音響学会春季大会 (1972-5).
14) 坂井，大谷，中川，崎村：会話音声のセグメンテーションと音韻識別について，電気四学会連合大会 (1973-10).
15) G. Nagy: Classification Algorithms in Pattern Recognition, IEEE Trans. Vol. AU-16, pp. 203～212 (1968).
16) N. J. Nilsson: Learning machines, McGraw-Hill (1965).
17) 坂井，田畑：VCV音節の多変量解析，信学論，Vol. 56-D, No. 1, pp. 63～70 (1973).
18) 市野，平松：統計的認識系の特徴評価関数，信学論，Vol. 53-C, No. 10, pp. 748～755 (1970).
19) 中川，坂井：単語系列情報の利用，電気四学会連合大会 (1974-10).
20) 坂井，中川：連続音声中の単語同定法，音響学会春季大会 (1975-5).
21) 坂井，中川，林：音韻スペクトルの個人差の予備学習による限定語彙単語音声の認識，信学会，電気音響技報 EA 75-61 (1976-1).
22) S. Nakagawa: A Speeh Understanding System of Simple Japanese Sentences by Computer, 京都大学博士論文(準備中).

(昭和50年12月19日受付)
(昭和51年3月5日再受付)