標準パターンDP合成法による多数話者の単語音声認識 (2)

標準パターン DP合成法による 多数話者の単語音声認識（II）

高 橋 文 彦* 矢 田 茂 郎**

Speech Recognition for Multi-speakers by Standard Patterns Synthesized by DP Matching Method (II)

Fumihiko TAKAHASHI* Shigerou YADA**

ABSTRACT

This paper is a report for a research of speech recognition in case of unspecified multi-speakers. In the former report, we developped a new Synthesis Method for Standard Pattern, averaging many patterns by utilizing DP Matching Method. In this research, peculiar parameter of voice consists of two components, main component is time series of PARCOR Coefficient and subcomponent is time series of voice power. Experimental results of speech recognition are fairly improved by adding voice power component, the sore of recognition is about 98% in case of 50 words (Japanese cities) by unspecified male voice.

Key Words: Recognition, Speech, Standard Pattern, Multispeakers

1. まえがき

音声認識や音声合成などの音声関連技術は，ここ数年で急速に進歩してきた．特に合成音声は，自動販売機から家電製品，電話の番号案内に至るまで広範囲に利用されるようになった．これらは音声の情報圧縮技術の進歩と，そのLSI化の成功によるものである．

音声認識技術に関しても実用化は進んでいる．簡単な音声認識用ボードが市販され，ホビー的な用途にまで使われるようになってきている．しかし，これらはすべて使用者が予め自分の音声を登録しておくタイプのもので特定話者音声認識と呼ばれる分野のものである．音声認識が広く利用されるためには，誰の声でもすぐ聞き分けるような認識装置が要求される．不特定多数話者の音声認識の研究が待たれるわけである．

筆者らは，この方向に沿って研究を進め，さきに標準パターンDP合成法[1]を提唱した．これはDPマッチングの手法を利用して，多数のパターンを平均化して標準パターンを合成し，個人特徴のバラツキを抑えた汎用の標準パターンを得るものである．前報では特徴パラメータにPARCOR係数を用い，50単語で92%程度の認識率（成年男子）を得た．

しかしながら，PARCOR係数には本質的に音の強弱や高低に関する情報は含まれない．したがって今回は，強弱に関する情報，すなわち音声パワーの時間変化を特徴パラメータとして追加し認識率の改善をはかった．その結果，認識率を98%程度まで向上させることができたので報告する．

2. 音声認識システムの概要

このシステムの概要を図1に示す．入力音声を予め登録された標準パターンと比較することは，特定話者の場合と同じであるが標準パターンの作成に工夫がほどこされている．標準パターン作成用に入力された音声信号は分析過程で特徴抽出され，各特徴パラメータの時系列パターンとなる．これを同一語彙毎に平均化して標準パタ

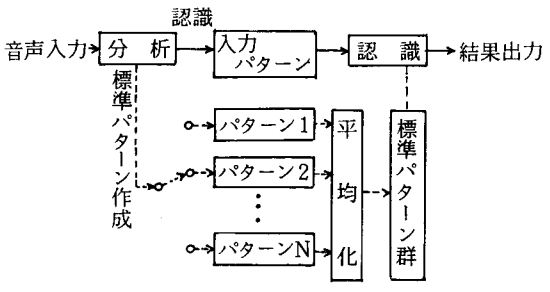


図1 音声認識システムのブロック図

平成2年4月9日受理

*電気工学科(Department of Electrical Engineering)
** 東京大学工学部(Dep. of Engineering, Tokyo Univ.)

ーンを作成し登録する．この平均化の際に，各時系列パターンを単語グループ毎にDPマッチング法で最適伸縮させて重ね合わせ平均化して，標準パターンを作成するわけである．

3. 音声の収録と特徴パラメータ

音声収録装置の外観を**図2**に示す．日本電気製パーソナルコンピュータ・PC 9801を主体とするもので，これに音声入力用A/Dコンバータと音声出力用のD/Aコンバータを附加したシステムである．音声を10kHzのサンプリング間隔でA/D変換（12 bit）し，ハードディスク（20MB）に格納するようになっている．

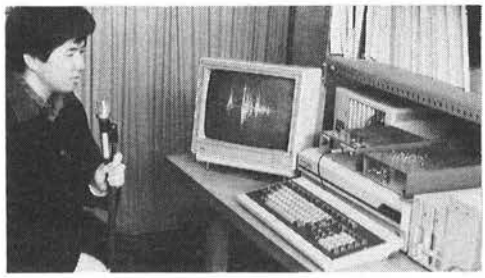


図2　音声入出力装置

入力音声は50個の単語（日本の都市名）で11名の男性が発声したものである．50単語は連続して発声するが，単語間の区切りと単語内の無音部を機械が混同しないように，区切りの休止期間は十分長くとるよう注意して発声した．

認識に用いる特徴パラメータはPARCOR係数（1～15次）と音声パワーで，共に25.6 msec毎（データ個数256）にフレーム化して計算している．したがって単語音声は，25.6 msec毎の時系列パターンに変換されることになる．

4. パターンマッチングと時間軸整合

マッチング法はパターン認識の汎用的手法である．あらかじめ登録された標準パターンと別の未知パターンを重ね合わせ，パターンの間のズレ（距離）を計算し，最もズレの小さいものを選び出して，そのパターンに属すると判定する手法がパターンマッチング法である．

音声信号の入力パターンは，PARCOR係数群と音声パワーの時系列として次式で表わされる．

$$\mathbf{A}=[\ \mathbf{a}_1,\ \mathbf{a}_2,\ \mathbf{a}_3\ \ldots\ \mathbf{a}_i\ \ldots\ \mathbf{a}_I]\qquad(4.1)$$

Iは入力単語音声の長さ（フレーム数）で，$\mathbf{a}_i$ は第iフレームのPARCOR係数群（1次～M次）と音声パワーパラメータからなる，（M＋1）次元ベクトルである．

また，予め登録されている N 個の標準パターンの中で，第n番目のパターンを次の式で表わすことにする．

$$\mathbf{B}^n=[\ \mathbf{b}_1^n,\ \mathbf{b}_2^n,\ \mathbf{b}_3^n,\ldots\mathbf{b}_j^n\ \ldots\mathbf{b}_J^n\ ]\qquad(4.2)$$

ここで入力パターン $\mathbf{A}$ と標準パターン $\mathbf{B}^n$ との間の距離関数としてスカラ量 $D(\mathbf{A},\mathbf{B}^n)$ を定義すると，N個のパターンの中で，この値を最小にするパターンに属すると判定するのがパターンマッチング法である．当然の事ながら，判定結果はこの距離関数 $D(\mathbf{A},\mathbf{B}^n)$ の定義の仕方によって大きく影響を受ける．

さて音声パターン $\mathbf{A},\mathbf{B}$ を較するとき，まず発声時間の長短が問題になる．たとえば，オハヨゴザイマスとオハヨーゴザイマースは後者の方が長いが同じ言葉である．したがってパターンマッチングの際の時間軸整合は単純で一様な線形伸縮ではなく，上記のことを考えた非線形伸縮で重ね合わせる必要がある．この整合法は時間軸の変動を吸収するのに好都合であるが，あまりに伸縮の幅を広くすると異なる単語でも同一単語と誤認識する恐れがあり，一定の制限を設ける必要がある．

5. 距離関数とDPマッチング

ここでは距離関数として，重みつきユークリッド距離を採用した．これに時間軸伸縮の概念をとり入れて関数を定義するわけであるが，**図4**に時間軸整合の模様を示す．$\mathbf{A},\mathbf{B}$ 両パターンの対応部分は図中の曲線で示される．時間軸の対応は格子点（i, j）の系列Fで表わすことができる．

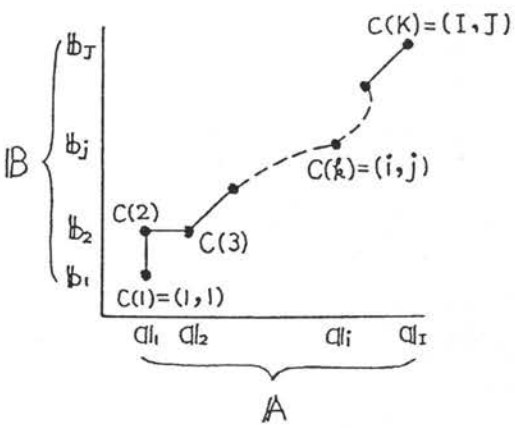


図3　時間軸整合とDPパス

$$F=[C(1),\ C(2),\ldots C(k)\ \ldots\ C(K)] \qquad (5.1)$$

ただし，$C(k)=(i(k),j(k))\ldots k$ 番目の格子点

ここで対応する部分の特徴ベクトル $\mathbf{a}_i$ と $\mathbf{b}_j$ の距離をベクトルのユークリッド距離として定義すると，その部分の距離 d(c) は

$$d(c)=d(i,j)=|\mathbf{a}_i-\mathbf{b}_j| \qquad (5.2)$$

対応経路Fに沿って計算したパターン間距離Eは d(c) の荷重平均となり

$$E(F)=\frac{\sum_{k=1}^{K} d(C(k))\,W(k)}{\sum_{k=1}^{K} W(k)} \qquad (5.3)$$

$W(k)$ は重みを表わし，その値は対応の仕方によって変わるものである．すなわち経路Fが斜め（伸縮のない部分）では $W(k)=2$，垂直又は水平（伸縮のある部分）では $W(k)=1$ とすると格子点の数に関係なくパターン間距離は正規化されることになる．

この距離E(F)を最小にするよう部分伸縮するのが前節の非線形マッチングであり，マッチングの経路FをDPパス（path）という．計算の効率化のためダイナミックプログラミング（DP）を使うためである．

このDPマッチング[1]は伸縮の幅を大きくしすぎると異種パターンにまでマッチングし誤認識の恐れが生じる．したがって一定の制限条件を設ける必要がある．許容されるマッチングの範囲を整合窓といい，窓寸法rを指定するとマッチングの範囲は制限され $|i-j|\leq r$ となる．また部分伸縮についてもその度合いを制限する方が良い．たとえば，1/2～2などのようにである．伸縮制限指数Lを設定すると許容伸縮限度は（$1/L\sim L$）となる．

6. 個人別音声パターンの平均化

不特定多数話者の単語音声認識において，もっとも重要な鍵を握るのが汎用標準パターンの作成法である．筆者らは多数話者の特徴パターンを平均化する方法で話者の個性を抑え，汎用的な標準パターンを合成した．

平均化の際にまず問題になるのが時間軸の取扱いである．**A**, **B** 2つのパターンの整合には，DPマッチングで時間軸を部分的に伸縮して対応部分を探したが，平均パターンの作成に際しては，この時間軸の伸縮も2つのパターンの中間的なものにしなければならない．各部の対応関係を明確にしておくため，DPマッチングと同時にDPパス（経路F）を記憶しておく必要がある．

このDPパスを使って，時系列パターン **A**, **B** を時間軸伸縮も含めて平均化したパターン **C** を求めるアルゴリズムについて説明する．

まず **A**, **B** パターンとその平均パターン **C** を次のように表わすことにする．

$$\mathbf{A}=[\mathbf{a}_1,\ \mathbf{a}_2,\ \mathbf{a}_3,\ldots\ \mathbf{a}_i,\ldots\ \mathbf{a}_I] \qquad (6.1)$$

$$\mathbf{B}=[\mathbf{b}_1,\mathbf{b}_2,\mathbf{b}_3,\ldots\mathbf{b}_j,\ldots\mathbf{b}_J] \qquad (6.2)$$

$$\mathbf{C}=[\mathbf{c}_1,\mathbf{c}_2,\mathbf{c}_3,\ldots\mathbf{c}_h\ \ldots\ \mathbf{c}_H] \qquad (6.3)$$

ただし

$$H=\mathrm{int}[(I+J)/2] \qquad (6.4)$$

ここでintは小数部を除去する演算子で，これは離散系に合わせる処置である．$\mathbf{a}_i$, $\mathbf{b}_j$, $\mathbf{c}_h$ はそれぞれ第i, j, h番目の特徴ベクトルである．式（6.4）は **A**, **B** 両パターンを平均したパターン **C** のフレーム数（時間長）が両者のほぼ中間になることを示す．

次に平均パターン **C** の各要素 **c** を求める方法を考える．パターン **A**, **B** の各要素 **c** と **b** の最適対応（DPパス）が図4のようであったとする．両者の平均パターンの要素 **c** は，図上の対応する要素 **a** と **b** を結んだ線分の中央に相当する時刻に，$\mathbf{c}=(\mathbf{a}+\mathbf{b})/2$ となることが原則である．

ただし，この時系列はフレーム単位の離散系であるため図の□点のような中途半端な位置はとり得ないので，このような場合は半フレームだけ左に寄せて，すぐ左の・点と考えて近似処理（図4の・←□の部分）をする．

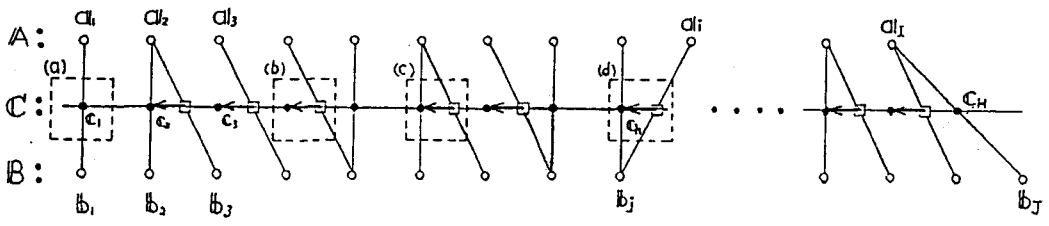


図4　パターン要素間の対応と時間軸の平均化

以上のことから，要素 $\mathbf{c}_h$ の計算には $\mathbf{a}$，$\mathbf{b}$ の対応関係も考慮して，次のように定義すればよい．

(1) 1：1 対応

$\mathbf{a}_i \leftrightarrow \mathbf{b}_j$ の対応（図4のⓐ，ⓑ部では

$$\mathbf{c}_h = (\mathbf{a}_i + \mathbf{b}_j) / 2$$

(2) 1：2 対応

$\mathbf{a}_i \leftrightarrow \mathbf{b}_j, \mathbf{b}_{j+1}$ の対応（図4のⓒ部）では

$$\mathbf{c}_h = (2\mathbf{a}_i + \mathbf{b}_j + \mathbf{b}_{j+1}) / 4$$

$\mathbf{a}_i, \mathbf{a}_{i+1} \leftrightarrow \mathbf{b}_j$ の対応（図4のⓓ部）では

$$\mathbf{c}_h = (\mathbf{a}_i + \mathbf{a}_{i+1} + 2\mathbf{b}_j) / 4$$

以上の手順で，類似した2つの時系列パターンについて時間軸伸縮も含めて平均化したパターンを合成することができる．さらにこの方法を拡張すれば多数パターンから1個の標準パターンが合成できる．まずパターン $\mathbf{A}$ と $\mathbf{B}$ を平均し，それと他のパターン $\mathbf{C}$ と $\mathbf{D}$ を平均したものをもう一度平均する，という様に平均化を進めれば，2^n 個（n整数）の原パターンから1つの標準パターンを合成することができる．

この方法で得られる標準パターンは，話者の特徴を平均化して打ち消し合うため，話者独立性の高い汎用性のあるパターンと考えられる．

7. 音声パワーのパラメータ化

さきに述べたようにPARCOR係数には，音声の強弱や高低の情報は含まれていない．これは発声のときの口腔の形状に相当するものである．したがってPARCOR係数だけで音声認識を行うことは，口の動きだけを見て判断するようなものである．自然言語では常に強弱や高低の変化があり，これらを認識に利用するのが妥当であろう．

図5はこの例を示したものである．話者が異なっても

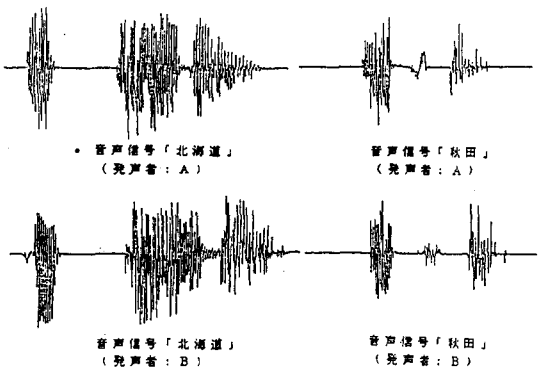


図5　音声パワーの変動

「北海道」「秋田」それぞれの振幅パターンは類似した形状を示す．したがって筆者らは，音声パワーをパラメータ化し，従来のPARCOR係数に追加することにした．

フレーム単位の音声パワー p は（振幅）^{2} のフレーム長毎の積分値であるから，パワーパラメータの時系列要素 P_i は正規化して次のように定式化される．

$$P_i = \frac{n p_i}{2\sum_{j=1}^{n} p_j} \qquad (7.1)$$

ここで n はフレーム数（単語長），p_i は第 i 番目のフレームの音声パワーである．

実験の結果，パワーパラメータ P_i は，通常の単語で0～6くらいの範囲で変化することがわかった．

8. プログラムの構成

プログラムは前処理段階から認識まで，つぎの4ブロックから構成される．使用言語はFORTRANである．

(1) 単語切り出しプログラム

一連の音声信号から，それぞれの単語部分を切り出すプログラムである．音声レベルが一定のしきい値より低い場合を無音部とし，基本的にこの部分で単語音声を切り出すようにしている．

ただし，単語内の無音部と混同をさけるため，発声の際に単語間の区切りの休止期間を十分長くとるようにしている．その上このプログラムでは，まずすべての無音部を検出し，それを長さの順に並べ，上位から（単語数＋1）番目までを単語間無音部とする方式をとった．切り出した後も誤切断がないか耳で聞いて確認をしている．

(2) 分析プログラム

切り出した音声データを一定長（256）でフレーム化し，PARCOR係数と音声パワーパラメータを計算して発声者毎のパターンファイルに登録する．

(3) 平均化プログラム

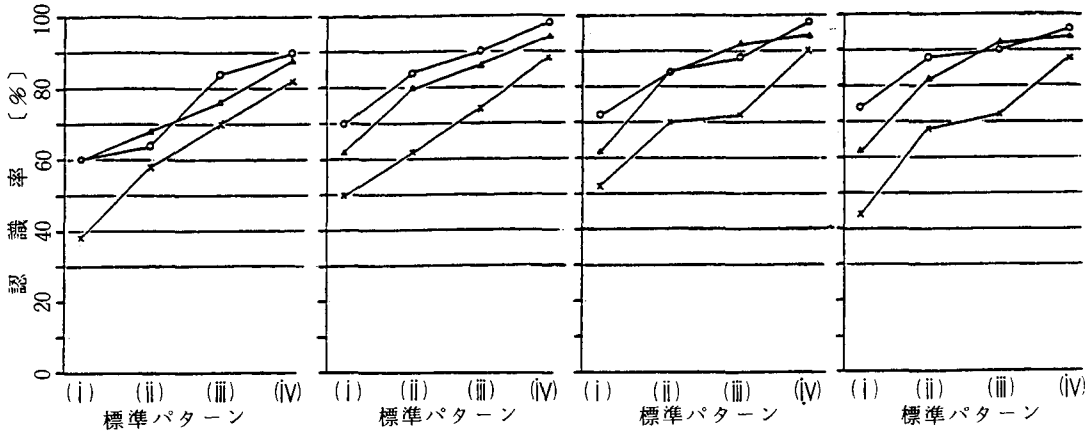
パターン群をDPマッチングによって平均化し，それぞれの標準パターンとして登録する．

(4) 認識プログラム

入力パターンに対して，登録された標準パターンを順次DPマッチングで比較し，距離最小のパターンを認識結果として選出する．

9. 認識実験

音声の収録段階はオフラインの音声収録装置（パーソナルコンピュータ）で行い，収録されたデータは，1MB



(a) 音声パワーパラメータの重み係数0（添加なし）の場合

(b) 音声パワーパラメータの重み係数1の場合

(c) 音声パワーパラメータの重み係数1.5の場合

(d) 音声パワーパラメータの重み係数2の場合

図6　認識実験

のフロッピーディスクで情報処理センターに持ち込み，中型計算機FACOM M360Rで処理する方式をとった．

音声データは「札幌」「青森」「秋田」など50都市名を11人の男性（20才前後）が発声したものである．このうち8名分は標準パターン作成用とし，他の3名分を独立パターンⒶⒷⒸとし，それぞれについて認識実験を行った．**図6**はその結果を示す．図中のマーク○，△，×はそれぞれ被験者Ⓐ，Ⓑ，Ⓒの音声を示す．

ここで横軸の(i)，(ii)，(iii)，(iv)は標準パターンのレベルを示すもので，(i)は単独パターン，(ii)は2個のパターンを平均化したもの，(iii)，(iv)はそれぞれ4個，8個のパターンを平均化した標準パターンを示す．

特徴パラメータはPARCOR係数（1～15次）に(7.1)の音声パワーパラメータを加え，この重み係数を種々変更して実験したものである．

標準パターンのレベルが高くなるほど（参加人数が多くなるほど）認識率は向上している．また重み係数は1.5附近がもっとも認識率が高くなるようである．

10. 考　察

標準パターンの合成に参加する人数が多くなるほど良好な成績が得られる．人数を16か32くらいまで増やすとかなり良質の汎用パターンが得られるであろう．

音声パワーパラメータを加えた効果は明らかで，左端(a)（パワーパラメータなし）と比べると右の3つのグラフは何れも認識率が高くなっており，(c)の重み1.5の場合と比較すると6～8%向上している．

この結果からみると，7・で述べたようにPARCOR係数だけで認識を行う手法は，いわば口の動きだけで言葉を推定するようなもので，本質的に欠陥があると言えよう．

○，△，×の個人差によるバラつきもかなり認められるが，標準パターンのレベルが上るほどバラつきも少なくなっているようである．

パワーパラメータの重みは1.5程度が適当となっているが，このパラメータ自身の変動振幅が大きいので，PARCOR係数に置きかえると，3～4個分くらいと思われる．パワーパラメータのとり得る値は0～6くらいで平均値は0.8程度，一方PARCOR係数のとり得る値は－1～＋1で絶対値の平均は0.3～0.4となっているからである．

11. むすび

前報[1]につづいて，筆者らの提唱した標準パターンDP合成法が，不得定多数話者の単語音声認識に大変有効であることが再確認できた．

また，従来のPARCOR係数に音声パワーパラメータを加えることによって，認識率が飛躍的に向上することがわかった．標準パターンの合成に参加する人数をもう1～2段階ふやして16名か32名にすれば，50単語で100%近い認識率が得られると思う．

音声データ収録に協力して頂いた電気工学科の学生諸君や便宜をはかって頂いた情報処理センターの職員各位に感謝の意を表する．

参 考 文 献

1）高橋ほか　標準パターンDP合成法による多数話者の単語音声認識　府立高専研究紀要（1987）
2）板倉　統計的手法による音声の特徴抽出　音声処理シンポジウム（1971）
3）梅崎，中川　単音節を標準パターンとした連続音声認識における発声速度の問題　音声研究資料（1985－6）
4）藤崎博也　連続音声認識のための音節の変動に関する検討　音声研資料（1984－12）