中間累積距離と音節間類似度を用いた単音節音声認識†

西村雅史†† 松田安弘††

日本語音声ワードプロセッサを目的として，特定の話者が単音節単位に発声した音声の認識方式について論ずる．まず，単音節の認識方式としては，中間累積距離マッチング法を提案し，従来必要とされていた子音・母音境界の正確な抽出なしに，高い認識精度が得られることを示す．次に，発声の変動による認識率の低下に対処するため，候補音節間の類似度に基づくテンプレートの教師付学習方法を提案する．また，音声認識部の最終的な出力である候補音節列の最適化を図り，候補単語数を効果的に削減するためにも，この音節間類似度が有効であることを示す．実験の結果，男性話者3名が日本語の68音節を10回発声したデータについて平均認識率95.3%，第2候補まで含めるならば平均98.0%の認識率を得た．また，出力候補音節列の最適化を図った場合，平均98.3%の精度を保ちながら，候補音節数を平均1.24に削減できた．これは4音節からなる単語を認識対象とする場合，通常の方法に比べ候補単語数を1/7以下に削減できることを示している．

1. はじめに

情報処理システム，特に日本語ワードプロセッサの普及に伴い，簡便な日本語入力手段の一つとして，音声入力が注目を浴びている．しかしながら現在の音声認識には十分でない部分が多く，対象語彙数，話者，発声方法等に少なからず制限を与えたものにならざるをえない．特に，大語彙を認識対象とする日本語音声ワードプロセッサの実現のためには，現在のところ音声を単音節単位に区切って発声する単音節入力方式が有効な手段の一つである．

この方式では，日本語発声の最小単位であり，かつ仮名文字とほぼ一対一の対応をなす音節を用いるため，少ない登録音声によって大語彙の認識が可能となる．しかしながら，単音節認識では，数十ミリセカンド程度の持続長しかない語頭子音の正確な識別を必要とするため，発声全体を対象とする従来の単語音声認識技術をそのまま適用したのでは高い認識率は得られない．そのため，認識が容易な後続母音の判定で音節候補を絞った後，先行子音部に対してマッチングを行うのが一般的である．特に子音の認識に関しては，前もって正確に抽出した子音部に対して，線形マッチングを行う方法[5]や端点フリーのDPマッチングを行う方法[2),4)]などが提案されている．しかし，子音部から母音部への移行は漸増的であり，子音部の正確な抽出は一般に困難である．これに対し，古井[3]は，語頭部に対する線形シフトマッチングと，ケプストラムを用いた大局的なマッチングを組み合わせた，子音部の正確な抽出を必要としない方法を提案している．本論文では，従来単語認識に用いられてきたDPマッチングの中間累積距離を用いた方法を提案し，本方法が，子音部の正確な抽出を必要とせず，しかも比較的簡単なアルゴリズムであることを示す[8]．

一方，単音節認識においては識別のための特徴が微細なものであるため発声上の変動を生じやすく，初期に登録された少数のテンプレートのみでは認識精度の劣化を起こしやすい．これに対処するためには，標準テンプレートに対して入力音声の学習を行うことが望ましい[7]．しかし，すべての入力音声を単純に標準テンプレートとして追加あるいは平均化したのでは，使用者の誤操作や，不良発声が生じた場合に，認識精度の大幅な劣化を起こす危険性がある．これに対しては，入力音声と，正答となる候補音節との照合距離によって学習の操作（平均化，追加，却下）を選択する方法が提案されている[6]．本論文ではまず，初期登録音声中に含まれる不良発声の除去を行った上で初期テンプレートを作成し，従来あまり検討されていなかった，学習のための基準を明確にする．さらに，単一の候補との距離のみに着目するのではなく，候補となった音節間の類似度を定義することにより，入力音声と複数の候補音節との総合的な距離関係から，入力音声の品質や使用者の操作内容についての判断を行い，それによって学習の操作を選択する教師付の学習方法を提案する[10]．

また，音声認識の結果を言語処理部で処理する場合，音声認識部から各音節に対し一定数の複数候補を出力することによって正答の脱落を避け，言語処理部で辞書中の最適な単語に絞るという方法をとることが

† Monosyllable Recognition by Using Intermediate Cumulative Distance and Normalized Distance Similarity by MASAFUMI NISHIMURA and YASUHIRO MATSUDA (Science Institute IBM Japan, Ltd.).

†† 日本アイ・ビー・エム(株)サイエンス・インスティチュート

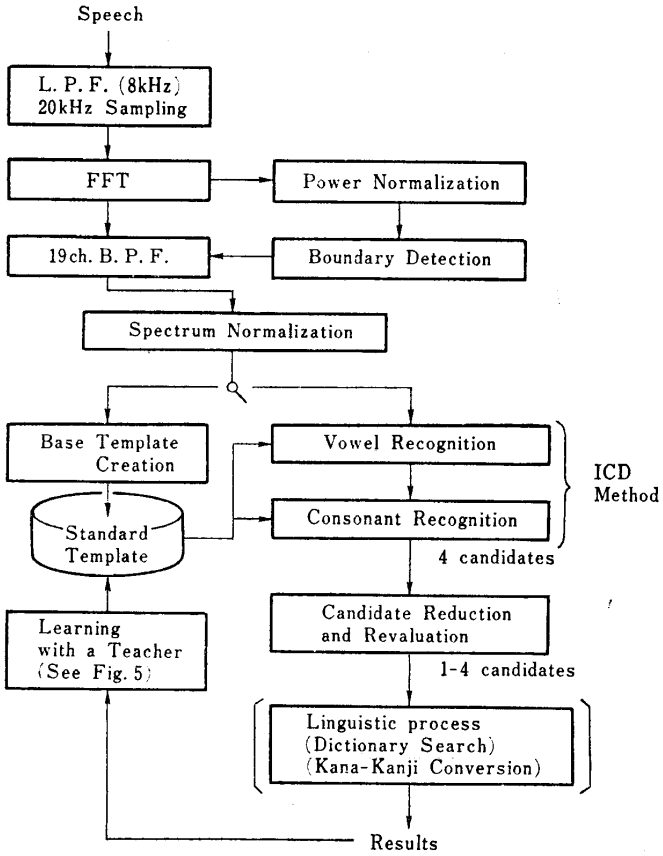

図 1 単音節認識システムの概要
Fig. 1 Outline of a Japanese monosyllable recognition system.

多い[11)~13)]．しかし，これらの方法では一部の正答の脱落を避けるために，不必要な音節候補を言語処理部に与えることになり，辞書の検索に多くの処理を必要とし，必ずしも高い精度は得られない．本論文では，テンプレートの教師付学習のために導入したところの，音節間類似度に基づく候補音節列の出力方法について検討し，その有効性について触れる[10)]．

2. 単音節入力音声認識システムの概要

2.1 音声分析部

今回検討を行った単音節入力音声認識システムの概要を**図 1** に示す．

音声は 8 kHz の低域通過フィルタを通過後，20 kHz，12 bit で標本化，量子化される．このデータについて，一階差分による高域強調を行ったのち，窓幅 25.6 msec，フレーム周期 10 msec のハミング窓を用いた FFT を行う．さらに，この出力を 19 チャンネルにまとめ，対数変換したものを 8 bit に正規化して特徴ベクトルとする．

2.2 音声認識部

入力音声は，19 次元の特徴量に変換されたのち，第 3 章で述べる中間累積距離マッチング法により，後続母音の大分類および子音認識が行われ，第 4 位までの候補が出力される．次に，この過程で得られた累積距離に基づいて，候補音節間の類似度が計算され，不必要な候補が削除される．この候補の絞りこみの結果，候補が一意に定まらない場合については，入力音声の局所的な特徴による判定を行い，最終的な候補音節列を出力する．

このあと，候補音節列は言語処理部において辞書との照合，子音の一字訂正などが行われたのち，仮名・漢字変換処理され，最終的な仮名・漢字混り文を得る[14)]．

一方，標準テンプレートについては，初期登録用音声から学習の基準となるテンプレートを作成した後，認識結果を基にした教師付の学習を行う．

3. 中間累積距離マッチング法による認識

3.1 認識アルゴリズム

ここでは，**図 2** に示すような入力パターン側に依存した形の非対称型パスをもつＤＰマッチング[1)]を行い，あらかじめ設定した一時点 u_1 までの中間累積距離 (Intermediate Cumulative Distance : ICD) により子音を識別する．u_1 は子音の特徴を多く有する部分に設定されていることが望ましいが，パスおよび終端点がともに入力パターンに大きく依存する形で設定されているため，u_1 の位置によ

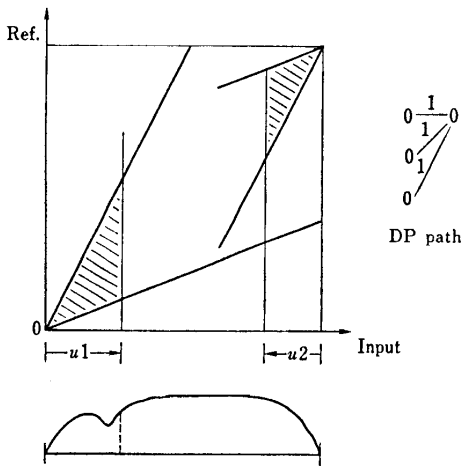

図 2 中間累積距離マッチング法
Fig. 2 Intermediate cumulative distance (ICD) method.

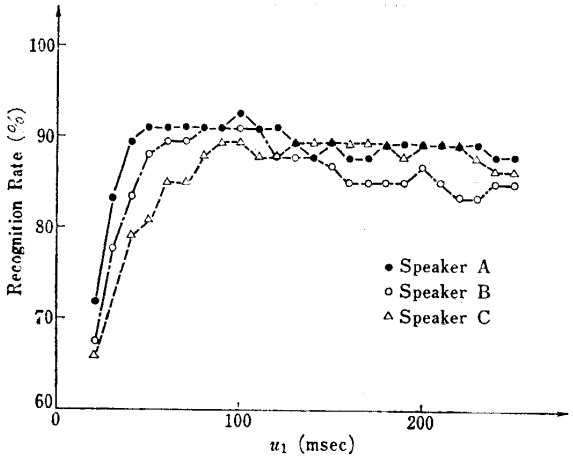

図 3　中間累積距離マッチング法において u_1 が認識率に与える影響
Fig. 3　Effect of u_1 on recognition rate with ICD method.

る認識率の変動は小さいと考えられる．後続母音の認識についても，各発音の終端点から，時間と逆方向に同様の手法を適用した．母音テンプレートには孤立発声された5母音，撥音に加え，無声化，鼻音化しやすい/si/, /su/, /ci/, /cu/, /mi/, /mu/の後続母音を用いている．

図3に，u_1 と認識率の関係を示す．ただし，後続母音の認識率は 100%であった．この図からわかるように，高い認識率を示す u_1 の範囲は広く安定しており，話者による違いも小さい．一方，発声速度の変動に伴い発声長は大きく変化するが，子音長は比較的安定しているため 150 音節/分以下の発声速度では，発声速度の変動が u_1 の最適値に与える影響は小さいものであった*．実験結果を**表1**に示す．なお，単音節の発声速度は，実用上約 100 音節/分程度が望ましく，150 音節/分以上の発声は，話者にとって多大の負担となるようである[9]．これらの予備実験の結果から，以降の実験では $u_1=110$, $u_2=40$ msec に設定した．

また，視察によって正確に切り出した単音節の子音部分（ホルマント遷移部を含む）について DP マッチングを行った結果（**表2**）との比較からも，本方式による認識精度の高いことがわかる．なお，子音部の切り出しはテンプレートおよび入力音声の双方について行っている．

* 文献3)では線形シフトマッチングの範囲を，$u_1=\min(aL, L_{max})$ により決定しているが，DPマッチングを用いるわれわれの方法では，発声長の変化に対しても，u_1 を一定値として十分であった．全長に比例する1時点をとった場合については文献8)で検討している．ただし L は音節長，a と L_{max} は実験定数を示す．

表 1　発声速度と u_1 の最適値の関係
Table 1　Relationship between utterance speed and u_1.

発声速度 (音節/分)	平均音節長 (msec)	最適値	
		u_1(msec)	認識率(%)
50	260	80～110	89.7
100	240	70～110	91.1
150	160	100～120	88.2
200	140	120～140	75.0

男性話者1名，68単音節

表 2　視察データを用いたDPマッチングと中間累積距離マッチング法の比較
Table 2　Comparison between manual segmentation and ICD method.

認識方法	マッチング部位	子音の累積認識率(%)		
		第1候補	第2候補	第3候補
DPマッチング	音節全体	73.5	89.7	92.6
	子音部	83.8	92.6	97.1
中間累積距離	u_1：最適値	88.2	95.6	97.1

男性話者1名，68単音節

3.2 認識実験

3名の男性話者が，毎分 70～110 音節の発声速度で，アイウエオ順に発声した 10 回分の 68 音節を用いて認識実験を行った．発声はすべて1時間以内に発声されたものである．1回目の発声を標準テンプレートとし，2回目以降を入力パターンとして，中間累積距離マッチング法を用いて認識実験を行った結果を**図4**に示す．図4は経時に伴う認識率の変動を示している．

1テンプレートのみの使用により，平均 87.2% 第

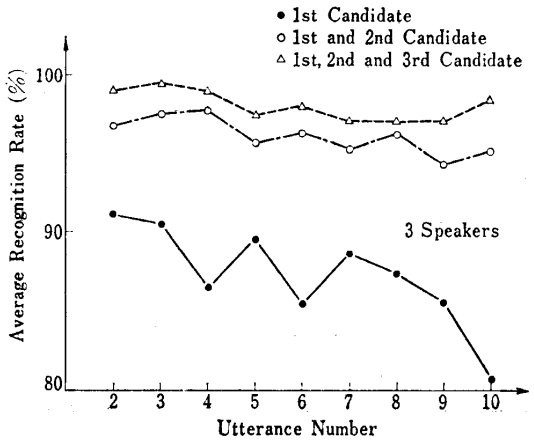

図 4　中間累積距離マッチング法による認識実験結果
Fig. 4　Recognition results by ICD method.

2候補まで含めるならば平均96.0%の認識率を得た．これは日本語における音節の出現頻度を考慮した場合，第1候補で平均90.7%の認識率に相当する（都市名，人名，新聞，小説，会話等から抽出した約27,000音節の分布に基づいて算出）．誤認識の傾向としては，/p/→/b, t/, /d/→/t, b/, /h/→/b, k/, /cu/→/su/, /a/→/ba/, /r/↔/b/等，破裂音に関する誤りが多かった．誤認識の原因としては呼気の混入，語頭の切り出し誤り，有声破裂音の語頭の無声化等があげられる．

4. 中間累積距離と音節間類似度を併用した認識

音声認識実験の結果得られる（中間）累積距離は，入力音声の種類（主に子音と母音の占める割合等）に依存して大きく変化する．このため認識実験結果を入力音声によらず対等に評価するためには，入力音声ごとに累積距離の正規化を行う必要がある．本論文では式(1)に示すように，第n候補の累積距離を第1候補の累積距離によって正規化し，これを音節間類似度と呼ぶ．このような尺度については，先に中川ら[4]によって，不特定話者認識における母音候補の削減に有効であることが示されている．

$$S_n = D_n / D_1 \quad (n=1\sim4) \qquad (1)$$

S_n：音節間類似度，D_n：第n候補の累積距離

4.1 テンプレートの教師付学習

発声の変動に伴う認識精度の劣化に対処するため，入力音声の教師付学習方法について検討する．基本的には，既存のテンプレートに対して，入力音声の追加，あるいは平均化を行うが，以下に示すような原因で，テンプレートの品質が大幅に劣化する可能性がある．

（1） 使用者の誤操作（出力候補選択あるいは訂正の際のキー操作の誤り）

（2） 不良発声（ここで言う不良発声には，呼気の混入，セグメンテーション誤りの他，有声音の無声化等の発声変動の大きな音声も含む．）

ここでは音節間の類似度に基づく学習方法により，(1)に対しては入力音声の却下，また(2)の多くについては入力音声の平均化を避けるようにする．

4.1.1 基準テンプレートの作成方法

学習を行うにあたって，キー操作の誤りや，不良発声を検出するための判断基準をあらかじめ明確にしておく必要がある．ここではまず，初期登録用として68音節の発声を3セット用意し，同名の音節相互間の累積距離を比較することによって，呼気混入や大幅なセグメンテーション誤りを生じた音声を検出する．このような音声の除去を行った後，残りの音声に対してDPマッチングの最適パスに沿った平均化を行い，これを学習のための基準テンプレートと呼ぶ．

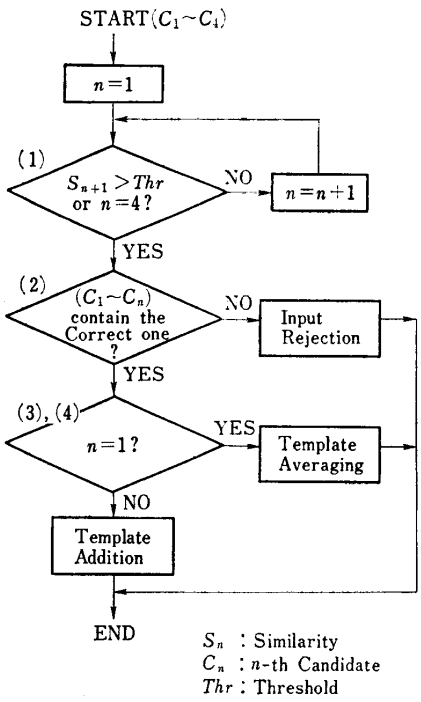

図5 テンプレートの学習方法
Fig. 5 Learning process for creating templates.

4.1.2 教師付学習方法

入力音声が，テンプレートの与える空間の中で占める位置によって，入力音声の品質を判断し，学習方法（平均化；追加，却下）を選択する．この判断の尺度として式(1)のような音節間の類似度を用いる．本学習方法の手順を以下に示す（**図5**）．

（1） 音節間の類似度に基づいて候補を絞りこむ．これは入力音声から一定の範囲内のテンプレートを求めることに対応する．この方法については4.2節で詳述する．

（2） この範囲に，正答が含まれない場合，使用者の誤操作あるいは完全な不良発声であると考え，入力音声を却下する．

（3） 候補が一つだけに絞られ，かつそれが正答である場合には，記憶量および計算量の節約と，テンプレートの適応化のため，DPマッチングの最適パスに沿った平均化操作を行う．

（4） 複数の候補が出力され，かつその中に正答が含まれる場合，各音節クラスタの境界付近のテンプ

入力音声：長浜(NA GA HA MA) --> 発声順

第1候補	WA	KA	HA	MA
第2候補	NA	GA	BA	NA
第3候補	MA	PA	YA	RA
第4候補	YA	TA	RA	WA

図 6 候補音節列
Fig. 6 Candidate monosyllables.

レートを充実させるために，この入力音声を新たなテンプレートとして追加する．

4.2 候補音節列の最適化

言語処理部の負担軽減と認識精度の向上を目的として，単音節認識の結果得られた候補音節列の最適化を図る．ここでは，音節間類似度に基づく候補音節の絞りこみと，音節の局所的な特徴による候補の再評価を行っている．

まず，音声認識部から得られる第4候補までの候補音節列は，それぞれ式(1)に基づいて第1候補との音節間類似度が計算され，閾値以下のものが**図6**に示すような候補音節列として出力される．ここで第4候補までを用いたのは，第3章での実験から，認識率がほぼ収束すること，発声変動に対しても高い精度を維持できることを確かめたためである．

図7に，本方法と，累積距離が閾値以下のものをすべて出力する方法との比較を行った結果を示す．なお図中の数字は使用した閾値を示している．この図よりわかるように，式(1)に基づく本方法では出力候補音節数に対する認識率の収束が早く，候補絞りこみの効果が大きい．また，話者間における最適な閾値の差も小さいと言える．

次に，先の候補絞りこみで複数の候補が出力された場合のみ，これらの出力候補に対して，**表3**に示すような語頭部の局所的な特徴による再評価を行い，出力候補音節列の最適化を図った．このとき，正答の脱落が新たに生ずることを避けるため，候補音節内での候補順位の修正のみを行っている．

4.3 認識実験

本認識システムの有効性を確認するため，第3章で用いたのと同じ，話者3名 10 回分の発声を用いて，認識実験を行った．基準テンプレートは最初の3回分の発声を用いて作成した．また，学習及び候補絞りこみに用いている閾値については，4.2 節での予備実験の結果から 1.2 とした．

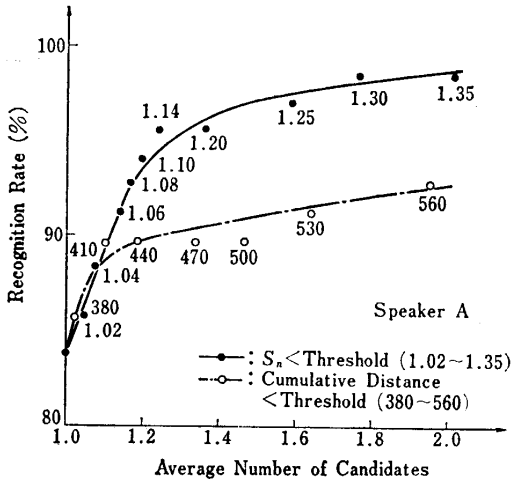

図 7-1 候補絞りこみ方法の比較 (1)
Fig. 7-1 Comparison between the candidate reduction methods (1).

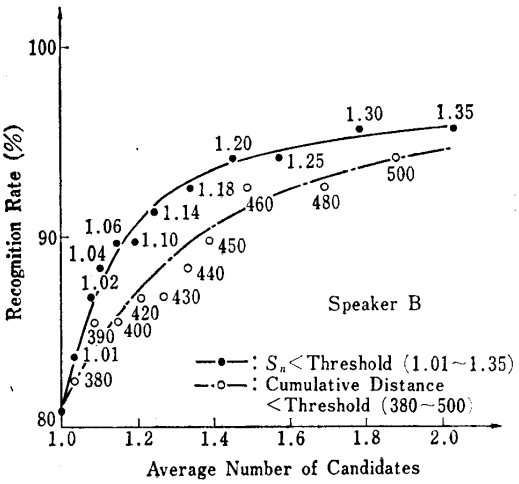

図 7-2 候補絞りこみ方法の比較 (2)
Fig. 7-2 Comparison between the candidate reduction methods (2).

表 3 候補再評価方法
Table 3 Methods for revaluating candidates.

順序	実行条件	評価方法	特徴量	主な識別対象
1	複数候補中に(V)を含む	中間累積距離マッチング	低域スペクトル(~1.8kHz)	(V)↔b (V)↔p
2	複数候補の出力	重心法によるクラスタリング	語頭パワーの2次微分量	k↔h t↔s
3	複数候補の出力	有声・無声判定	語頭スペクトル形状	p↔b t↔d

(V): 単母音

まず，以下に示す3方法について認識実験を行い，学習の効果を見た．

(方法1) 1回目の発声のみを標準テンプレートと

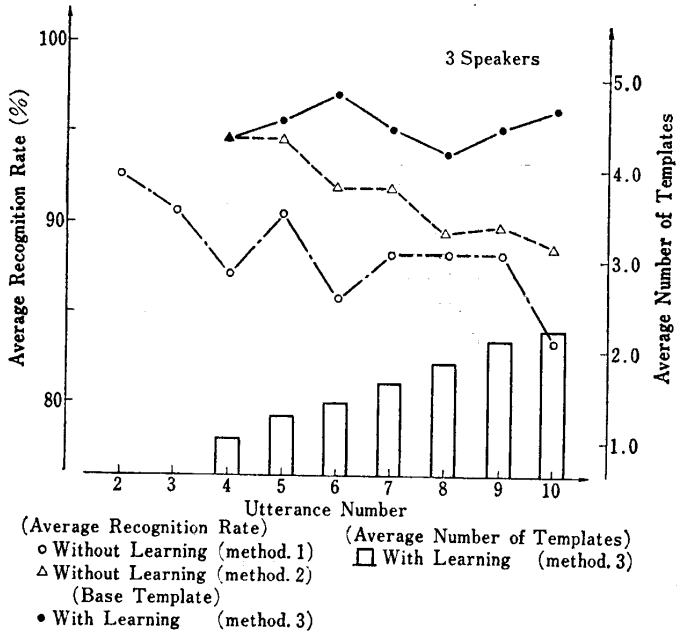

図 8 教師付学習方法による認識実験結果
Fig. 8 Recognition results by a learning method with a teacher.

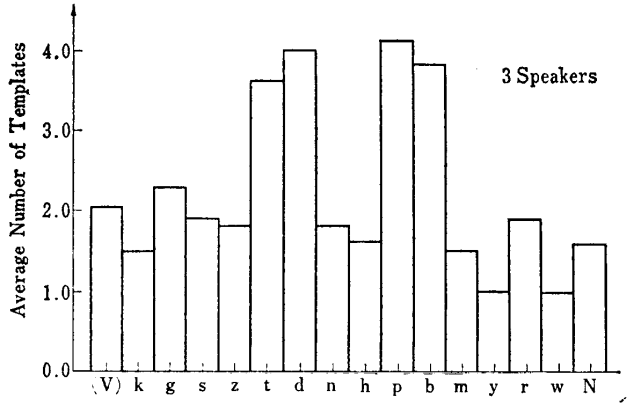

図 9 音素別テンプレート数 (10回発声時)
Fig. 9 Average number of templates for each phoneme (at 10 utterances).

し，2～10 回目を入力音声とする．

(方法2) 基準テンプレートを標準テンプレートとし，4～10 回目を入力音声とする．

(方法3) 学習開始時には基準テンプレートを標準テンプレートとする．入力音声（ここでは 68 音節音声のセット）を認識するごとに，その結果に基づく学習操作を行って新たな標準テンプレートを作成し，次の入力音声の認識に用いる（4～10 回目を入力音声とする）．

図 8 に認識結果を示す．また，学習終了時の音素別テンプレート数を**図 9** に示す．本学習方法を用いた結果，7回分の入力音声に対して平均 95.3% の認識率を得，経時に伴う認識率の低下が防止されている．特に，発声が不安定で，不良発声を生じやすかった/p/, /t/, /b/, /d/に関して，多くのテンプレートが追加されたのが効果的であった．なお，テンプレートを削減するアルゴリズムが導入されていないため，平均テンプレート数は単調増加するが，その傾向は図8に示すように緩やかで(1回の学習あたり平均0.2テンプレートの増加)，各テンプレート中で上位の候補として使用される頻度の低いものを削除すればよいと考えている．また，学習の内容は，1回の 68 音節入力に対し，平均化が 78%，追加が 20%，却下が 2% 程度であり，却下の大部分は呼気の混入によるものであった．

一方，候補絞りこみを行った結果，平均候補音節数 1.24 で平均累積認識率 98.3% が得られた．実験の結果を，学習を行わなかった場合との比較の形で，**図 10** に示す．候補音節列から単語候補を出力する場合，単語の音節数 L，一音節あたりの候補音節数を N とすると，候補となりうる単語の数は N^L で与えられるから，これは4音節からなる単語を想定した場合，出力候補単語数 2.36 ($=1.24^4$) 語に相当する．この場合の認識精度は，第2候補ないしは第3候補までの累積認識率に相当するが，第2候補まで出力する場合の候補単語数は 2^4 で与えられることから，候補単語数は約 1/7 ($=2.36/2^4$) 以下に削減されたと言うことができる．

表 4 にこれらの実験の結果をまとめておく．

最後に，本論文で提案した認識方法が，実際の単語あるいは文章の入力に対して，どの程度有効であるかの目安を与えるため，単音節単位発声された 100 都市名の認識実験を行った．標準テンプレートとしては認識実験で示した（方法3）を，10 回分の 68 音節発声に適用した結果得られたものを用いた．実験の結果，音節の認識率としては第一候補で 92.5%，平均候補音節数 1.30 で 98.5% の認識率を得た．音節の出現頻度のばらつきがあるため厳密に 68 音節のセットに対して得られた結果と比較することはできないが，誤認識率および，同レベルの認識精度を得るために必要な候補音節数がともに若干増加した．これは主に，入力音声の音節環境が多様化したため，調音結合の影響が現れたものと考えられる．

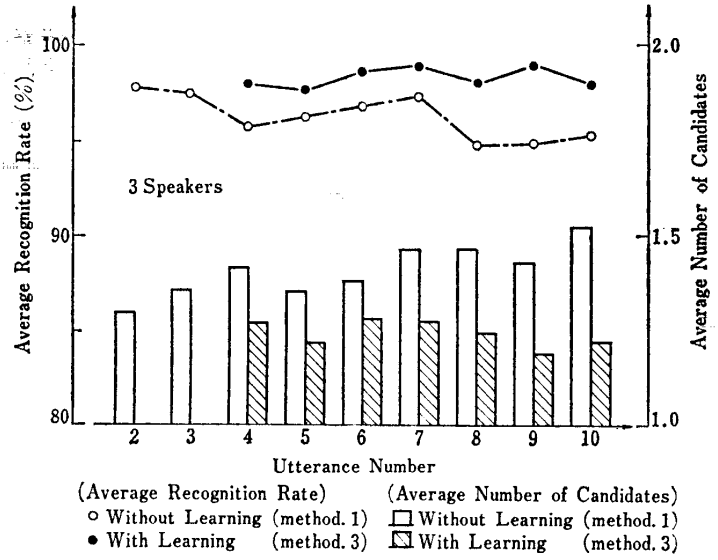

図 10 候補絞りこみの効果
Fig. 10 Effect of candidate reduction.

表 4 本システムによる音声認識実験結果（まとめ）
Table 4 Summarized recogniton results.

候補音節数	累積認識率 (%)
1	95.3
2	98.0
3	98.7
1.24(平均)	98.3

男性話者3名，68単音節

5. おわりに

本論文では，日本語音声ワードプロセッサを目的として開発した，特定話者用単音節音声認識方式の検討を行った．

まず，基本となる単音節の認識方式としては，DPマッチングの中間累積距離によるマッチング法を提案した．この方法は子音・母音間の正確な抽出を必要とせず，簡単なアルゴリズムで実現できることを特徴としている．実験の結果，この方法だけでは発声の変動に十分対処しきれないものの，話者3名，10回分の68音節の発声に対して1テンプレートのみで平均87.2%という高い認識精度を示した．

また，発声の変動による認識精度の劣化に対処する方法として，音節間類似度を用いたテンプレートの教師付学習方法を提案した．この方法は，入力音声の発声不良や，使用者の誤操作にも効率良く対処した方法であり，1回の学習について平均0.2テンプレート程度の追加により，認識精度の劣化が防止されることを示した．

さらに，音声ワードプロセッサ実現のために，言語処理部において必要となる辞書検索などの処理を軽減し，単語認識精度の向上を図るために，音声認識部から出力される候補音節列の最適化を試みた．

これらの方法を併用することにより，男性話者3名が68音節を10回発声したデータに対して，平均テンプレート数1.6で第1候補の平均認識率95.3%，平均1.24個の候補音節の出力時には，平均98.3%の累積認識率を得た．これはすでに提案されている種々の認識手法[2),3),5)~7)]と比較しても，十分な認識精度を与えているものと思われる．

今後は，言語処理部と音声認識部との整合を図った上で，実使用下での本システムの有効性について検討を行う予定である．

謝辞 本システムの開発にあたって貴重なご援助，ご助言をいただいた金子担当，手塚氏，狩野氏をはじめとする音声グループの諸氏に深謝いたします．

参考文献

1) Itakura, F.: Minimum Prediction Residual Applied to Speech Recognition, *IEEE Trans.*, Vol. ASSP-23, No. 1, pp. 67-72 (1975).
2) 吉田，迫江，千葉：日本語単音節音声認識実験，日本音響学会講演論文集，3-2-16 (1979.6).
3) 古井：単音節認識とその大語彙単語音声認識への適用，電子通信学会論文誌，Vol. J 65-A, No. 2, pp. 175-182 (1982).
4) 中川，中本：不特定話者の単音節単位入力による大語彙単語音声認識，電子通信学会論文誌，Vol. J 65-D, No. 12, pp. 1558-1565 (1982).
5) 似鳥，伊福部，吉本：単音節音声の実時間認識装置，日本音響学会誌，Vol. 39, No. 2, pp. 75-81 (1983).
6) 松井，山田，渡辺，藤本，佐藤：再照合を用いた実時間単音節音声認識システム，日本音響学会音声研究会資料，S 83-56 (1983).
7) 楠原，松井，相良，前原：単音節認識における辞書構成の検討，日本音響学会音声研究会資料，S 83-71 (1984).
8) Matsuda, Y., Tezuka, S., Kanoh, M., Nishimura, M., Kaneko, T.: A Method for Recognizing Japanese Monosyllables by Using Intermediate Cumulative Distance, Proc. IEEE 84 ICASSP (1984).

9) 西村，松田，手塚：単音節認識における発声上の諸要因に関する検討，情報処理学会第28回全国大会講演論文集，2L-5 (1984.3).
10) 西村，松田，手塚：中間累積距離を用いた単音節の認識実験，日本音響学会音声研究会資料，S 84-23 (1984).
11) 並木，浜田，中津：音声認識を用いた日本語入力方式，電子通信学会論文誌，Vol. J67-D, No. 4, pp. 405-412 (1984).
12) 嶋田，津田，三橋，平塚：音声入力用かな漢字変換，情報処理学会第29回全国大会講演論文集，5J-9 (1984.9).
13) 村田：音声入力による文章作成に関する一考察，情報処理学会第29回全国大会講演論文集，3J-3 (1984.9).
14) 斉藤，大深，大河内：音声入力のための仮名漢字変換法，情報処理学会第30回全国大会講演論文集，1G-6 (1985.3).

(昭和60年3月7日受付)
(昭和60年7月18日採録)