多段階分割復元法による誤りの多い文字列からの原文の復元†

荒木健治†† 宮永喜一††† 栃内香次†††

一般的な日本語文を音声により入力する場合には数万単語を対象とせねばならないが，これは，音声の個人差，時期変動などにより，現在は困難である．この問題に対して，連続音声から単音節を切り出す手法の研究が行われているが，現状では，その認識結果に誤りを多く含むことは避けられない．そこで，本論文では，日本語音声を単音節単位で認識し，その結果得られた単音節列の誤りを復元しながら単語単位に分割するという手法を提案し，さらに本方式の前提となる認識結果から候補文字をある範囲に限定できることを確認した後，本方式を用いた実験システムを開発し，その性能評価実験を行った結果について述べたものである．本方式は，音声認識によって複数の文字候補を取り出し，その組合せによって得られる多数の単語候補について確実性の高いものから順次段階的に確定していく方式である．実験により複数回出現する誤りが全誤りの 90% を占めることがわかり，本方式による日本語文書の音声認識における誤りを復元するシステムを開発した．さらに，情報処理分野の学術文献を資料として性能評価実験を行った結果，音声認識の正認識率が 67.0% のものを 85.9% へ 90.0% のものを 96.2% に復元できることがわかり，本方式の有効性を示すことができた．

1. はじめに

日本語文を計算機へ入力する手段として，現在はキーボードからローマ字あるいはかなで入力し，その後かな漢字変換によって漢字かな混じり文に変換する方式が主流である．しかし，この方式はキーボードの操作に習熟する必要があり，これに代わる方式の一つとして，音声入力の実用化が望まれている．

一般的な日本語文を取り扱うものとすると数万単語を対象とする音声認識が必要であるが，音声の個人差，時期変動などにより，現在のところ実用レベルで数万単語の単語認識を行うことは困難である[1]．一方，最近，連続音声を単音節に切り出し，単音節単位で認識する手法の研究が進められている[2),3)]．しかし，この方法では文字レベル（単音節認識結果）で多数の曖昧さ（誤り）が存在するので，それらの組合せによりきわめて多数の単語候補が出現する．そこで，これら多数の候補単語の中から正しい単語を一意に決定する必要がある．しかし，このような誤りの多い文字列に対しては，従来の文法規則，意味情報を用いた構文解析および意味解析を適用することは困難である．

ところで，人間がこのような誤りを含む文字列から原文を復元する場合は，すべての可能性を総当たりに考えるのではなく，比較的少数の手掛りとなる部分を最初に決定し，前後のつながりに矛盾がなければ同様の操作を順次続けることによって復元を進めていると考えることができる．これに基づき，本論文では連続音声より切り出される単音節認識結果より（誤りを含む）複数の文字候補を取り出し，その組合せによって得られる多数の単語候補について確実性の高いものから順次段階的に確定してゆく方式を提案する．以下，この方式を多段階分割復元法と称する．

2. 多段階分割復元法

2.1 概　　要

多段階分割復元法は，音声認識結果より得られる複数の文字候補を対象に，文字列中で前後の部分と重なり合うことなく一意に単語を確定できる部分から順次，段階的に単語を決定していく手法である．すなわち，まず単音節認識結果から認識情報辞書を用いて候補文字列を取り出し，次いでこの組合せによって得られる多数の単語候補の中から最適な単語を選択し，一意に決定する．以下，本論文ではこれらの処理を復元と呼ぶ．多数の単語候補の中から単語を決定する際には，文字列中で前後の部分と重なり合うことなく一意に決定できる部分から順次段階的に復元を行ってゆく．ここで，一番最初の段階で使用される単語をキーワード（以下，KW と略記する）と呼ぶ．音声認識により得られたべた書き文字列は最初に KW によっていくつかのブロックに分けられる．KW はべた書き文

† Multi-Stage Segmentation and Recovery Method for Recovery of Sentences from Erroneous Character Strings by KENJI ARAKI (Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai Gakuen University), YOSHIKAZU MIYANAGA and KOJI TOCHINAI (Department of Electronic Engineering, Faculty of Engineering, Hokkaido University).

†† 北海学園大学工学部電子情報工学科
††† 北海道大学工学部電子工学科

字列中で前後の部分と重なり合うことなく確実に分離でき，かつ高頻度の語であることが条件となる[4]．KW は，あらかじめ例えば表 1 に示すような一定量の資料より抽出され，KW 辞書に蓄えられている．なお，このような KW は，対象分野を限定した場合にのみ存在する．KW によって分割された各ブロックに対し，以下，KW の拡張処理，カタカナ語，一般単語，接辞，助詞，一字漢字語の順に復元処理を進める．また，本方式は，対象分野を限定する必要があるので，対象分野が変化すると，性能が低下する．そこで，辞書の自動更新[4]により使用につれて辞書を対象分野に適応させ，この問題を解決している．

本方式は，べた書き文のかな漢字変換法として著者らが先に提案した多段階分割法[4]に基づいており，これを音声認識において発生する文字単位の誤りの復元に拡張したものである．本手法では，上位の階層で単語を決定することにより下位の階層の単語候補の数は急速に減少する．それゆえ，数万語を対象とし，しかも音声認識結果に曖昧さを含む音声入力テキストプロセッシングにおける単語分割に有効である．

2.2 予備実験[5]

前述のように，現在の音声認識技術では，音節の認識において相当数の誤りが発生する．したがって，逆にある認識結果が得られた場合，入力音節は一意に決まらず，複数の候補が存在することになる．ここで，あらかじめ認識誤りの発生の傾向を把握しておけば，候補を限定することが可能になる．そこで多段階分割復元法による音声入力かな漢字変換の実験を行う前に，認識誤りの傾向を調査した．その結果，複数回出現する誤りが全誤りの 90% を占めることが確認された．つまり，同じ誤りが何回も出現することがわかり，文字候補をある程度限定できることが確認された．

3. 実験システム

3.1 システム構成

実験システムの構成を図 1 に示す．このシステムは，音声認識部および復元処理プログラムの二つのサブシステムからなる．実験に使用した音声認識装置は，連続音声より単音節を切り出し，認識する手法が確立されていないため，市販の単音節認識装置を用い

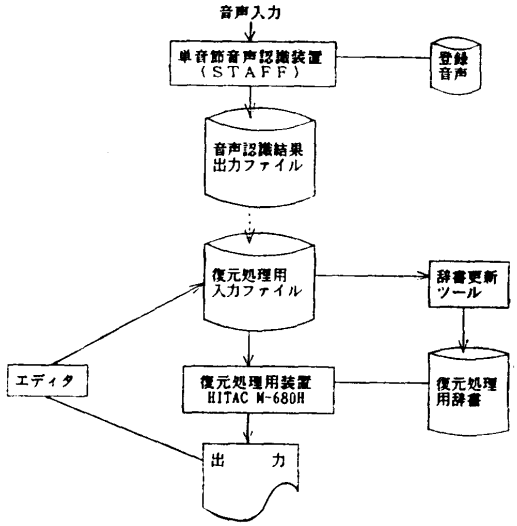


図 1 システム構成
Fig. 1 Construction of the experimental system.

表 1 辞書の作成に使用した資料
Table 1 Collected data for creating the dictionaries.

No.	論文名	著者名	巻号	文字数
1	高集積マイクロコンピュータに適したマイクロプログラム制御方式	前島他	Vol. 23, No. 1	10,594
2	COBOL マシンとその設計思想—ハードウェア構成について—	山本他	Vol. 23, No. 1	8,451
3	フーリエ変換を用いたテクスチュアの構造解析	松山他	Vol. 23, No. 2	7,723
4	日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析	木村	Vol. 23, No. 2	8,468
5	インテリジェント・コンソール—OS の機能拡張の一方法—	有田	Vol. 23, No. 3	9,343
6	ポータブル画像処理ソフトウェア・パッケージ SPIDER の開発	田村他	Vol. 23, No. 3	9,624
7	グラフィック・ディスプレイ・ターミナルのための端末作画システム	高藤他	Vol. 23, No. 4	6,423
8	オペレーティング・システムのファームウェア化対象選定法	長岡他	Vol. 23, No. 4	7,594
9	プログラム階層構造の生成，処理，文書化能力を有するテキスト・エディタ	酒井他	Vol. 23, No. 4	7,916
10	パステストに本質的な分岐に着目した網ら率尺度の提案	中所他	Vol. 23, No. 5	10,970
11	計算機システムにおける性能管理の一方式とそれを用いた実験	吉住他	Vol. 23, No. 6	9,123
12	高速パケット伝送路用前置処理装置の一構成法	寺田他	Vol. 23, No. 6	8,401
13	ソフトウェア生産過程の評価実験に関する考察	有澤他	Vol. 23, No. 3	6,453
14	文節数最小法を用いたべた書き日本語文の形態素解析	吉村他	Vol. 24, No. 1	9,658
	文字数合計			120,741

た．使用した装置は(株)ビー・ユー・ジー製マイクロコンピュータ STAFF とこれに接続して使用する音声認識基板からなり，特定話者用単音節認識型のものである[6]．また，復元処理プログラムは，北海道大学大型計算機センターの HITAC-M 680 H 上に作成されており，使用言語は PL/I である．なお，音声認識部と復元処理プログラムの間は現在はオフラインである．

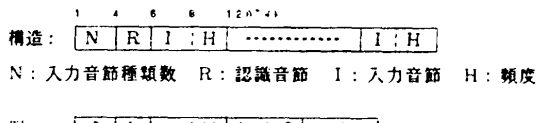


図 2　認識情報辞書の構造と例
Fig. 2　Construction and example of the recognition information dictionary.

処理過程としては，まず，発声された単音節を音声認識部で認識する．この結果はローマ字列の形で出力される．次に，このローマ字列を復元処理プログラムに入力し，多段階分割復元法により単語列に分割し，その結果を漢字かな混じり文で出力する．

3.2　復元処理用辞書

復元処理用辞書として，1種 KW 辞書，2種 KW 辞書，カタカナ語辞書，単語辞書，単語補助辞書，接辞辞書，助詞辞書，一字漢字語辞書，一字漢字語補助辞書，認識情報辞書の 10 種類を用いる．認識情報辞書を除く 9 種の辞書の構造は文献 4) で述べたものとほぼ同様である．KW は後述のようにその安定性によって1種，2種の2種類に分けられている[7]．**図 2** に認識情報辞書の構造とその例を示す．これは，単音節音声認識を行った結果から，認識音節Rに対する入力音節Iの組 (R, I) とその出現頻度を記録したものである．この例でいえば，認識結果が「do」に対する入力音節として，「ro, do, yo」の3種類が出現し，例えば「ro」であった出現頻度は，16回ということである．この辞書を用いて，認識結果より文字候補（入力音節候補）を取り出す．

①入力文字列

こうしゆうせきまいくろこんぴゆうたにてきしたまいくろふろぐらむせいぎよほうしき

②認識文字列

こうし？うせきまいくぼこ？びいうたにべきしたわいくぼふほぐらほせいぎよほもしぎ

正認識率 69．2%　注）下線部は誤認識、？は未認識を示す．

③認識情報辞書より得られる文字候補

認識結果	こうし？うせきまいくぼこ？びいうたにべきしたわいくぼふほぐらほせいぎよほもしぎ
文字候補	こうしちゆうるつくすちしよんきふねほさずへふ せきつ まいゆ くろどぼのよ こゆるつくすちしよんきふねほさずへふ びいゆ うたにとさ にべきててで したまちとさ まいわゆ くろどぼのよ ふろぐらふどぼのよ ほせいぎよむつゆき ぼとこもう ほもしぎよむうちき ぼとこもう

④原文の復元結果

a．結果

：高集積：[マイクロコンピュータ](に)：出来：@した@[マイクロ]@プログラム@@制御@@方式@

@@：1種KW　[ ]：2種KW　( )：助詞　：：一般単語

正認識率　97．4%

b．正解

高集積マイクロコンピュータに適したマイクロプログラム制御方式

図 3　復元の例
Fig. 3　Example of recovery.

3.3　復元処理アルゴリズム

3.3.1　処理の概要[5]

前述のように，入力された音節とその音節の認識結果との組がその頻度とともに認識情報辞書に記録されており，認識結果が与えられると各文字ごとに文字候補を取り出すことができる．この様子を**図 3** を例にして説明する．今，①のような文字列が単音節ごとに発声され，②のような認識結果が得られたとする．ここで，「？」は認識装置がどの音節か認識できなかった場合（未認識）を示し，また下線部は誤認識された部分を示す．この段階での正認識率は 69.2% である．この認識結果をもとに認識情報辞書から③のような複数の文字候補が得られる．この文字候補を組み合わせて多数の単語候補が得られるが，その中から単語を一意に決定することにより④ a のような復元結果が得られる．ここに付加した記号は，どの段階で復元されたかを示すものである．ここで，入力された原文は b であったので，復元された後の誤りは「適」を「出来」とした1カ所だけであり，正認識率は，97.4% に向上したことになる．

3.3.2　単語の決定

本項では，上述のようにして得られた文字候補を組み合わせて得られる多数の単語候補の中から単語を一意に決定するアルゴリズムについて述べる．

図4に，単語復元処理の流れを示す[8]．以下，各段階における処理の概要を述べる．

i） KW による分割，復元

まず最初に，KW による分割および復元を行う．KW は，例えば表1のような一定量の資料中において，べた書き文中にその読みが出現した場合無条件にその語に決定しても誤りとならないことがあらかじめ確認されている語である．KW は1種，2種の2種類に分けられている．初めに，文字候補を組み合わせて得られる単語候補中の1種 KW の有無を検査し，もしあれば KW 同士が重複してあてはまるものを除いて分割を行う[4]．なお，一方が他方全体を含む重複については文字数の多いほうをあてはめる．

次に，2種 KW について同様の処理を行う．2種 KW のあてはめには制限を設けているが，これについては3.3.4項で述べる．

ii） KW の拡張

次に，KW によって分割，復元された部分を拡張する．これは，KW を前後に拡張した文字列について，一致する語が単語辞書中にあればその語も KW と同程度に確実性が高いと考え，決定するものである．この際にも，認識結果から得られるすべての文字候補について一致を検査する．

iii） カタカナ語による分割，復元

カタカナ語には，外来語など総りに特徴があるものが多く，KW に次いで確実性が高いと考えられるので，次にカタカナ語による分割および復元を行う．

以上述べたように，この段階までは該当する単語が候補文字列中にあれば直ちにその単語であると決定している．

iv） 一般単語による分割，復元

次に，一般単語による分割および復元を行う．この場合は，種々の分割可能性を考慮する必要がある．そこで，単語辞書中の各単語ごとに付加されている連接情報を利用して単語を決定する．連接情報は，文章中でその単語の前後に連接する各1文字の組である[4]．連接情報の一致を検査する方法については3.3.6項で述べる．なお，数字，記号，句読点およびすでにあてはめ済みの単語で完全に挟まれた単語については上記の手順によらず，直ちに決定する．

v） 接辞の処理

次に，接辞の処理を行う．これは，すでに確定した漢字語，カタカナ語および数字が存在し，その前後の未変換部分が接辞の読みと一致する場合にそれを接辞とするものである．同一箇所に複数の接辞候補がある場合は，文字数の最大のものをとる．

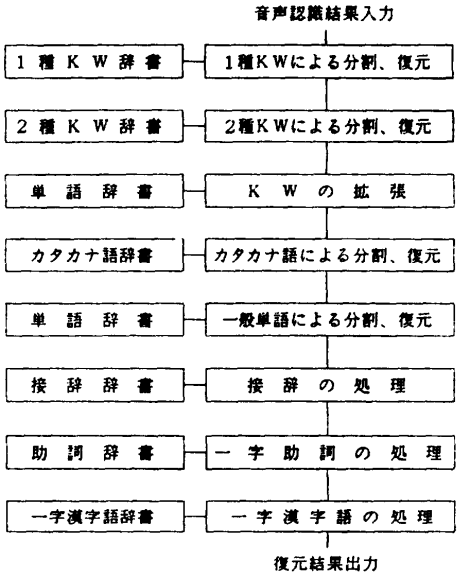


図4 単語復元処理の流れ
Fig. 4 Process of recovery.

vi） 一字助詞の処理

次に，一字からなる助詞の処理を行うが，これについては3.3.5項で述べる．

vii） 一字漢字語の処理

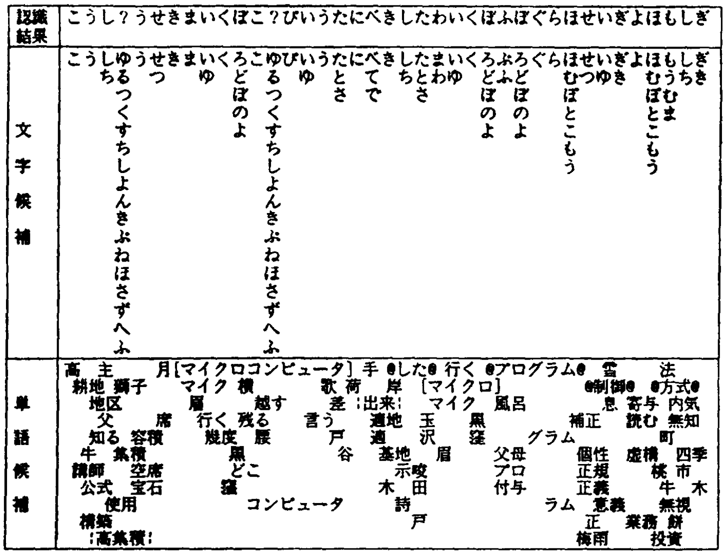
最後に，一字漢字語の処理を行う．一字漢字語は前後の文字との結合可能性が大きく，したがって独立性は最も低い．そこで，一字漢字語の処理は最後に行っている．一字漢字語の決定は一般単語と同様，連接情報を用いて行う．

単語復元処理過程の例を図5に示す．この図を見るとわかるように，単音節認識結果から認識情報辞書を用いて複数の文字候補が得られ，これらから得られる復元処理前の単語候補の数は 92 個である．ここで1種 KW を決定すると，1種 KW と重複する単語候補が除かれ，その数は 52 個に減少する．さらに，2種 KW を決定すると同様にして初期状態の 30% 程度の 30 個に減少する．したがって，種々の分割の可能性がある一般単語による分割および復元では対象とする単語候補の数は最初の候補数の 30% 程度で済み，単語の決定が容易となることがわかる．

3.3.3 辞書の更新

本手法は，対象分野を限定した場合に有効となる．そこで，辞書を対象分野および使用者に自動的に適応させる必要がある．辞書の自動更新は，校正済みの正

（a）単語候補の例



（b）単語候補数の推移

段階	処理結果	単語候補数
初期状態	こうし？うせきまいくぼこ？びいうたにべきしたわいくぼふぼぐらほせいぎよほもしぎ	92
1種KW	こうし？うせきまいくぼこ？びいうたにべき@した@わいくぼ@プログラム@@制御@@方式@	52
2種KW	こうし？うせき[マイクロコンピュータ]にべき@した@[マイクロ]@プログラム@@制御@@方式@	30
一般単語	¦高集積¦[マイクロコンピュータ]に¦出来¦@した@[マイクロ]@プログラム@@制御@@方式@	9
助詞	¦高集積¦[マイクロコンピュータ](に)¦出来¦@した@[マイクロ]@プログラム@@制御@@方式@	8

@ @：1種KW　[ ]：2種KW　¦ ¦：一般単語　( )：助詞

図 5 単語復元処理過程の例
Fig. 5 Example of recovery process.

しい字種指定文字列を用い，辞書中の単語について頻度，履歴の更新，誤変換数の更新，新語の登録，連接情報の更新等の処理を行い，さらにキーワード辞書，カタカナ語辞書を更新し[4]，この後に，認識情報辞書の更新を行う．なお，認識情報辞書の更新は，校正済みの正しい字種指定文字列とその音声認識結果の組を単音節単位にこの辞書に記録することによって行われる．

3.3.4 KW の階層化[7]

本方式では，最上位の階層である KW による分割および復元の精度が下位段階での単語の確定の際に影響を与えるので，この処理には，よりいっそうの正確さが要求される．そこで，KW による分割をより正確にするために，KW の階層化を行っている．すなわち，KW を対象とする文献にかかわらず安定なものと文献単位で変動する未だ不安定なものに分けて各々1種および2種とし，2種については，その適用に制限を設けている．

2種 KW は，その KW が抽出された時期によってグループに分類されている．抽出時期は，辞書更新の際に KW 候補として抽出される度にそのグループに属するとして抽出時期フラグに‘1’を与えることにより，記録される．

2種類の KW のうち，2種 KW については，抽出時期の同じ KW が，本システムではキーボードから入力している句読点，記号，空白によって区切られた処理単位中に二つ以上あてはまる場合のみあてはめている．

ところで，KW は最初は必ず2種 KW として登録され，その後，一定の条件を満たしたときに1種 KW とする．この1種 KW となるための条件を，以下に示す．これは，実験[7]により最適なものとして求められたものである．

条件 1：2種 KW として登録後，30 回使用された．

条件 2：誤り数／頻度≦0.05

辞書更新の際，校正済みの文字列と復元結果を比較することにより誤り数をカウントし，2種 KW の中で条件 1，2 を満たすものを1種 KW 辞書に登録する．**表 2** に表1の資料より抽出された1種 KW 辞書の登録語を示す．なお，1種 KW 辞書は，検索の効率を向上させるために見出し語のローマ字表記時の先頭のローマ字によりグループ分けされて登録されている．

3.3.5 一字助詞の処理

認識結果より得られる文字候補は一般に複数個である．それらの中には，一字で助詞になるものも当然出現するので，助詞候補の数が非常に多くなる．そこで，一字助詞の決定および復元は，文字候補が十分減少している「接辞の処理および復元」の後で行い，しかも前後があてはめ済みの単語および記号で挟まれた助詞候補のみを助詞と決定している．この様子を**図 6** に示す．図6で（a）はここで対象とした助詞の一覧，（b）は助詞を決定する一般的規則，また，（c）は助詞とする場合としない場合の例である．（c）の①はカタカナ語の「コンピュータ」と1種 KW の「適」で完全に挟まれているので助詞とするが，②では助詞候補

表 2 1種 KW 辞書の登録語
Table 2 Registered words of the first rank key-words dictionary.

A	読み	表記	読み	表記	読み	表記
d	である	である	では	では	でえた	データ
	できる	できる	だけ	だけ	であり	であり
g	がぞう	画像	がめん	画面		
h	ひつよう	必要	ひょうか	評価	ほうほう	方法
	へんかん	変換	ふく	副	ほうしき	方式
i	いる	いる				
k	から	から	こと	こと	けいさん	計算
	これ	これ	この	この	こうそく	高速
	こまんど	コマンド	かいせき	解析		
m	めいれい	命令	もの	もの	まくろ	マクロ
	もじゅうる	モジュール	めもり	メモリ	まで	まで
	もんだい	問題	もち	持ち	める	める
n	なく	なく	ので	ので	にゆうりょく	入力
o	おこな	行	おいて	おいて	おける	おける
p	ぷろぐらむ	プログラム				
r	れべる	レベル	りよう	利用		
s	する	する	しすてむ	システム	した	した
	その	その	しめ	示	される	される
	そふとうえあ	ソフトウェア	それ	それ	されて	されて
	さくせい	作成	せいぎょ	制御	せつけい	設計
	すべて	すべて	そんざい	存在	さぶるうちん	サブルーチン
t	ため	ため	として	として	ついて	ついて
	という	という	となる	となる	たいおう	対応
	つぎ	次	ている	ている	たいしょ	対処
	ていぎ	定義				
y	よつて	よって	より	より	よる	よる
	ような	ような				
z	じっこう	実行	じょうほう	情報	じつげん	実現

A：グループの別

「は」の後ろに，未変換文字列「ん」があるので助詞としない．

なお，2字の助詞は単語辞書に登録され，一般単語による分割および復元の段階で決定される．

3.3.6 連接情報について

連接情報についても，認識結果より得られる文字候補を用いる．以下，この処理手順を示す．

（1） 認識結果より得られる候補文字列から単語候補を得る．

（2） この単語候補について，その前後各1文字を取り，単語辞書を検索して該当する連接情報を得る．ここで，前後連接文字も一意に確定しないので，連接情報も複数種得られる．

（3） 各連接情報の頻度の和を求め，これを当該単語候補に関する連接情報の頻度とする．

（a）助詞辞書に登録された助詞

「の、は、に、が、を、と、や、で、も、し、な、へ、て、ば、か」

（b）助詞と決定する条件

A （助詞候補） B

A：決定済みの単語及び句読点を除く記号
B：決定済みの単語及び句点を除く記号

（c）例

①助詞とする場合

◎マイクロ◎｛コンピュータ｝に◎適◎◎した◎

助詞とする

②助詞としない場合

◎その◎｜救助｜はん｜メンバー｜◎より◎｜指名｜｜される｜

助詞としない

◎ ◎：1種KW ｛ ｝：カタカナ語 ｜ ｜：一般単語

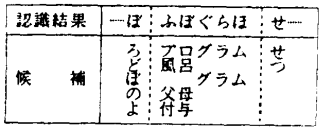
図 6 助詞の決定および復元の方法
Fig. 6 Method of particle determination and recovery.

（4） 単語候補のすべてについて上記の手順を適用し，連接情報の頻度が最大であるものを正解とする．

この例を**図7**に示す．この図は，図のような認識結果において多くの単語候補の中の一つである「プログラム」に対して得られる連接情報の頻度の求め方を示している．まず初めに，「プログラム」の前後の認識結果「ぼ」，「せ」に対して図のような文字候補を得る．その結果，①に示すような連接情報が得られる．次に，単語辞書中の「プログラム」の項の連接情報を参照することにより②のような連接情報とその頻度を得る．①，②を比較して一致するものは（ろ，せ），（の，せ）の二つなので，その頻度を合計することにより単語候補「プログラム」の連接情報の頻度は 17 となる．このような操作をすべての単語候補に行い，その連接情報の頻度が最大のものを採用する．

3.3.7 同音異義語について

本方式では，音声認識された文字列を単語分割および復元処理により字種指定文字列に変換し，これを字種指定方式によるかな漢字変換システムである KKH[9] に入力し，同音語の自動選択を行い，漢字かな混じり文に変換している．ここで，同音語自動選択に用いた手法は一般単語による分割および復元に用いた方法とほぼ同様であり，単語Wとその前後文字（または記号，数字，空白などを含む）a，b との三組 aWb の出現頻度により同音語を選択するものである[10]．例えば，字種指定文字列で「こうえん」という文字列が漢字と指定された時「公園」や「講演」という同音語が考えられる．ここで，「〜について講演する」とは言うが，「〜について公園する」とは言わない．そこで，「講演」と前後の一文字「こ」および「と」があらかじめその出現頻度とともに辞書中に記録されていれば，文字列と辞書中の連接情報を比較することにより「講演」に決定することができる．



①文字列中より得られる単語候補「プログラム」の連接情報

（ろ、せ）（ろ、つ）（ど、せ）（ど、つ）
（ぼ、せ）（ぼ、つ）（の、せ）（の、つ）
（よ、せ）（よ、つ）

②単語辞書中の単語候補「プログラム」の項の連接情報

前文字	ろ	の	と	も	り	を	ん
後文字	せ	せ	に	に	が	そ	ぐ
頻度	12	5	3	3	2	1	1

③得られる単語候補「プログラム」の連接情報頻度

（ろ、せ）　12　（の、せ）　5　合計　17

図 7 文字候補を用いた連接情報の適用例
Fig. 7 Example of the use for the triplet of the word with character candidates.

4. 性能評価実験

4.1 実験方法

実験の手順を以下に示す．

（1） 市販の単音節認識装置[6]を用いて音声認識結果を得る．資料は**表3**の資料を四つに分けて用いた．また，各資料を入力する前にあらかじめ単音節 68 音を発声して学習させる必要がある．

（2） 得られた認識結果の正認識率は，最低が表3，資料4の 60.7% で最高が資料1の 66.5% であったので認識誤りをランダムに訂正することにより，各資料の正認識率を 67，70，75，80，85，90%の6種類に変化させたデータを実験用に人工的に作成した．

（3） これらのデータを用いて性能評価実験を行った．

また，復元処理用辞書作成に用いた資料は，表1に示す情報処理に関する論文 14 編（120,741 文字）である．認識情報辞書を除く9種の辞書は，単語辞書のひらがな語[11]（1,089 語）以外の項目は，空の状態から，多段階分割法によるべた書き文かな漢字変換システム[4]を用いて論文1編ごとに更新を行い，その結果得られたものである．**表4**に，この9種の辞書の登録語数を示す．

また，**表5**に実験に用いた認識情報辞書の1文字に対する文字候補数の平均値と最大値を示す．認識情報辞書は，表3の各資料を復元する前に，あらかじめ各資料を音声認識装置に入力し，その認識結果と対応づけて記録したものである．すなわち，入力音節とその

表 3 性能評価実験に使用した資料
Table 3 Collected data for the performance evaluation experiment.

No.	章	音節数
1	第1章前半	639
2	第2章後半	888
3	第2章前半	721
4	第2章後半	705
	計	2,953

出典：「高集積マイクロコンピュータに適したマイクロプログラム制御方式」情報処理学会論文誌，Vol. 23, No. 1

表 4 辞書の登録語数
Table 4 Number of the words in the dictionaries.

辞書名	語数
1種KW辞書	74
2種KW辞書	350
カタカナ語辞書	200
単語辞書	3,262
単語補助辞書	200
接辞辞書	37
助詞辞書	15
一字漢字語辞書	116
一字漢字語補助辞書	8

表 5 認識情報辞書の認識結果1文字当りの文字候補数
Table 5 Number of candidate characters for a recognized character in the recognition information dictionary.

復元前の正認識率 (%)	文字候補数	
	平均	最大
67	2.4	22
70	2.3	22
75	2.2	19
80	2.0	18
85	1.8	15
90	1.7	13

認識音節を一組として出現頻度とともに情報を辞書に保存している．したがって，心ず正しい文字が認識情報辞書中に存在する．なお，1入力音節に対して複数の認識結果が得られる付加や何も認識されない脱落が誤認識として発生するが，ここでは誤った文字に置き代わるあるいは一つの文字に特定できない置換のみを対象として実験を行った．付加，脱落の問題については今後検討を進める予定である．

4.2 実験結果

表6に正認識率，誤認識率，未認識率および未分割率の推移を示す．ここで，正認識率，誤認識率，未認識率，未分割率の定義は，次式による．

正認識率＝正しく認識された音節数／全音節数

誤認識率＝誤って認識された音節数／全音節数

未認識率＝一意に特定できなかった音節数／全音節数

未分割率＝未分割による誤り音節数／全誤り音節数

表6より，復元前に90%程度の正認識率があれば，96%以上まで復元できることがわかる．また，誤認識率については，本手法を用いて誤認識の約1/2を正しく復元できることがわかる．また，未認識率については，未認識のほとんどすべてを何らかの文字と特定することができたことがわかる．また，本方式では，単語分割を行うと同時に復元を行うが，分割の条件を満たす単語が存在しなければ，単語分割を行わない．このために，もし，その部分に誤りが存在しても復元は行われない．未分割率とは，このような原因により復元処理後も残る未分割による誤りの割合である．

表 6 認識率の推移
Table 6 Change of the recognition rate.

復元前の正認識率	67.0	70.0	75.0	80.0	85.0	90.0
復元後の正認識率	85.9	87.2	88.4	90.3	93.5	96.2
復元前の誤認識率	23.4	21.0	17.5	13.8	10.2	7.0
復元後の誤認識率	14.1	12.6	11.6	9.7	6.6	3.9
復元前の未認識率	9.6	9.0	7.5	6.2	4.8	3.1
復元後の未認識率	0.2	0.3	0.2	0.1	0.1	0.0
未分割率	4.6	5.1	4.7	3.7	3.6	2.7

注) 単位は%

表 7 復元状況
Table 7 State of recovery.

正認識率	67%	70%	75%	80%	85%	90%
①正→正	61.9	65.2	69.8	75.3	81.8	87.8
②誤→正	24.8	22.8	19.0	15.2	11.9	8.5
③誤→誤	8.6	7.6	6.4	5.2	3.5	1.8
④正→誤	4.8	4.4	4.7	4.3	2.9	1.9

注) 単位は%
①復元前に正しかった文字を復元後も正しい文字として認識した．
②復元前に誤っていた文字を復元によって正した．
③復元前に誤っていた文字が復元によっても正されなかった．
④復元前に正しかった文字が復元によって誤りとなった．

表7に復元状況を示す．ここで，①～④の定義を次式に示す．なお，単語分割音節数とは，単語分割された部分の音節数で，表7に示す未分割の部分の音節数を全認識音節数から引いたものである．

①～④＝①～④のいずれかの音節数／単語分割音節数

4.3 考察

正認識率については，表6より現在の単音節音声認識技術で可能である90%程度の正認識率があれば，キーボード入力ワードプロセッサの実用基準といわれる95%以上[12]にまで誤りを復元できることがわかった．また，誤認識率，未認識率についても良好な減少を示していることがわかる．これは，多段階分割復元法が誤りの多い文字列から原文を復元するのに有効であることを示していると考えられる．単語分割されないために復元後も誤りとなる音節の割合は，復元前の全誤り音節数のうちの2.7～5.1%で良好な値を示している．

復元状況については，表7より単語分割を行うことにより復元された全音節のうち復元後正しい認識となる①と②の復元が占める割合が，復元前の正認識率を67%から90%まで上昇するにしたがって，86.7%から96.3%まで上昇している．また，全復元のうち復元後誤った認識となる③と④の復元は，13.4%から3.7%まで減少し，特に，正しいものを誤りに変えてしまう④の復元が4.8%から1.9%まで減少している．復元前の正認識率が，85%の時について見てみると，全復元のうちで復元後正しい認識となる復元(①と②の和)が占める割合は93.7%におよび，正確な復元を行っていることがわかった．

また，1種KWによる復元段階では，すべての文字候補を検索の対象とするので，文字候補の組合せの爆発が考えられる．この問題に対しては，文字候補の組合せによって作られるすべての単語候補を対象として1種KWを検索するのではなく，1種KWの読みの先頭のアルファベットによるグループごとに，文字候補中で1種KWの各語の読みを先頭より順に検索する．したがって，ある1種KWの読みがその先頭より検索して，1字でも文字候補中に存在しなければその1種KWは，単語候補中に存在しない．このような方法により検索を行うので文字候補の組合せの爆発という事態にはならない．しかし，このような方法を用いてもHITAC-M 680 Hを使用した表3, No. 1の復元処理のCPU時間は，表8に示すようにかなり長い．これは，先頭一字のアルファベットによるグループ分けだけでは曖昧さを含む文字列と辞書中の単語とのパターンマッチングの高速化には十分ではなかったためと考えられる．特に，未認識については非常に多くの文字候補が存在し，処理時間を大きく増大させている．また，今回の実験システムでは，辞書から単語を取り出し文字候補中でその読みを探す方式であるために，各段階で文字候補より単語候補を生成することや，辞書検索の方法が十分でないことも処理時間増大の原因と考えられる．

今回の実験では，約3,200語の単語を用いて実験を行ったが，本手法は，階層的に復元を進め上位の階層で正しい単語を決定することにより下位の単語候補の数が急激に減少するので，たとえ，辞書の収録語数が2万語程度になったとしても最上位の階層のKW辞書の語数は一定であり，単語辞書の語数が増えても下位の単語候補の数はさほど上昇しないので，その性能に変化はないものと考えられる．

表8 復元処理時間
Table 8 CPU time for recovery process.

復元前の正認識率 (%)	CPU 時間
90	2分39.60秒
85	3分32.83秒
80	4分42.26秒
75	4分33.63秒
70	5分33.63秒
67	6分01.96秒

注) 資料は，表3, No. 1に示す639音節からなる文章．

5. おわりに

音声による日本語文入力においては，連続音声の単音節認識結果に曖昧さがあるために，べた書き文の単語分割より非常に多くの単語候補が現れる．したがって，文字レベルでの曖昧さを解消し，そこから一意に正しい単語を決定する必要がある．本論文では，この問題に対して，音声認識の傾向を学習することにより，認識結果より文字候補を取り出し，その結果得られる単語候補の中から頻度統計に基づき確実度の高いものより順次，段階的に復元する方法を提案した．また，本手法は，連続音声を文節単位または単語単位に認識する音声認識手法に対しても文節候補または単語候補より正しい単語を選択する際の手法としても有効であると考えられる．

今後の課題としては，以下の3点が挙げられる．第1点としては，処理の高速化が考えられる．このためには，文字候補を減少させる必要があり，特に，文字候補の多い未認識については別処理を行うことや文字候補の出現確率に基づく絞り込みや辞書検索の高速化を行うことが考えられる．第2点としては，現在一字助詞の決定の際にしか用いていない認識情報辞書より得られる文字候補の頻度情報の活用およびその他の辞書中の各単語に付加された頻度，履歴，エラー情報を復元の際に活用することが挙げられる．第3点としては，現在は復元処理部のみで行っているシステムの使用者および対象分野への適応を，音声認識部においても行うようにすることが考えられる．これは，復元処理部や校正段階で訂正された情報を音声認識部にフィードバックし，システムがその情報を自動的に学習することにより，使用につれてシステムが使用者および対象分野に適応し，音声認識の能力も同時に向上させるというものである．これらについては，今後検討を進める予定である．

謝辞 本研究に際し，貴重なご意見をいただいた釧

路工業高等専門学校長永田邦一先生に感謝いたします．また，STAFF および音声認識基板を貸与していただいた(株)ビー・ユー・ジーに感謝します．なお，本研究の一部は文部省科学研究費補助金（一般研究(C)第 63580017 号）の補助により行われた．

参考文献

1) 中津良平：音声認識技術，情報処理，Vol. 24, No. 8, pp. 984-992 (1983).
2) 広重，宮永，栃内：知識工学的手法を導入した適応信号処理，第2回ディジタル信号処理シンポジウム講演論文集，pp. 337-342 (1987).
3) 溝口，田中，福田，辻野，角所：連続音声認識エキスパート・システム—SPREX，電子情報通信学会論文誌，Vol. J70-D, No. 6, pp. 1189-1198 (1987).
4) 荒木，栃内，永田：多段階分割法によるべた書き日本語文のかな漢字変換，情報処理学会論文誌，Vol. 28, No. 4, pp. 412-421 (1987).
5) 荒木，栃内，永田：誤りの多い文字列からの原文の復元，電子情報通信学会創立 70 周年記念総合全国大会講演論文集，6-126 (1987).
6) 伊福部：音声タイプライタの設計，CQ 出版社，東京 (1983).
7) 荒木，内田，山田，栃内，永田：キーワード方式べた書き文かな漢字変換システムの変換性能の向上，昭和 60 年度電子通信学会情報・システム部門全国大会講演論文集，3-169 (1985).
8) 荒木，宮永，栃内：多段階分割法による原文復元システムの性能評価，第 35 回情報処理学会全国大会論文集，pp. 1291-1292 (1987).
9) 栃内，伊藤，荒木，鈴木，永田：研究者向き日本語ワードプロセッサ KKHⅡの開発，北海道大学工学部研究報告，No. 109, pp. 119-126 (1984).
10) 栃内，伊藤，鈴木：前後連接文字を利用した同音語選択機能を有するかな漢字変換システム，情報処理学会論文誌，Vol. 27, No. 3, pp. 313-320 (1986).
11) 荒木，内田，山田，栃内，永田：キーワード方式べた書き文かな漢字変換システムの性能評価，第 32 回情報処理学会全国大会論文集，pp. 1169-1670 (1986).
12) 森，天野：日本語ワードプロセッサとテキストエディタ，電子通信学会誌，Vol. 63, No. 7, pp. 729-733 (1980).

（昭和 63 年 3 月 16 日受付）
（昭和 63 年 11 月 14 日採録）

荒木 健治（正会員）

昭和 34 年生．昭和 57 年北海道大学工学部電子工学科卒業．昭和 63 年同大学院工学研究科博士課程修了．工学博士．現在，北海学園大学工学部電子情報工学科助手．自然言語処理，音声情報処理の研究に従事．電子情報通信学会，IEEE，人工知能学会，日本認知科学会，ACL 各会員．

宮永 喜一（正会員）

1956 年生．1981 年北海道大学工学部電子工学専攻修士修了．工学博士．現在，北海道大学工学部電子助教授．並列計算機システム，ディジタル信号処理等の研究に従事．電子情報通信学会，日本音響学会，IEEE 各会員．

栃内 香次（正会員）

昭和 14 年生．昭和 37 年北海道大学工学部電気工学科卒業．昭和 39 年同大学院工学研究科修士課程修了．現在同工学部電子工学科教授．工学博士．自然言語処理，音声情報処理および信号処理プロセッサなどの研究に従事．電子情報通信学会，日本音響学会各会員．