音声言語情報処理 35－1
（2001．2．2）

最適な音声認識器選択手法を用いた音声入力インタフェース

加来 文伸　　森山 高明　　小川 均

立命館大学理工学部情報学科
〒525-8577 滋賀県草津市野路東 1-1-1

あらまし　ユーザが計算機に対し音声で指示する場合には，完全な構造の文ではなく，ユーザの指示の対象と，動作および対象の属性が伝われば良い．本研究では，このような単純な構造の文音声の音声認識システムとして，キーワードスポッティングおよび連続音声認識を組み合せることで高い認識精度を得る手法を提案する．提案システムでは，まずキーワードスポッティングによってユーザの指示対象を検出し，次にユーザの指示対象ごとに用意された専用連続音声認識器を用いて認識を行う．本研究では，タスクとしてマルチモーダル機器操作説明システムを用い，システムの音声入力として実装する．提案手法を適用したタスクにおいて評価実験を行った結果，単語正解率 88.8%，正解コマンド特定率 90.7% を得ることができた．

Selecting Method of a Suitable Speech Recognition System and its Applications

Fuminobu KAKU　　Takaaki MORIYAMA　　Hitoshi OGAWA

Department of Computer Science,Faculty of Science and Engineering,
Ritsumeikan University
1-1-1, Nojihigashi, Kusatsu, Shiga, 525-8577 JAPAN

Abstract　When a user orders to a system by speech, the user need not speak by a complete structural sentence, and only has to speak the instruction object, operation, and the attribute of the object. In this report, we proposed the speech recognition system for simple structural sentence, in which high recognition accuracy is obtained because continuous speech recognition is combined with a keyword spotting. The system is composed of the following two passes; The user's directed object is detected by keyword spotting in the first pass. The object, the operation, and the attribute of the object are recognized by the special continuous speech recognition system prepared for the object in the second pass. The instruction sentence used with the equipment operation explanation system was adopted as a task of the system proposed in this report. We obtained 88.8% word correct rate, 90.7% specified correct command rate as a result of the evaluated experimentation.

1　はじめに

現在，機器への入力に音声を利用したシステムが実用化し，一般に普及し始めている．例えば，カーナビゲーションシステムでは機器の操作を音声で行うことで，ドライバーの安全性，利便性を確保している．

このような音声認識システムに用いられる代表的な認識手法として，キーワードスポッティングや連続音声認識があり，音声対話 [1][2]，話題決定 [3]，音声の書き起こし [4] など，用途に応じて使い分けられている．しかしこれらの認識手法には各々長所，短所がある．例えばキーワードスポッティングは，

認識時に起こりやすい助詞の欠落・誤認識に対してロバストである一方，語彙の活用への対応が複雑である．また連続音声認識は，語彙の活用への対応が容易である一方，類似した語彙を誤認識する確率が高い．さらに，連続音声認識システムの汎用性を高めるためには，大量のテキストから音響モデル，言語モデルを学習または構築し，大語彙の単語辞書を作成する必要がある．したがって，認識対象となるテキストを増加させるためにはそれらの再学習，再構築が必要になる．

一方，特定の目的において使用語彙数および使用方法 (文法) が限定される場合がある．もし，文の対象が分かっていれば，その文中に出現する他の語彙が限定されることが多い．このような場合，対象語彙を限定した専用の音声認識器を用いれば，認識率の向上が期待できる．しかも認識対象となるテキストを変更，追加する場合，認識システム全体の辞書や音響モデル，言語モデルを変更することなくその対象専用の音声認識器を変更するだけで対応できると考えられる．

本研究では，文の対象をオブジェクトと捉え，このオブジェクト，および，これに関するメソッドと変数を専用に扱う連続音声認識器 (オブジェクト専用音声認識器) を用意し，オブジェクト専用音声認識器の選択にキーワードスポッティングを用いた手法を提案する．

2 認識対象システムと発話語彙の分類

本研究では，タスクとして操作説明のマルチモーダルプレゼンテーションシステム (以下，操作説明システム)[5] を採用し，提案する最適な音声認識器選択手法を用いた音声入力インタフェースを適用する．操作説明システムは電気機器の取り扱いを動画やテキストを用いて説明し，かつ，ユーザからの様々な要求（部品の場所指定，説明の速度調整）に対し，適切なコマンドを選択，適用する機能を持つ (図 1)．

ユーザの発話意図は発話に含まれる語彙の組合せから構成できる [6]．操作説明システムに対する要求を利用者が音声で行う際にも，その発話音声中に指示の対象，動作および対象の属性を示す語彙が含まれることが多い．したがって本研究ではこれらの要素をそれぞれオブジェクト，メソッド，変数というキーワードで表すことにする (図 2)．さらにシス

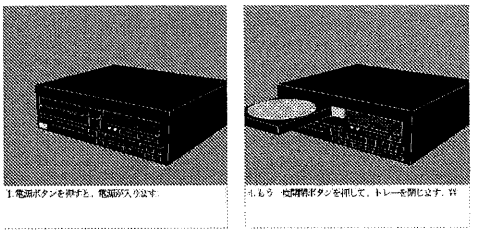

図 1: 操作説明システム

テム中ではオブジェクトに相当する語彙を一意に決定できる．そこでオブジェクト毎に複数のメソッドと変数を合わせて 1 つの群として扱う．例えば，操作説明システムにおいて説明が遅い場合，ユーザは「(説明の速度を) はやくして」といった発話を行う [7]．この場合，“説明” がオブジェクト，“はやくして” がメソッド，“速度” が変数に分類し，オブジェクトが “説明” では，変数には “速度” “1 番”，メソッドには “はやくして” “遅くして” “戻して” などが属し，1 つの群とする．同様にしてオブジェクトごとに群を用意する．

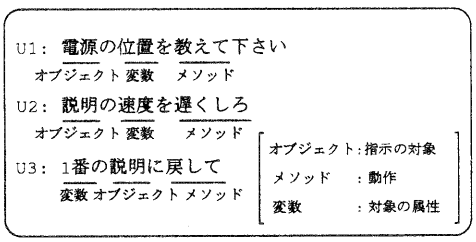

図 2: 発話語彙の分類例

3 オブジェクト専用音声認識器

大語彙連続音声認識は様々な状況で出現する語彙・文法を想定し，大量の語彙・文法データベースを構築，高精度の認識を実現する．しかし特定の状況で出現する語彙・文法が決まっている場合には，状況ごとに語彙・文法データベースを構築し，各状況でそれに合った語彙・文法データベースを用いて認識を行う方が，高い精度の認識が期待でき有効であると考えられる．

そこで本研究では，2 章で述べた各オブジェクト群に対応した連続音声認識器を作成する．すなわち，オブジェクト群に含まれる語彙を単語辞書に登録し (図 3)，オブジェクト群を認識するために必要な文法だけを記述した連続音声認識器である．これをオブジェクト専用音声認識器と呼ぶことにする．

3.1 オブジェクト専用音声認識器の作成

オブジェクト専用音声認識器はHTK[8]を用いて作成した．音韻モデルは情報処理振興事業協会(IPA)の独創的情報技術育成事業の研究成果物である「日本語ディクテーション基本ソフトウェア99年度版」[9]のmonophoneモデルを使用した．言語モデルは，本タスクに適応した単語辞書，文法を作成した．オブジェクト専用音声認識器の数は7，各オブジェクト専用音声認識器の語彙数は30から44となった．また，また同じ語彙が複数のオブジェクト専用音声認識器に属することもあり，最大で5つのオブジェクト専用音声認識器に属する語彙が現れた．

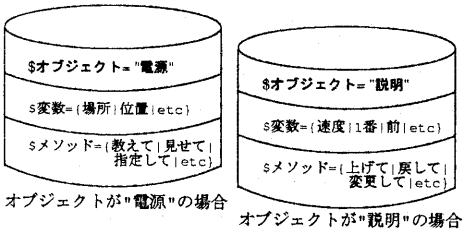

図 3: オブジェクト専用音声認識器の単語辞書例

3.2 オブジェクト専用音声認識器の有効性

認識時に，オブジェクト専用音声認識器を用いることの有効性を検証するため，事前実験として，オブジェクト専用音声認識器と全語彙・文法を網羅した認識器を用いて，単語正解率，単語正解精度による評価実験を行った．全ての語彙数は，7つのオブジェクト専用音声認識器をあわせ106となった．単語正解率 W_{CR}，単語正解精度 W_{AC} の定義をそれぞれ式(1)，式(2)に示す．音声データは，操作説明システムに対する命令調・依頼調30発話を被験者10名に発話してもらい実験に用いた．1発話あたりの平均単語数は5.3である．

実験の結果，オブジェクト専用音声認識器では単語正解率88.8%，全語彙・文法網羅の認識器では単語正解率77.5%であった(表1)．300発話中179発話は同じ認識結果，121発話で認識結果に違いが現れた．認識結果を考察したところ，全語彙・文法網羅の認識器では，正解単語を複数の助詞に置き換えた出力(図4)や，類似した語彙を出力する傾向がみられた．これら評価実験より認識時にオブジェクト専用音声認識器を用いることの有効性がいえる．

表 1: 各認識器の単語正解率と単語正解精度の比較

	N	S	D	I	W_{CR}	W_{AC}
OBJ	1640	107	102	53	88.8%	84.0%
ALL	1640	254	109	82	77.5%	72.9%

$$W_{CR} = \frac{N-S-D}{N} \times 100\% \quad (1)$$

$$W_{AC} = \frac{N-S-D-I}{N} \times 100\% \quad (2)$$

S: 置換誤り単語数　　I: 挿入誤り単語数
D: 脱落誤り単語数　　N: 全単語数
W_{CR}: 単語正解率　　OBJ: オブジェクト専用認識器
W_{AC}: 単語正解精度　　ALL:全語彙・文法網羅の認識器

a-1 正解認識例（オブジェクト専用音声認識器）
```
Read 43 physical / 43 logical HMMs
Read lattice with 91 nodes / 218 arcs
Created network with 612 nodes / 742 links
File: /Record/xxx.mfc
PAUSE NIBAN NO SETSUMEI NI IQTE KUDASAI PAUSE
...
```
a-2 認識誤り例（全語彙・文法網羅の認識器）
```
Read 43 physical / 43 logical HMMs
Read lattice with 111 nodes / 428 arcs
Created network with 835 nodes / 1155 links
File: /Record/xxx.mfc
PAUSE NI WA NO SETSUMEI NI IQTE KUDASAI PAUSE
...
```
図 4: 正解単語を複数の助詞に置き換えて出力した例
「2番の説明にいって下さい」

4 キーワードスポッティング

最終的に発話に合った最適なオブジェクト専用音声認識器候補の絞り込みを行い易くするため，発話中に存在する任意のキーワードを検出するキーワードスポッティングを用いる．キーワードスポッティングには，2種類の言語モデルのもとで入力発話を連続音声認識し，その結果を利用したスポッティング手法[10]を使用した．キーワード辞書にはオブジェクト・変数・メソッドあわせて95の語彙を登録した．

通常，キーワードスポッティングは予め閾値を設け，キーワードとの類似度を表すスコアが閾値以上であれば正解キーワード候補として検出する．スコアが大きいほどより正解キーワードらしいことを意味する．本研究では，キーワードスポッティングの認識結果から発話ごとのスコアにばらつきを観測し

たため，一定の閾値を設けずに，正解キーワード候補としてスコアの大きい順から相対順に検出する．相対順に正解キーワード候補をいくつ検出するかを決めるため，事前実験として被験者 10 名による 1 発話に 3 つのキーワード，オブジェクト・変数・メソッドを含んだ 30 発話を発話してもらい，キーワードスポッティングを行った．実験の結果，相対順位 2 までを検出する場合では，キーワード検出に失敗する発話が全体の 5%(300 発話中 15 発話) 存在し，認識器の絞り込みを行えない発話が多い．また 4 までで は，検出したキーワードのなかで湧き出し誤りキーワードの占める割合が，正解キーワードの占める割合を越え，認識器を絞り込む意図が反映されない (図 5)．そこで相対順位 3 までを検出することにした．

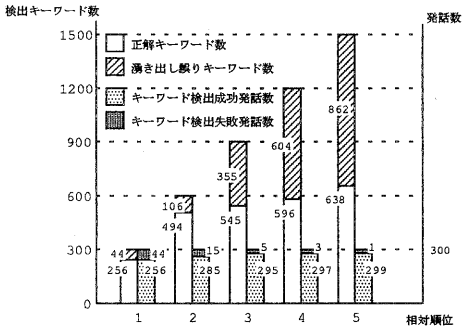

図 5: 相対順位ごとのキーワードスポッティング結果

5 音声認識器選択手法を用いた音声入力インタフェースの構成

5.1 最適な音声認識器選択手法

入力音声に対しどのオブジェクト専用音声認識器を用いて認識を行えば最適であるか，キーワードスポッティングを行い，得られたキーワードから判断する．例えば，得られた任意のキーワードがオブジェクトであれば，そのオブジェクト専用音声認識器を選択する．また，オブジェクト以外の変数・メソッドがみつかれば，それらが属するオブジェクトの専用音声認識器を選択する．この提案する手法を最適な音声認識器選択手法と呼ぶことにする．

5.2 音声認識器絞り込み段階 (第 1 パス)

第 1 パスでは，入力音声に対し予めキーワード辞書に登録しておいたオブジェクト・変数・メソッドのキーワードスポッティングを行う．得られたオブジェクト・変数・メソッドのいずれかより音声認識器選択手法を用いてオブジェクトを決定し，オブジェクト専用音声認識器を絞り込む．キーワードが得られない場合は，音声認識器の絞り込みができないため，全てのオブジェクト専用認識器を認識候補とする (図 6)．

5.3 コマンド特定段階 (第 2 パス)

第 1 パスで得られたキーワードだけでは文全体を判断できない．さらに絞り込まれたオブジェクト専用音声認識器から 1 つの正解候補を決めるため，第 2 パスでは改めて連続音声認識を行う．このとき認識器の数は第 1 パスで絞り込んだ数だけ用いるため，1 つの場合もあれば複数の場合もある．

まず，複数の認識結果の発話語尾に注目し，発話がシステムに対する要求を表す命令調・依頼調かどうかの判別を行う．命令調・依頼調であればコマンド実行の対象とし，それ以外であればコマンドの実行対象から削除する．

次に，コマンド実行の対象発話の認識結果を対応表で参照し，オブジェクト・変数・メソッドを検出する．ここで対応表とは，認識結果として現れた語彙がオブジェクト・変数・メソッドのいずれに属するか参照するものである．オブジェクト・変数・メソッドが決まらない発話は，検出できない各キーワードを出力する．

さらに，複数候補の中からオブジェクト・変数・メソッドが 3 つ決まり，助詞の数が 2 つにより近い候補を正解発話と判断する．これは発話を最適な音声認識器を用いて認識を行った場合，認識結果は正しく出力されるが，異なる音声認識器では，単語辞書にない語彙を単語辞書内の語彙から出力するため，類似した語彙での置き換えや，複数の助詞での置き換えが起こり，助詞の数が顕著に増えるといった特徴を利用したものである．

最後に，正解候補の 3 つのキーワードの組合せを対応表から 1 つのコマンドにマッチさせ，コマンドを実行する (図 6)．

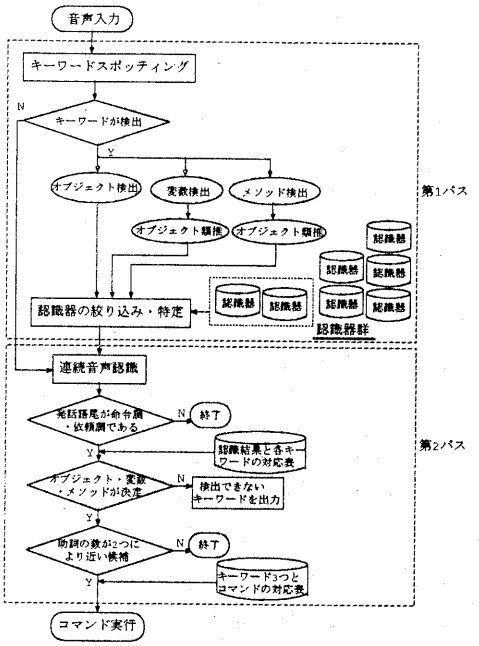

図 6: 音声入力インタフェースの処理フロー

6 評価実験

5 章で述べた音声認識器選択手法を用いた音声入力インタフェースを使用して，発話からのオブジェクト特定実験，コマンド特定実験を行った．音声データは，被験者 10 名の命令調・依頼調 30 発話を用いた．

6.1 オブジェクト特定実験

第 1 パスでキーワードスポッティングを行い，オブジェクトを検出，または得られた変数・メソッドからのオブジェクト類推による，オブジェクト特定の実験を行った．

表 2 にキーワードスポッティングで検出した正解キーワードの組合せと発話数の比較，表 3 にキーワードスポッティングにより絞り込んだオブジェクト専用音声認識器数と発話数の比較を示す．実験の結果，98.3%(300 発話中 295 発話) といった高い確率でキーワードスポッティングから正解オブジェクトを検出，類推できることが分かった．また 74%(300 発話中 222 発話) の確率で最適な 1 つのオブジェクト専用音声認識器を特定できることが分かった．

さらに認識結果を考察したところ，相対順位 3 までにオブジェクトが 1 つだけ現れた 206 発話中，198 発話でオブジェクトは正解であった．また相対順位 3 までにオブジェクトが 2 つ現れた 13 発話全てで，正解オブジェクトは 2 つの候補のうちのいずれかであった．これらを考慮し，第 1 パスでオブジェクトが 1 つに絞られた発話は，第 2 パスでそのオブジェクト専用認識器のみを用いて認識し，また 2 つに絞られた発話は，第 2 パスでその 2 つのオブジェクト専用音声認識器のみを用いて認識を行う．

表 2: 相対順位 3 でのキーワードスポッティング結果

正解キーワード検出の組合せ	発話数
オブジェクト・変数・メソッド	36
オブジェクト・変数	26
オブジェクト・メソッド	117
変数・メソッド	30
オブジェクト	32
変数	11
メソッド	43
正解キーワード無し	5

表 3: 絞り込んだ音声認識器数と発話数の比較

認識器数	1	2	3	4	5	6	7	誤り
発話数	222	22	1	0	8	27	10	10

6.2 コマンド特定実験

実験では，第 1 パスで絞り込んだオブジェクト専用音声認識器を用いて，それぞれ認識結果を出力し，最も認識結果が正解候補と思われる発話のオブジェクト・変数・メソッド 3 つから特定したコマンドの正解率を調べた．また最適な音声認識器選択手法を用いることの有効性を調べるため，オブジェクト専用音声認識器でのコマンド特定と，全語彙・文法網羅の音声認識器でのコマンド特定を比較した．

表 4 にオブジェクト専用音声認識器でのコマンド特定率を，表 5 に全語彙・文法網羅の認識器でのコマンド特定率を示す．実験の結果，全語彙・文法網羅の音声認識器では 59.3%であったが，オブジェクト専用音声認識器では 90.7%という高い確率でコマンドを特定できることが分かった．

また追加実験として，命令調・依頼調以外の単にシステムに対しての感想を述べた発話など，コマンド対象外の 6 発話を被験者 10 名に発話してもらい，コマンド対象外と判別できるか実験を行った．発話内容は「説明の速度が早い」「説明の速度が遅いな」「電源の場所が光るのでは」「電源の場所が光るらしい」などである．

表 4: コマンド特定率 (オブジェクト専用音声認識器)

コマンド特定		発話数
成功		272(90.7%)
失敗		28(9.3%)
失敗原因	認識器選択誤り (第 1 パス)	10
	キーワード欠落 (第 2 パス)	11
	キーワード剰余 (第 2 パス)	6
	語尾に助詞挿入 (第 2 パス)	1

表 5: コマンド特定率 (全語彙・文法網羅の認識器)

コマンド特定		発話数
成功		178(59.3%)
失敗		122(40.7%)
失敗原因	認識器選択誤り (第 1 パス)	10
	キーワード置換 (第 2 パス)	98
	キーワード欠落 (第 2 パス)	7
	キーワード剰余 (第 2 パス)	6
	語尾に助詞挿入 (第 2 パス)	1

実験の結果,「電源の場所が光るらしい」以外ではキーワードが 3 つ揃うことはなく，コマンド特定に結び付くことはなかったが,「電源の場所が光るらしい」では「電源の場所を光らせて」と誤認識し，コマンド対象として判断した．これは語尾の「らしい」が「らして」と認識誤りしたと考えられる．

6.3 考察

本研究では当初，最適なオブジェクト専用音声認識器の選択を目的に，キーワードスポッティングを使用した．しかし，オブジェクト専用音声認識器で認識した結果，キーワードが欠落・剰余であった発話のなかに，キーワードスポッティングの認識結果から欠落したキーワードが得られる例が 1 つ，また，剰余したキーワード (メソッド) のうちどちらが正解か判断できる例が 4 つ観測された．これらから，連続音声認識でのキーワード欠落や剰余の場合に，キーワードスポッティングの認識結果を利用できることが分かった．

7 まとめ

本稿ではユーザが計算機に対して音声で指示する場合の単純な構造の文音声を対象とした音声認識システムとして，キーワードスポッティングおよび連続音声認識を組み合せた音声認識手法を提案した．本手法を適用したタスクにおける評価実験を行った結果，単語正解率 88.8%，正解コマンド特定率 90.7%という高い精度の音声認識，コマンド特定が実現できた．

今後の課題としては，現インタフェースでの，オブジェクト・変数・メソッドからのコマンド特定を，オブジェクト・メソッドの組合せからのコマンド特定，また変数の数を増加させたときのコマンド特定も考慮し改良していく．

また，音声入力インタフェースは 2 つのパスで構成するため，各パスで用途に応じた認識手法の利用，切り替えが可能である．例えば，ユーザの発話のなかで強調した語彙に対し，第 1 パスで韻律情報に基づいたキーワードスポッティングを行い，その後，第 2 パスで強調した語彙が属するオブジェクト専用音声認識器を用いて認識を行うといった認識手法の切り替えも検討していく．

参考文献

[1] 屋野武秀, 笹島宗彦, 河野恭之：“文テンプレートハッシュ方式によるラティス解析方式と質問-応答システム”, 情報処理学会研究報告, 98-HI-80-13 98-SLP-23-13, pp. 71–76 (1998-10).

[2] 桐山伸也, 広瀬啓吉：“文献検索音声対話システムの機能拡張とその評価”, 情報処理学会研究報告, 2000-SLP-30-10, pp. 45–50 (2000-2).

[3] 櫻井光康, 河原達也, 堂下修司：“キーワードスポッティングに基づくニュース音声の話題同定”, 情報処理学会研究報告, 95-SLP-6-3, pp. 15–20 (1995-5).

[4] 今井亨, 小林彰夫, 尾上和穂, 安藤彰男：“ニュース番組自動字幕化のための音声認識システム”, 情報処理学会研究報告, 98-HI-80-11, 98-SLP-23-11, pp. 59–64 (1998-10).

[5] 官上大輔, 小川均：“操作説明のマルチモーダルプレゼンテーションシステムの構築”, 人工知能学会誌, Vol. 13, No. 2, pp. 221–230 (1998-3).

[6] 田中克明, 河原達也, 堂下修司：“汎用的な情報検索音声対話プラットフォーム”, 信学技報, NLC98-109,SP98-109, pp. 9–16 (1998-12).

[7] 官上大輔, 小川均：“教示生成におけるユーザの要求に関する研究”, 信学技報, NLC98-27, pp. 9–16 (1998-10).

[8] S.Young, J.Jansen, J.Odell, D.Ollason and P.Woodland: *The HTK BOOK*, Entropic Cambridge Research Laboratory (1995).

[9] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏：“日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価”, 情報処理学会研究報告, 2000-SLP-31, pp. 9–16 (2000).

[10] 山下洋一：“音声認識シミュレーションに基づくワードスポッティング精度 の予測”, 信学技報, SP98-95, pp. 65–70 (1998-11).