平成１２年度

卒業論文

音声認識による機器の制御
（音声認識における認識・誤認識についての一考察）

指導教員

井上喜雄　教授

甲斐義弘　助手

高知工科大学工学部

知能機械システム工学科

松谷　融

目次

図表

１章　緒言

近年，ロボットや機器類の発展にともない，操作や制御の手段が複雑になりつつある．しかし，一般に普及し万人が扱う機器類を操作，制御するには容易さが求められる．特に，介護機器や家電製品などのように不特定多数の人が使用するような機器においては，より使用者に負担の少ない容易な操作，制御が必要不可欠となる．

人は複雑な情報を伝達する手段として言葉を用いる．その中でも発声によって情報を伝達する手段をもっとも多用する．つまり人にとって音声による情報の伝達はもっとも負担の少ない伝達手段の一つと言える．また，キーボードやボタンなどによる操作に比べ熟練が必要なく，入力速度もそれらと比較しても数倍速い．体の他の器官を並列的に使用することも可能であるため，別の作業をしながらでも，機器の操作を行うことができる，などのメリットがある．

しかし，人同士であっても，聞き違いや聞き取りにくいというようなことがおこるように，人と機器の間でも同様のことが起こり得る．現在では音声入力によるテキスト入力や，携帯電話のボイスコントロール，カーナビゲーションシステムなどのように比較的危険の少ないシステムに使用されているが，機器の誤作動がおこった場合，機器が人に重大な危害を加える危険性のあるものには多用されていない．そこで，そのような機器を用いる場合は，音声認識の確実性が必要となる．現在音声認識の手法には大きく分けて，HMM(Hidden Markov Model)，ニューラルネットワーク，DP(Dynamic Programming)法の三通りがある．HMM はマルコフモデルを利用した数学的手法の一つであり，マルコフモデルのプログラムを上手に行えば学習能力も高くなり，確実性もあらわれてくるが，マルコフモデルの構造によっては学習能力が発揮されないなどプログラムに関してやや複雑さがある．ニューラルネットワークに関しても同様のことが言われている．

そこで，本研究では誤作動がおこった場合，人に重大な危害を加える可能性のあるような機器に他の手法に比べ比較的プログラムが容易な DP 法を使用した音声認識ボードが適応可能かどうかを検討する．また，万人が扱う機器の容易な制御を目的とし，実験の対象を不特定多数の人による制御とし，一つ一つの命令単語を入力するものとし，不特定話者，孤立単語音声認識方式の市販の音声認識ボードを用いる．

上述の音声認識ボードは，登録されていない音声に対しても音声の入力があればハード上で候補を選出するので，登録されていない単語も認識として扱われることがある．そこで，本ボードに付属の音声認識用サンプルプログラムをベースに，認識，誤認識の区別が可能なプログラムの作成を試みる．また，本ボードから与えられる距離値より認識，誤認識を区別する方法として，距離値にもとづいたしきい値を採用し，そこから認識，誤認識を選別する手段を用いる．それにより作成したプログラムを用いた認識実験を複数の被験者に適応し，しきい値，距離値による認識，誤認識について一考察を行う．

２章　音声認識プログラムの作成

２－１　音声認識ボード

有限会社スカラベ社の音声認識ボード AUDIO-98VOICE を使用した．このボードは音声認識の処理をおこなうために NEC 社製 MOS 集積回路μPD77524, サウンドコーデックに同社のμPD63310 を使用している．不特定話者対応の離散単語音声認識，認識方式には NEC 社開発の半音節音声認識方式を使用しており，DP 法によるパターンマッチング処理をおこなっている．

選定理由には,

・他の音声認識開発ソフトに対し比較的安価である．

・標準波形の作成が容易で，プログラムも比較的容易である．

・不特定多数の人間による機器の制御，孤立単語の認識により操作をおこなうことを目的としている．

という点があげられる．

またこのボードはμPD77524 より提供された距離値をもとに候補の選定をおこなう．本ボードに登録した単語の波形と，マイクなどから入力された音声の波形とを比較し，離れ具合を距離値という値で表している．この距離値は，登録された単語の波形と入力された音声を比較した際の一番近い候補との距離を示し，値が大きくなるにつれ，その候補とは異なることを示している．

２－２　音声認識プログラム

全体のプログラムの流れ

全体のプログラムの流れを図２－１に示す．

マイクレベル入力

入力時のマイクレベルの設定をおこなう．

マイクレベルはサウンドコーデック LSI を使用しておこなわれる．利得調整は６ビットでおこなわれ（図２－２参照），設定調整範囲は-46.5dB から 0dB，および-∞dB（ミュート）で，-1.5dB 刻みとなっている．プログラム中では 0dB（００）から-13.5dB（０９）までを使用しており，本研究では同社の推奨値である-7.5dB（０５）を使用している．

ゲイン入力

サウンドコーデックから音声認識 LSI に送る際の値の設定をおこなう．

マイクレベルの入力と同様の操作をおこなうこで設定する．

認識ファイルロード

読み込み用のファイルをロードする．

このファイルは音声入力時のキーワードとなる言葉が書かれている．データはテキスト形式のひらがなで書かれ，各単語の区切りはキャリッジリターンとしている．テキストはJIS コードを使用しており，音声認識 LSI が単語の読み方をダウンロードすることで認識単語の辞書（波形）を作成する．

表示用ファイルロード

表示用のファイルをロードする．

このファイルは本来，音声合成用ファイルであるが，本研究では音声認識を対象としたため，表示用ファイルとして使用した．認識ファイルの単語区切りに対応させることにより，入力した単語を表示させる．入力方法は認識ファイルと同様の操作をおこなう．

２－３　認識・誤認識判別プログラムの作成

２－１で述べた距離値によって認識もしくは，誤認識であるかを決定するため，音声認識ボードに付属の音声認識サンプルプログラムをベースに新プログラムを作成した．

新プログラムは図２－３に示すように

（１）プログラム開始前からしきい値を一定の値に決定しておき，それにより認識，誤認識を区別する．

（２）プログラムを開始する度にしきい値を決定，設定し，そのしきい値より認識，誤認識を決定する．

とした．

（１）のプログラムでは，あらかじめ決められたしきい値を入力しておくことでユーザーがその機器を使用する際にしきい値を検討する必要がなくなると同時に，ユーザーが新たに設定するという負担が少なくなる．

（２）のプログラムでは，細かなしきい値の設定が必要となるが，個人差を吸収することが可能となる．

また，両プログラムを作成する前に，共通の作業として，

・本研究で使用するプログラム環境（Borland 社 Turbo C for DOS）への変換を行った．
・ボードに登録する単語のファイルをユーザーが入力する手間を省くため，あらかじめプログラム内へ組み込みを行った．

本研究で作成したプログラムのソースについては省略する．

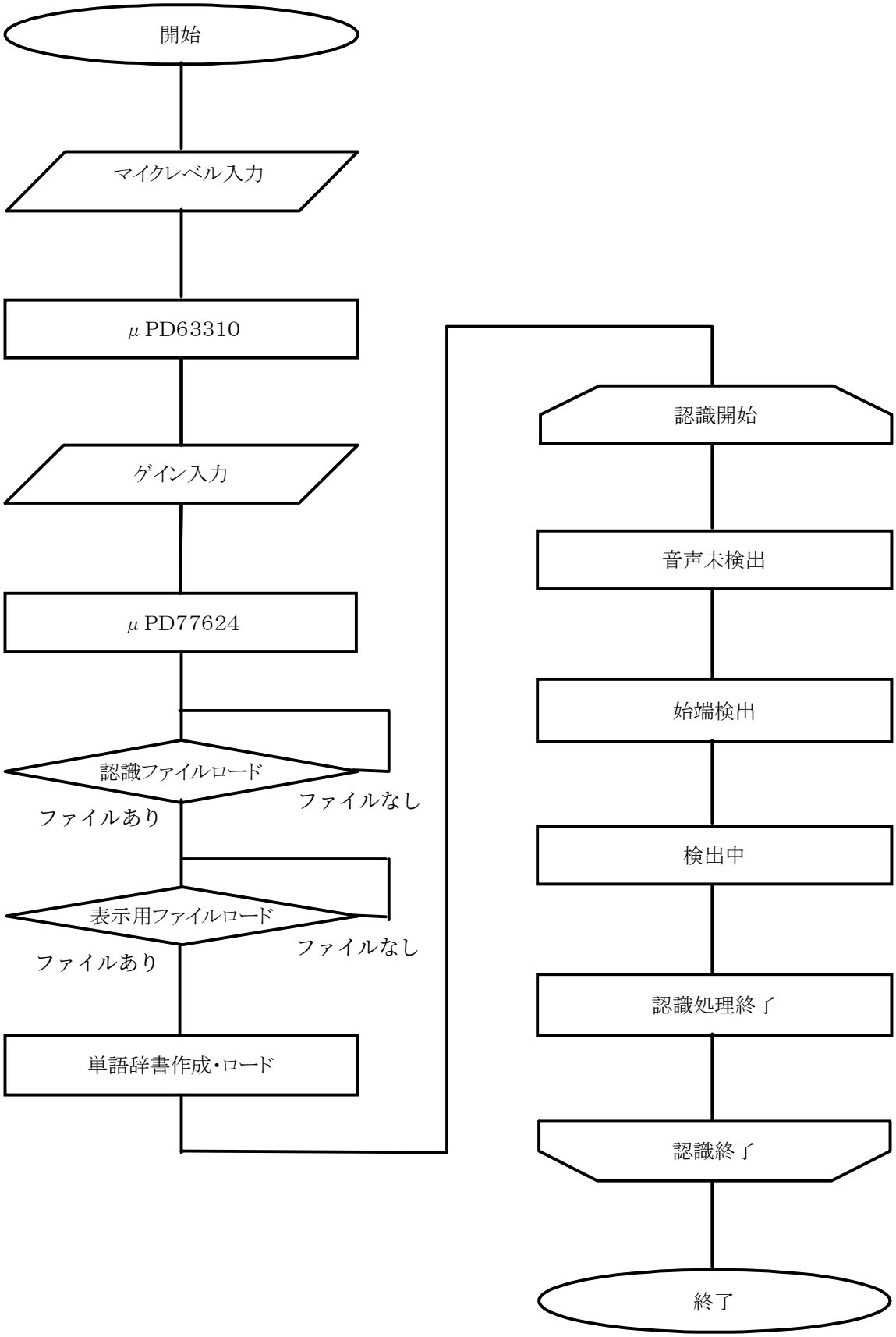

図２－１　プログラム全体の流れ

データ入力値	利得	データ入力値	利得
00	0.0	16	−24.0
01	−1.5	17	−25.5
02	−3.0	18	−27.0
03	−4.5	19	−28.5
04	−6.0	20	−30.0
05	−7.5	21	−31.5
06	−9.0	22	−33.0
07	−10.5	23	−34.5
08	−12.0	24	−36.0
09	−13.5	25	−37.5
10	−15.0	26	−39.0
11	−16.5	27	−40.5
12	−18.0	28	−42.0
13	−19.5	29	−43.5
14	−21.0	30	−45.0
15	−22.5	31	−46.5
		32	ミュート

図2－2　6ビットデータと利得値

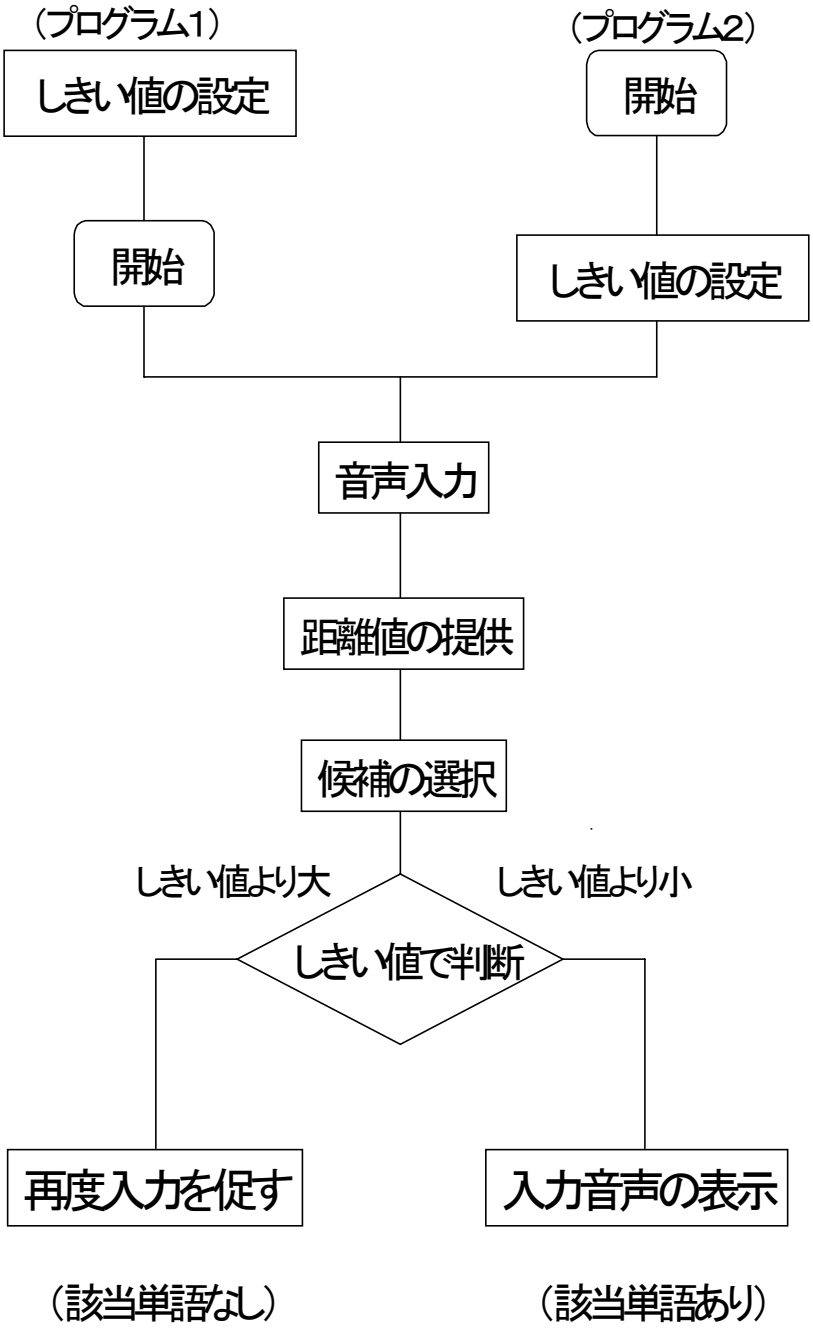

図2－3 プログラムの流れ

３章　実験

本研究では先に述べた距離値，しきい値，認識，誤認識の関係を調べるために

認識時の距離値・・・登録単語をマイクより入力した際の距離値

誤認識時の距離値・・・登録していない単語をマイクより入力した際の距離値

を記録し，それをもとにそれぞれをしきい値ごとに考察した.

３－１　実験装置

３－１－１　実験装置の構成

本研究では，実験を行うための装置を図３－１に示すような構成とした.

コンピュータに音声認識ボードを挿入し，それをプログラム（Borland 社 Turbo C for DOS）で作動させる．音声はマイクロフォンからおこない，音声認識ボードにより登録単語であるか否かを判断させ，その結果をディスプレイに表示させることにした.

３－１－２　コンピュータ

NEC 社製コンピュータ PC-9821Cs2S3 をベースに使用した. CPU のクロック数は High mode 33MHz に設定し，メモリは標準 5.6MB から 32MB 増設している．音声認識ボード以外のボードは取り付けていない.

選定理由には,

・本研究に使用した音声認識ボードが NEC 社製 PC-98 シリーズ対応であるため.

ということがあげられる.

３－１－３　マイクロフォン

アイワ社製ダイナミックマイクロフォン DM-H110 を使用した．指向特性は単一指向性，ｂ周波数帯域は 50～150000Hz，インピーダンスは 600Ω，感度-56dB となっている.

３－１－４　プログラム

距離値を記録するために，２－３で作成した新プログラムを使用した.

３－１－５　入力音声（発声単語）

音声認識ボードの認識方式に従い，音声入力時のキーワードとなる言葉の書かれているファイルを作成した．データはテキスト形式で一行につき一単語のキャリッジリターン方式となっている.

今回使用した「キーワード」は無作為に選出し，それぞれの発声単語は図３－２に示したものとする.

３－２　実験方法

３－２－１　被験者および実験環境

無作為に５人を選出し，登録した単語１９個，登録していない単語１９個，合計３８個を１０回ずつ発声してもらい，各回の最小値を記録した．無作為に選出した被験者は，２２歳男性A，２１歳男性B，２１歳女性C，２１歳男性D，２１歳男性Eの合計５人である．

また，サンプリングは比較的雑音のない部屋でおこない，単語入力一区切りごとにマイクロフォンのスイッチをＯＦＦにした．

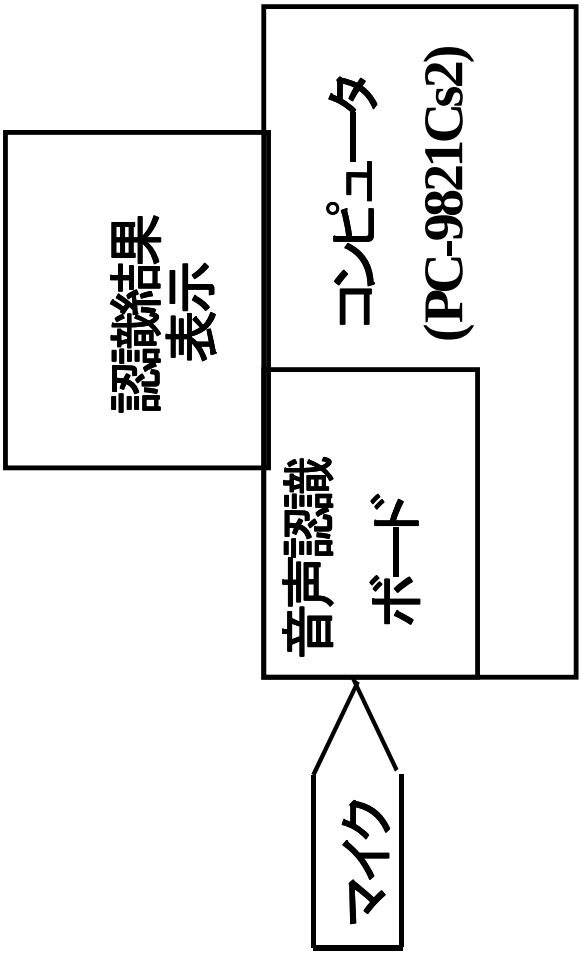

図3－1　装置の構成

大阪	おおさか		梅田	うめだ
福島	ふくしま		淀川	よどがわ
野田	のだ		姫島	ひめじま
西九条	にしくじょう		千船	ちぶね
弁天町	べんてんちょう		杭瀬	くいせ
大正	たいしょう		大物	だいもつ
芦原橋	あしはらばし		尼崎	あまがさき
今宮	いまみや		出屋敷	でやしき
新今宮	しんいまみや		尼崎センタープール前	あまがさきせんたーぷーるまえ
天王寺	てんのうじ		武庫川	むこがわ
寺田町	てらだちょう		鳴尾	なるお
桃谷	ももだに		甲子園	こうしえん
鶴橋	つるはし		久寿川	くすがわ
玉造	たまつくり		今津	いまず
森ノ宮	もりのみや		西宮東口	にしのみやひがしぐち
大阪城公園	おおさかじょうこうえん		西宮	にしのみや
京橋	きょうばし		香枦園	こうろえん
桜ノ宮	さくらのみや		打出	うちで
天満	てんま		芦屋	あしや

図３－２　発声音声
左）登録単語　右）未登録単語

４章　実験結果および考察

４－１　距離値からの評価

図４－１に被験者Ａが登録した単語を入力した際の距離値，図４－２に同被験者が登録していない単語を入力した際の距離値を示す．このグラフでは縦軸に各回の最小の距離値を，横軸に各発声単語の１回目から１０回目の距離値と各発声単語を示している．

これらの結果より登録単語に対し登録していない単語の方が，距離値が大きくなる傾向が見られた．

そこで，登録していない単語の最小距離値でしきい値を設定した場合，登録した単語のなかに認識することができない単語が現れる．ある被験者に対しての実験では，登録単語のうち，約４６％の単語が認識不可能とされた．そのなかには１０回の入力中１０回すべてにおいて認識できない単語もあった．

このことから，しきい値を低く設定した場合，本来認識するべき単語までが認識不可能となることがわかる．

また，すべての登録単語を認識可能にするため，認識率が１００％になるよう登録単語の最大距離値でしきいを敷いた場合，登録していない単語のなかにも認識として別の単語を表示する（誤認識を行う）単語もある．

これより，認識率を高めるためにしきい値を高く設定した場合，登録していない単語までを認識として最も近い単語を表示することがわかる．

他の被験者にも同様の結果が現れた．図４－３，５，７，９に，それぞれ被験者Ｂ，Ｃ，Ｄ，Ｅが登録した単語を入力した際の距離値，図４－４，６，８，１０にそれぞれの被験者準にに登録していない単語を入力した際の距離値を示す．

ここで被験者ごとに，入力単語によって距離値がばらつくという現象がでてくる．同じ単語を参照しても，個人個人で距離値に偏りがでてきている．これは個人ごとに同じ単語でも発声時間が異なると言うことを表していると思われる．

また，単語単位で見た場合，同一の人間であれば距離値が近い傾向がある．これは同一の人間が同じ単語を話す際は発声継続時間が近いということを表していると思われる．

４－２　しきい値ごとの評価

次に，しきい値ごとに認識，誤認識を分ける作業を行う．図４－１１に被験者Ａのしきい値ごとの認識率，誤認識率を表している．軸にしきい値を１００刻みに振り分け，縦軸にそのしきい値未満の全体の割合（パーセンテージ）を示している．この図から見られるように全体的に認識率が上昇するにつれ誤認識率も上昇する傾向にある．

他の被験者Ｂ，Ｃ，Ｄ，Ｅのしきい値ごとの認識率，誤認識率のグラフを図４－１２，１３，１４，１５に示す．

４－３　考察

以上の距離値，しきい値，認識，誤認識の結果より，各人，各音声に共通の完全な認識を行うしきい値は存在しない．そこで１章で述べた二つのプログラムについて考察を行う．

（１）のプログラムでは，ある程度の誤認識を許容される機器を制御する際においては，かなり不特定多数の人間にでも対応することが可能となる．しかし，さらに認識率を高める必要がある場合には，あらかじめ使用する単語の距離値を各人ごとに調べておき，それをもとに（２）のプログラムでしきい値を設定する必要がある．

また，しきい値を低く設定すると単語によってはまったく認識しない単語も現れる反面，しきい値を高く設定すると登録されていない単語が入力された場合，誤った認識を行うこともある．

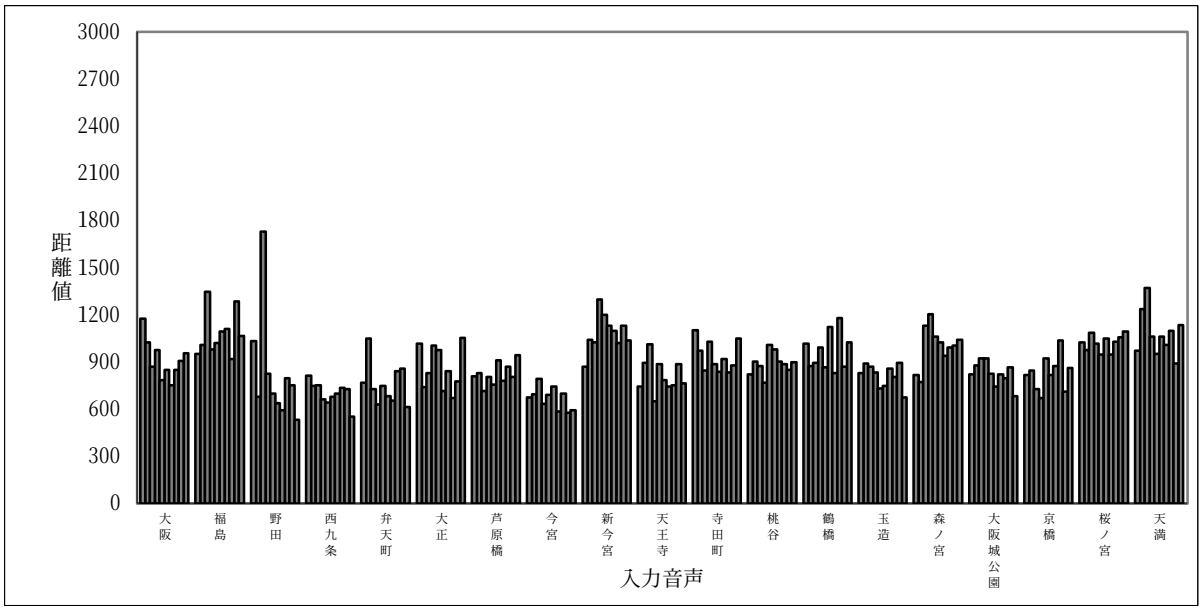

図４－１　登録単語と距離値（被験者Ａ）

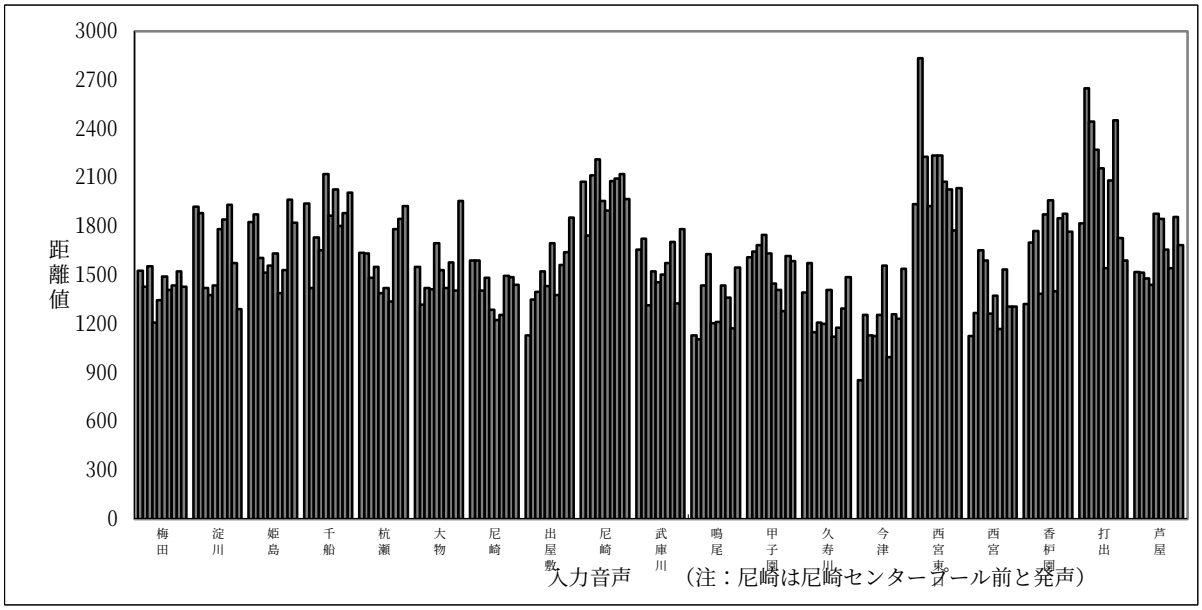

図４－２　未登録単語と距離値（被験者Ａ）

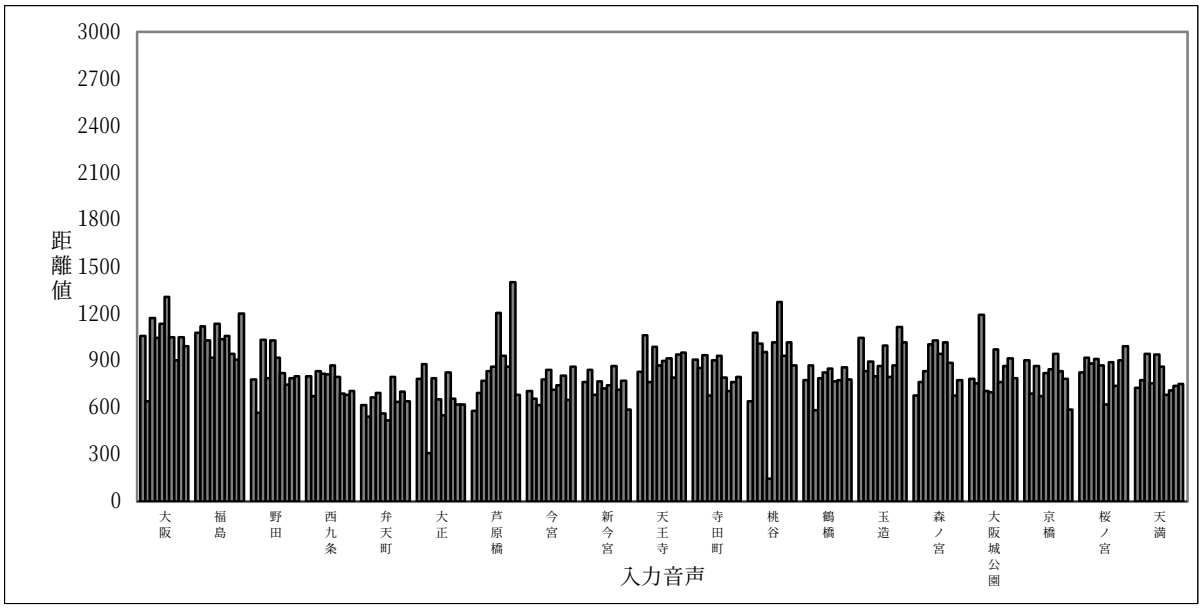

図４－３　登録単語と距離値（被験者Ｂ）

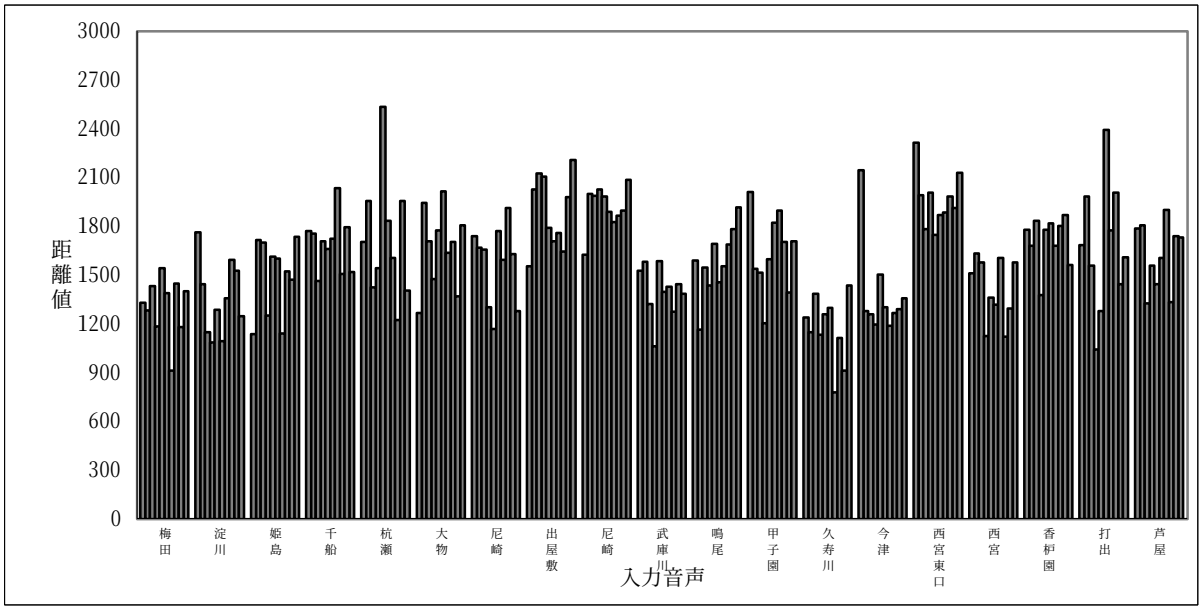

図４－４　未登録単語と距離値（被験者Ｂ）

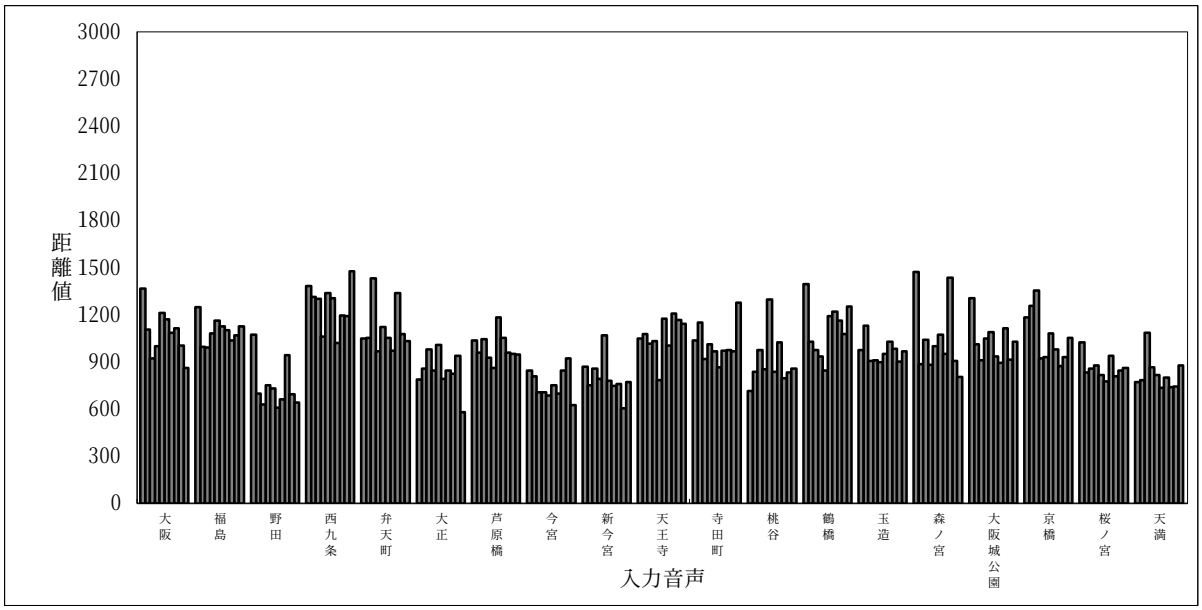

図４－５　登録単語と距離値（被験者Ｃ）

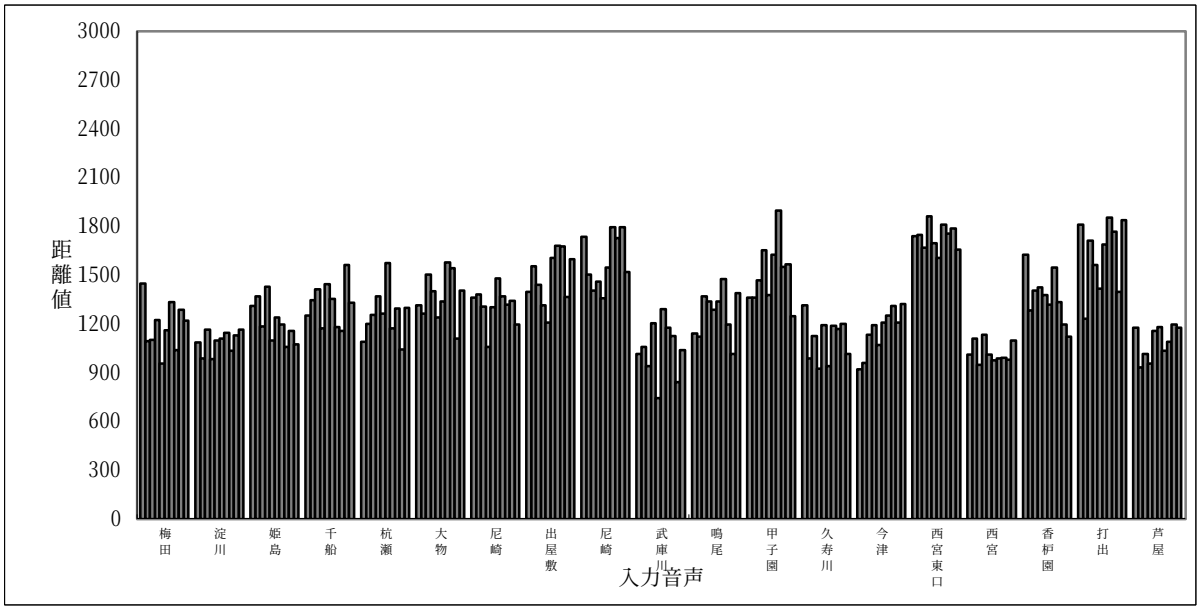

図４－６　未登録単語と距離値（被験者Ｃ）

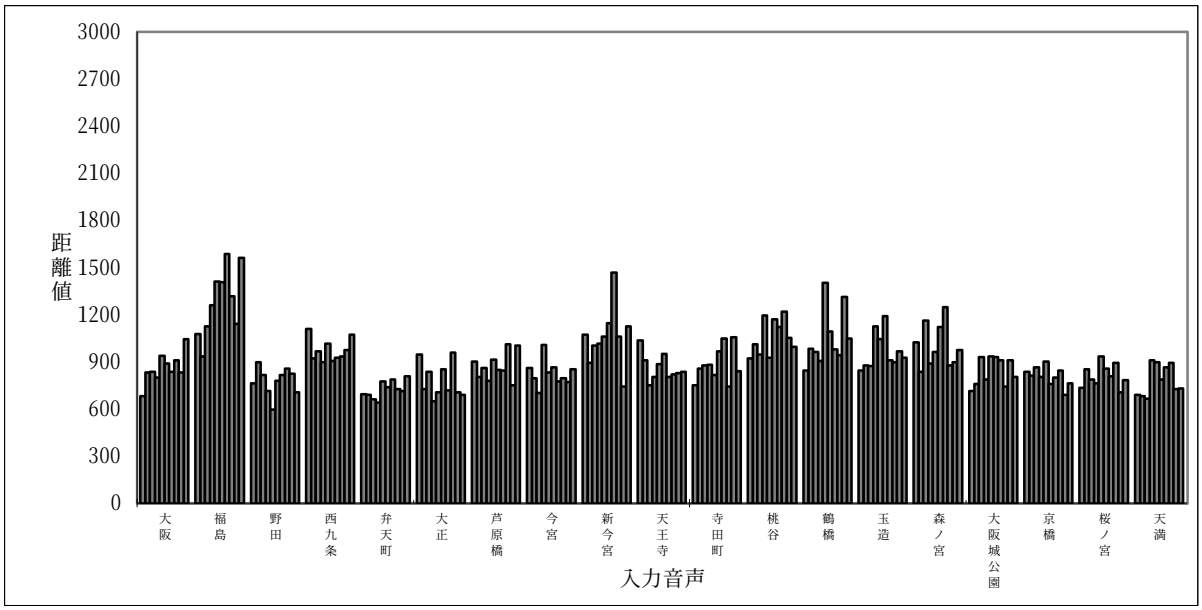

図４－７　登録単語と距離値（被験者Ｄ）

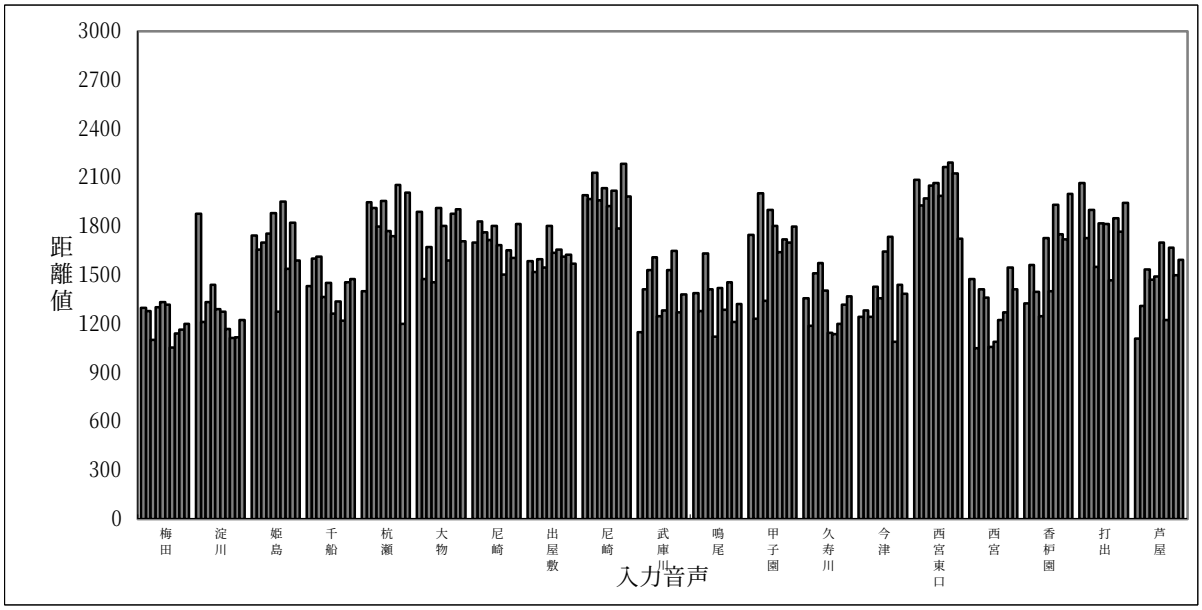

図４－８　未登録単語と距離値（被験者Ｄ）

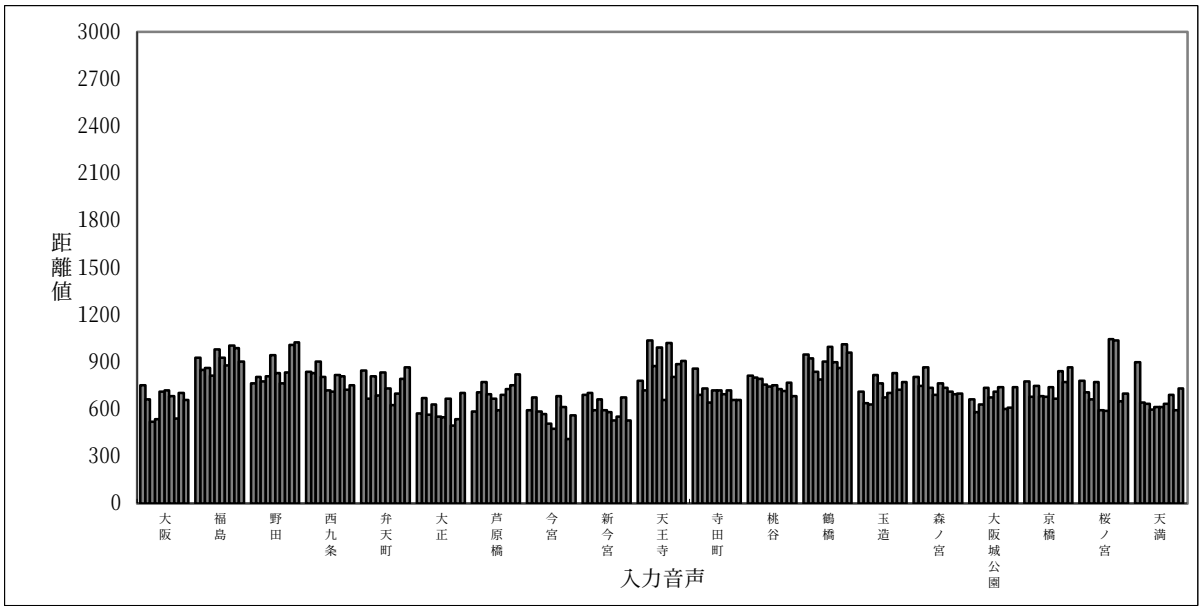

図４－９　登録単語と距離値（被験者Ｅ）

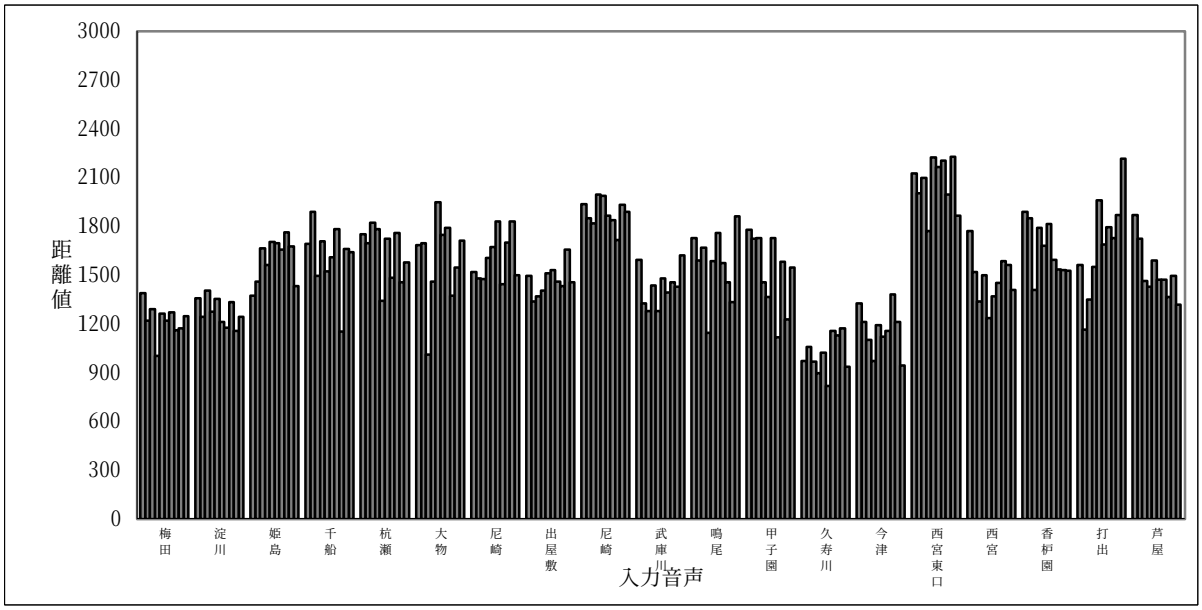

図４－１０　登録単語と距離値（被験者Ｅ）

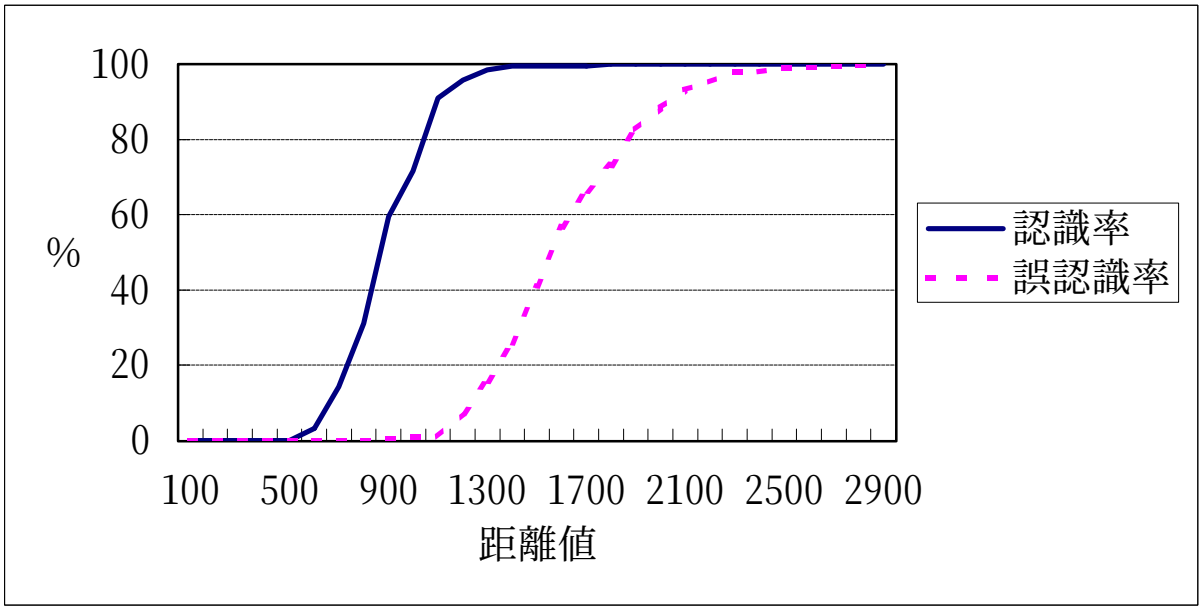

図４－１１　しきい値ごとの認識率・誤認識率（被験者Ａ）

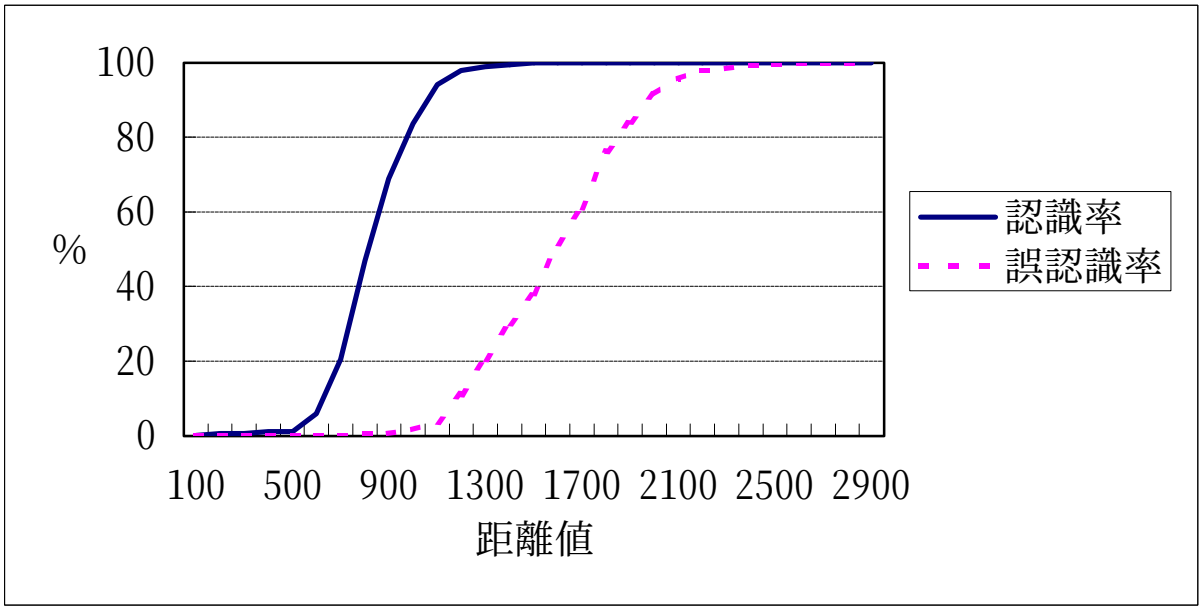

図４－１２　しきい値ごとの認識率・誤認識率（被験者Ｂ）

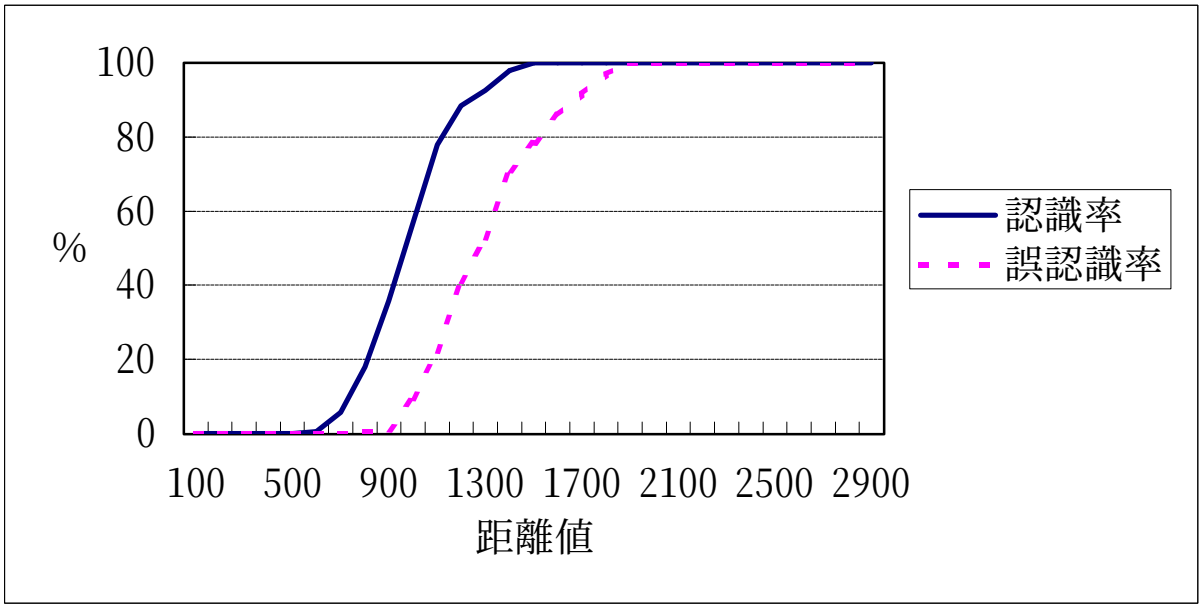

図４－１３　しきい値ごとの認識率・誤認識率（被験者Ｃ）

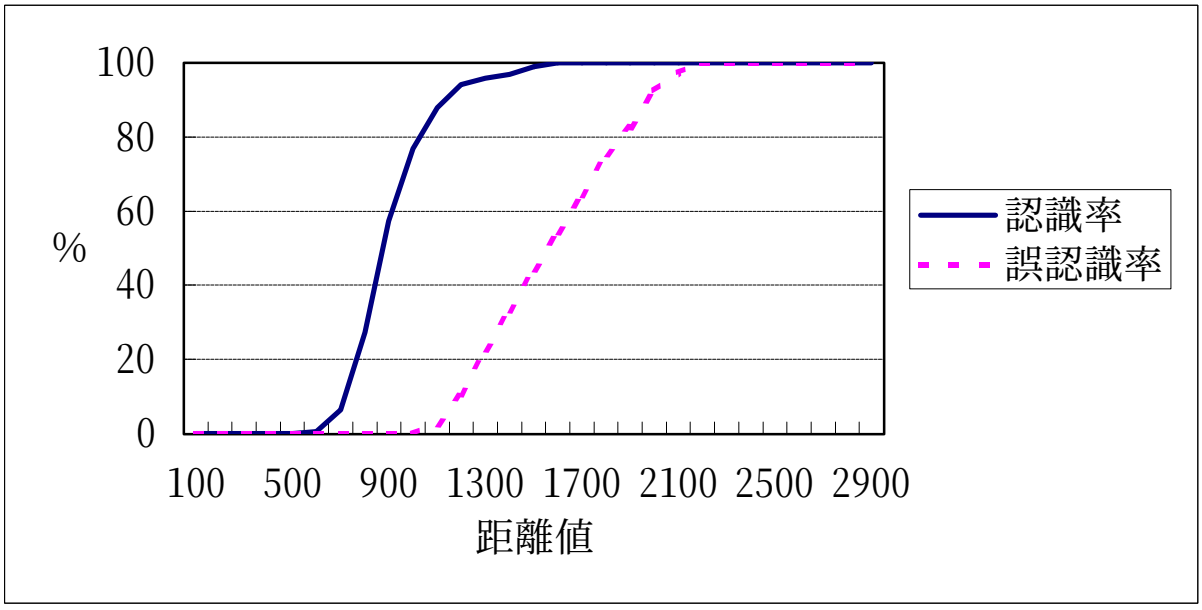

図４－１４　しきい値ごとの認識率・誤認識率（被験者Ｄ）

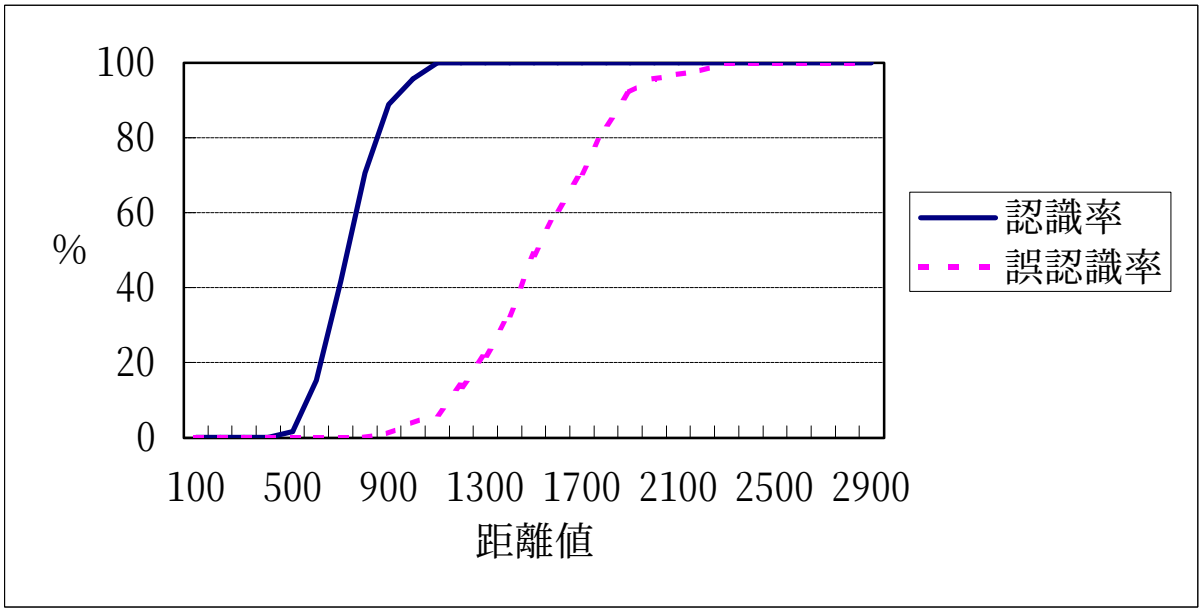

図４－１５　しきい値ごとの認識率・誤認識率（被験者Ｅ）

５章　結言

本研究では，パターンマッチング方式を用いた市販の音声認識ボードによる実験を行った．その結果から，認識，誤認識，しきい値について以下の結論を得た．

１．しきい値を低く設定することにより誤認識率は減少するが，認識率も同様に減少する．また，登録単語においては何度入力しても認識することができない単語が現れる．そこでそのような単語に関しては同義で，異音の単語を選択，登録する必要がある．

２．しきい値を高く設定することで認識率は上昇する．反面，誤認識率も上昇する傾向がある．そこで使用する機器に応じてしきい値の設定が必要となってくる．

３．特に動作に注意をはらう必要のある機器を扱う際は，さらに確実な認識が必要となる．そこで使う人間がある程度限定されている場合には，その人の特徴となる距離値を前実験で求めておくことが求められる．さらに，正しければボタンを押すなどの二次的な補助装置を設けるなどの対応が求められる．

謝辞

本研究を行うにあたり，本学工学部井上喜雄教授，甲斐義弘助手には絶えず御指導を賜りましたことを深く感謝し，厚く御礼申し上げます．また，数多くの助言，叱咤激励をくれた本学井上甲斐研究室の４回生のメンバー，その他本研究に協力していただいた皆様に，ここに感謝の意を表します．

参考文献

今井　聖　音声認識，共立出版，１９９５

北脇信彦　音のコミュニケーション工学，コロナ社，１９９６

古井貞熙　音響・音声工学，近代科学社，１９９２