# 音声誤認識データベースによる認識率向上とその効率的作成方法

三谷 健† 葛谷 紳† 渡部 広一† 河岡 司†

†同志社大学工学部知識工学科

## 1. はじめに

人間が知的生活を送るためにはコミュニケーション環境が必要であり，常に無数の情報を様々な手段によって収集している．その中で，人間と人間との対話は最も頻繁に遭遇するコミュニケーション環境である．一方，機械と人間との対話も多く存在する．街角や博物館などで見かける，ボタンを押すと対応した場所が光る電光掲示板や，切符などの自動販売機もその一種の対話システムといえる．このような対話システムでは， コンピュータが扱える処理が"あらかじめ決められた情報"のみで，それ以上の情報が提示されることはない．そのため，このようなシステムに対し，人間側はコンピュータに合わせることで対応している．

近年のロボットは，ソニーのエンターテインメントロボット「AIBO」や本田技研工業の HONDA ヒューマノイドロボット「ASIMO」のように，「ただの機械」から「親しみのあるロボット」へと発展しつつある．また，実用的な秘書ロボットや介護ロボットなど，人間とのコミュニケーションが可能な知能ロボットが注目されている．このような知能ロボットを実現するためには，知能ロボットが人間の発している音声を認識し，その発話内容を理解し，意図していることを常識的に判断する必要がある．

音声認識の分野では，これまで様々な研究が行われ，さまざまな製品が提供されている．これまでの音声認識ソフトのほとんどは，音声認識率の向上を主たる目的として開発が進められてきた．近年，ロボットとコミュニケーションをはかる手段の一つとして，対話形式による音声認識の実現が期待されている．現在行われている対話形式による音声認識の分野では，状況を限定したり，語彙数や文法に制限をかけたりした上で，音声認識を行っている．そのため，多種多様な表現には対応出来ない．

本稿では，不特定多数話者と日常活動型知能ロボット ROBOVIE[1]とのコミュニケーションを想定し，ROBOVIE で動作・移動可能な言葉，および挨拶などの使用頻度の高い語彙の単語に対し，あらかじめ誤認識データベースを作成しておくことにより，認識率の向上を図る一致度補正方式を提案する．その一方で，それ以外の語彙（使用頻度の低い語彙）についても，市販の音声認識ソフトの認識結果を利用し，限定語彙に限らず認識が可能なシステムの構築を提案する（図１）．また，この方式のキーとなる誤認識データベースの効率的な作成方法を提案する．

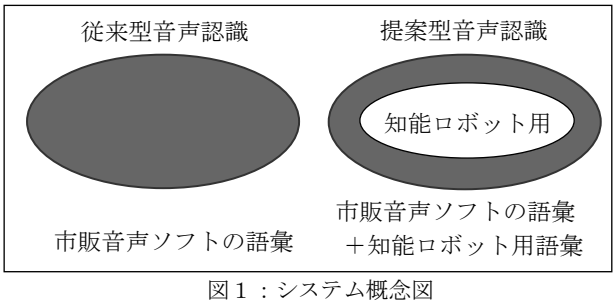


図１：システム概念図

## 2. 実験環境

### 2.1. 知能ロボットの構成

コミュニケーションを行う上で重要な上半身によるジェスチャー表現に対応するために，本研究で使用する知能ロボットROBOVIEは，人間と同程度の表現が可能な肩関節３自由度・肘関節１自由度の２リンク４関節腕マニピュレーターを両腕に備えている．また，コミュニケーションをはかる上ではさほど重要とされない脚部は，移動機能のみを可能とする２車輪および１車輪の補助輪を備えた台車（Pioneer2[2]）が搭載されている（図２）．

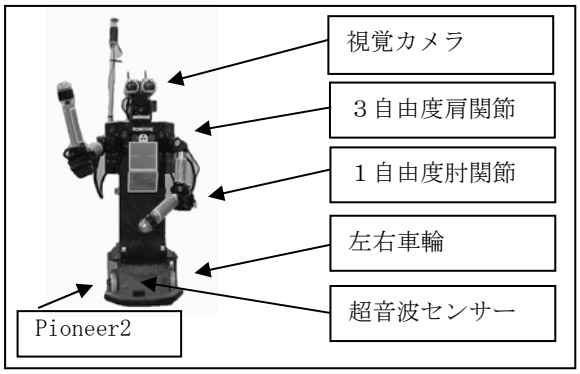


図２：知能ロボットの構成

### 2.2. 音声認識ソフト

#### 2.2.1. 音声認識ソフトについて

現在市販されている音声認識ソフトは，大別して不特定多数話者による音声認識，特定話者による音声認識，さらには単語を対象とした単語認識，および，連続音声を対象とした連続音声認識の４つに分類される．

本稿では特定話者の連続音声認識ソフトウェア D を用いて実験を行った．

#### 2.2.2. ソフトウェア D とその利用法について

ソフトウェア D は使用する前にユーザごとにエンロールを行う．「エンロール」とは，予め用意された文章を読み，そのユーザの声の特徴をコンピュータに学習させ，その人のエンロールデータを作成することである．そして，その自分のエンロールデータを用いることで特定話者の音声認識を行うことができる．

しかし，本研究の目的は特定話者かつ不特定多数話者に対する音声認識率の向上である．よって，本稿では他人のエンロールデータを用いることで擬似的な不特定多数話者音声認識を行う手法について述べる．

### 2.3. 入力装置

一般的な音声認識ソフトで奨励されている音声入力方法は，有線ハンドマイク・ヘッドセットマイクによるものが多く，本稿で使用するソフトウェア D においても，有

線ヘッドセットマイクによる入力が想定されている．

しかし，ROBOVIE が動作・移動を行いながらコミュニケーションを行う上で，有線による音声入力がその妨げとなる可能性がある．また，ユーザビリティの観点からみても，ヘッドセットの着用や，ピンマイクのセットなど，その利用方法は煩雑であると言わざるを得ない．そこで，本研究では無線ハンドマイクを利用し，音声認識専用のパソコンで入力音声を受信する方式を取ることとした．また，受信機を増やすことで，ほぼ同程度の入力音声を複数の音声認識専用端末で受信可能となる．受け取った入力波形を市販の音声認識ソフトで解析し，認識結果をネットワークを介してまとめ，１台の計算機で補正を行う．そして，最終的な結果を ROBOVIE が動作可能な命令に変換して，ROBOVIE へ送信する．本研究では３台の計算機と受信機を用意し，音声認識実験を行った（図３）．

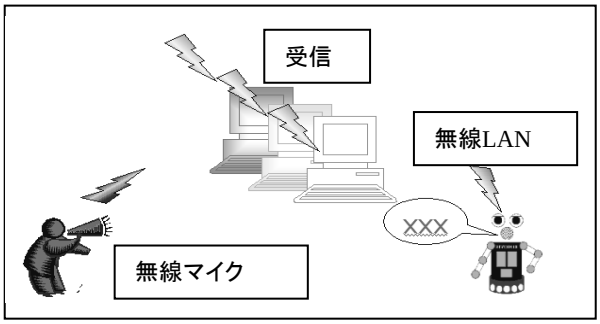


図３：実験環境

## 3. 誤認識データベースによる各種補正法

入力語Aに対して音声認識ソフトが返してきた認識語$a_i$とその出現回数$w_i$の対の集合として定義する．

$$A=\{(a_1, w_1), (a_2, w_2), \cdots, (a_i, w_i)\}$$

ここで，$a_i$を「属性」と呼ぶ．また便宜上，Aを「見出し語」と呼ぶ．このような属性が定義された見出し語を大量に集めたものを誤認識データベース（表１）と呼ぶ．

表１：誤認識データベースの例

| 見出し語 | 属性 1,重み | 属性 2,重み | 属性 3,重み | … |
|---|---|---|---|---|
| する | する，27 | つーる，8 | すぐ，7 | … |
| ふる | する，13 | ふうふ，6 | ふる，4 | … |
| くむ | くむ，14 | する，9 | すぐ，2 | … |
| すすむ | すすむ，34 | すすめ，5 | する，1 | … |

次に，音声認識ソフトが返した認識語を誤認識データベースで参照し，一致した属性があれば，その見出し語を最終的な結果として出力する．もし，誤認識データベースに複数個一致する属性が含まれていれば，それらの見出し語を全て補正対象とし，その属性の出現回数を比較し，最も出現回数の多い見出し語を最終結果として出力する．以下これを「単純補正方式」と呼ぶ．

また，３台の計算機から返される認識語{p1, p2, p3}と，任意の見出し語の属性集合との比較を行い，３つの認識語が属性集合の中に含まれていたら，それぞれの属性の出現回数の総和をその見出し語の一致度とし，全ての見出し語に対して，一致度を計算して一致度が最大となる見出し語を最終的な結果として出力する．以下，これを「一致度補正方式」と呼ぶ．

## 4. 誤認識データベースの構築手法

本研究の目的は，認識率向上と誤認識データベースを効率的により良いものにすることである．これは，ただ認識結果を適当に誤認識データベースに登録していては，無駄な入力や手間がかかるからである．これでは今後，誤認識データベースに登録する語彙数を増やすことが困難になる．よって，誤認識データベースを効率的により良く構築をする方法が重要である．具体的には，どのような条件で誤認識データベースを作成するかが要点になる．

そして，誤認識データベースを作成する際の要点は大きく分けて以下の三つが挙げられる．

１）誤認識データベース作成の際に利用するエンロール数の比較（4.1 節）
２）各単語における認識結果の解析（4.2 節）
３）各単語における必要な作成人数と入力回数（4.3 節）

本稿では，この三つの要点において解析を行う．

### 4.1. 誤認識データベース作成の際に利用するエンロール数の比較

#### 4.1.1. 実験目的

本節では，誤認識データベース作成の際に利用するエンロール数によっての正解率の差を検証する．この実験は，誤認識データベースを「１人分のエンロールデータ」で作成する場合と，「複数人分のエンロールデータ」で作成する場合を比較する．この実験により，誤認識データベースを作成する際に必要なエンロールの数がわかる．

#### 4.1.2. 実験方法

まず，誤認識データベースを作成した．誤認識データベースは，ROBOVIE との会話で頻繁に使用される１００単語を被験者１０名が入力した．エンロールの数は，①～③のように，１人分，２人分，３人分と３つの場合に分けて作成した．各単語における合計入力回数（１エンロール×６回入力×１０名＝６０回）を統一するために，①では６回入力，②では３回入力，③では２回入力とする．

①誤認識 DB１：１人分{α}のエンロール×６回入力
②誤認識 DB２：２人分{α,β}のエンロール×３回入力
③誤認識 DB３：３人分{α,β,γ}のエンロール×２回入力

次に，①～③の誤認識データベースを用いて補正を行うために，テストデータを用意する．テストデータは，３名の被験者{α,β,γ}が事前にエンロールを行った計算機で，被験者１０名が同様の１００単語を１回ずつ入力した．

そして，このテストデータを①～③の誤認識データベースをそれぞれ用いて，単純補正方式と一致度補正方式で補正を行った．結果は，4.1.3 節に示す．

不特定多数話者に対する音声認識が本研究の目的であるため，誤認識データベース作成の被験者１０名，エンロール行った３名{α,β,γ}，テストデータ作成の１０名は，それぞれ別人である．

#### 4.1.3. 実験結果

実験結果を示す（図４）．横軸が用いたテストデータのエンロールと補正方法を表し，縦軸が補正後の正解率を表す．黒色のグラフが①の DB１，灰色のグラフが②の DB２，白色のグラフが③の DB３である．

これより，単純補正方式の場合では，テストデータがα，β,γエンロールの３つの場合で，全て DB１と DB２の正解率に明らかな差（４～６％）が見られる．

これに対し，一致度補正方式の場合では，３つの DB において正解率に明らかな差が見られなかった．

これは，単純補正方式の場合では，テストデータ自体が１人分のエンロールであるため，誤認識データベースを作成する際のエンロールが，テストデータのエンロールと声質が異なると，正解率が下がるためだと考えられる．一致度補正方式の場合では，テストデータ自体が３人分のエンロールであるため，単純補正方式の様に誤認識データベースに異なるエンロールを用いることにはならず，正解率に差がなかったと考えられる．

以上のことから，誤認識データベースを作成する際に，「単純補正方式」ではエンロール数は複数必要であり，「一致度補正方式」ではエンロール数は正解率に影響しないと考えられる．

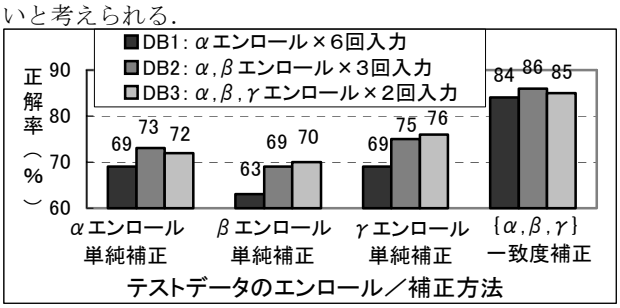


図４：エンロール数による比較

## 4.2. 各単語における認識結果の解析

### 4.2.1. 実験目的

本節では，誤認識データベースの効率的な構築をするために，「各単語における認識結果の特徴と収束」を検証する．収束とは，各単語を何度も入力しても，重み２以上の認識結果の属性の種類が増加しないこととする．

### 4.2.2. 実験方法

実験方法は，ROBOVIE との会話で頻繁に使用される１００単語において，３名の被験者{α,β,γ}が事前にエンロールを行った計算機で，各２００回の入力を行った．そして，各単語の認識結果の解析を行った．

### 4.2.3. 実験結果

この実験により，「音の長さによる収束の早さの違い」がわかった（図５）．

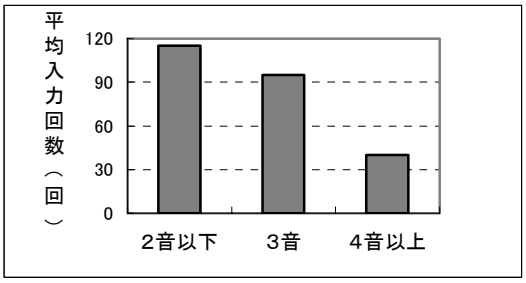


図５：単語の音の長さによる収束の早さの違い

これは，１００単語を，音の長さに分けて解析を行った結果である．音の長さは，２音以下の単語(右，押す，手 etc)，３音の単語(止まる，進む etc)，４音以上の単語(後退，直進 etc)の３つのグループに分けた．

この図５では，横軸は音の長さ別の単語で，縦軸はそれぞれの単語が収束するまでの平均入力回数である．この図からもわかるように，「２音以下の単語」では収束するまでに平均で約１１０回，「３音の単語」では約９０回，「４音以上の単語」では約４０回の入力が必要と言える．

これは，単語の音の長さが短い場合，音声認識ソフトは音声解析が正しく行えず，誤った認識結果を返すと考えられる．それに対し，単語の音の長さが長い場合，音声認識ソフトは音声解析を正しく行いやすく，認識結果にばらつきが少ないと考えられる．

## 4.3. 各単語における必要な作成人数と入力回数

### 4.3.1. 実験目的

この実験目的は，誤認識データベースの効率的な構築である．具体的には，効率良く誤認識データベースを作成する際の「各単語における必要な作成人数と入力回数」を検証することにより，無駄な入力を減らすことである．

### 4.3.2. 実験方法

まず，ROBOVIE との会話で頻繁に使用される１００単語を用いて，以下の①～③の誤認識データベースを作成した．「エンロール」は，３名の被験者{α,β,γ}が事前にエンロールを行った計算機を用いた．

そして，各単語における「入力回数」の検証を行うために，各単語を①では１回入力，②では３回入力，③では５回入力を行い，それぞれの誤認識データベースの入力回数を増加させた．

また，各単語における「作成人数」の検証も行うために，誤認識データベース①～③でそれぞれ被験者を１～１５名に増加させた．よって，合計４５個（３×１５人）の誤認識データベースを作成した．

①誤認識 DB1:3 人分のエンロール×１回入力(1～15 人分)
②誤認識 DB2:3 人分のエンロール×３回入力(1～15 人分)
③誤認識 DB3:3 人分のエンロール×５回入力(1～15 人分)

次に，①～③の誤認識データベースを用いて補正を行うために，テストデータを用意する．テストデータは，３名の被験者{α,β,γ}が事前にエンロールを行った計算機で，被験者１０名が同様の１００単語を１回ずつ入力した．

そして，このテストデータを①～③の誤認識データベースをそれぞれ用いて，単純補正方式と一致度補正方式で補正を行った．結果は，4.3.3 節に示す．

不特定多数話者に対する音声認識が本研究の目的であるため，誤認識データベース作成の被験者１５名，エンロール行った３名{α,β,γ}，テストデータ作成の１０名は，それぞれ別人である．

### 4.3.3. 実験結果

まず，単純補正方式の結果を示す（図６）．

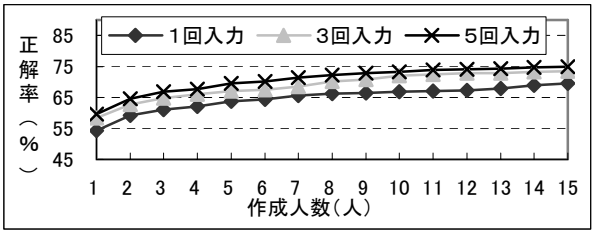


図６：単純補正方式の結果

横軸が誤認識データベースの作成人数を表し，縦軸が補正後の正解率を表す．３本のグラフがそれぞれ，①～③の誤認識データベースである．

図６において，正解率が約７４％で変化がほとんど無くなっている．よって，誤認識データベースを作成する際は，単純補正方式では（入力回数，作成人数）＝（3, 15），（4, 12），（5, 10）となるようにし，各単語において合計約 50 回の入力が必要と言える．

次に，一致度補正方式での結果を示す（図７）．

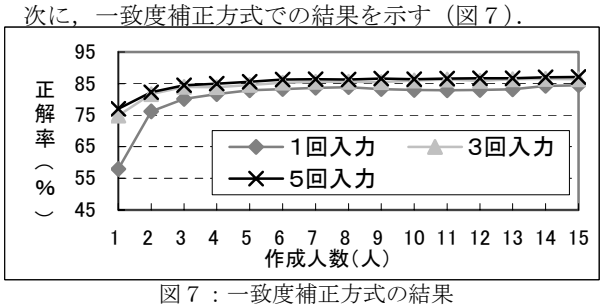


図７：一致度補正方式の結果

図７において，正解率が約８７％で変化がほとんど無くなっている．よって，誤認識データベースを作成する際は，一致度補正方式では（入力回数，作成人数）＝（3, 8），（4, 6），（5, 5）となるようにし，各単語において合計約２５回の入力が必要と言える．

この実験を 4.2.3 節で求めた音の長さ別で行ったところ，図８のようになった．各単語における入力回数は３回で実験を行った．この図より，「２音以下の単語」では約３０回（３回×１０人），「３音の単語」では約２０回（３回×７人），「４音以上の単語」では約１０回（３回×３人）の合計入力回数が必要と言える．よって，さらに合計入力回数を減らすことができた．

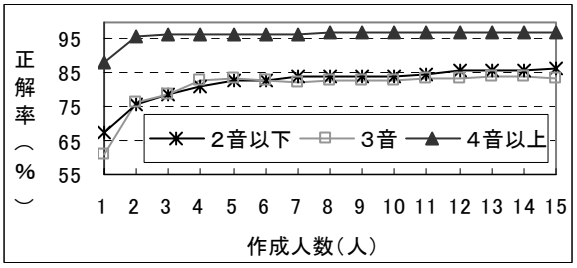


図８：一致度補正方式（単語の音の長さ別）

最後に，補正前，単純補正方式，一致度補正方式での正解率の比較を示す（表２）． これより，単純補正方式で約２７％，一致度補正方式で約４０％正解率が上がっており，一致補正方式が有効な補正手法であることが実証された．

表２：補正前，単純補正方式，一致度補正方式の比較

| | 正解率（%） |
|---|---|
| 補正前 | 約４７ |
| 単純補正方式 | 約７４ |
| 一致度補正方式 | 約８７ |

## 5. まとめ

4.1.3 節より，誤認識データベースを作成する際に必要なエンロール数は複数である方がよいと言える．また，4.3.3 節の表２より，補正方法は一致度補正方式が有効であると考えられる．

そして，各単語における必要な合計入力回数を検証するために，4.2.3 節で示した音の長さ別での「認識結果の収束」を利用する手法（図５）と「一致度補正方式」を利用する手法（図８）を比較する（表３）．

表３：誤認識データベースに必要な合計入力回数の削減

| 手法 / 単語 | 認識結果の収束 | 一致度補正方式 |
|---|---|---|
| ２音以下 | 約１１０ → | 約３０ |
| ３音 | 約９０ → | 約２０ |
| ４音以上 | 約４０ → | 約１０ |

この表より，「２音以下の単語」では約１１０回から約３０回，「３音の単語」では約９０回から約２０回，「４音以上の単語」では約４０回から約１０回，それぞれ約８０回，約７０回，約３０回の合計入力回数を減らすことができると言える．

以上で，本研究目的である「誤認識データベースの効率的な構築手法」を示した．今後，誤認識データベースの語彙数を増やす時に，この実験結果をもとに誤認識データベースを作成する．

## 6. おわりに

ロボットと人間が自然にコミュニケーションを行う際に，最も大きな障害として考えられるのは，人間が経験によって蓄積していく「知識」である．

本稿で扱っている誤認識データベースは，音声認識ソフトの認識語を元に作成されたデータを知識として蓄積し，音声認識ソフトで認識しきれなかった入力語を補正することで認識率の向上を図った．

しかし，この誤認識データベースにおける知識の蓄積には多大な時間と労力がかかる．そこで，効率的で無駄の少ない誤認識データベースの構築手法を提案し，その有効性を示した．

また，単純に誤りデータを補正する単純補正方式だけでなく，誤りの近さを表す一致度により，複数の計算機を用いて，入力語を推測し補正を行う一致度補正方式を提案し，その効果を示した．

今後は誤認識データベースの語彙数の拡張や，文章入力による ROBOVIE 制御法，ROBOVIE との対話システム構築の検討が必要である．

## 謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト「知能情報科学とその応用」における研究の一環として行った．

## 参考文献

[1] ATR 知能映像通信研究所，“日常活動型ロボット Robovie,” http://www.mic.atr.co.jp/~michita/everyday/

[2] ActivMedia Robotics LLC，“Pioneer robots from ActivMedia Robotics,” http://www.activrobots.com

[3] 葛谷紳，渡部広一，河岡司，“誤認識データベースを用いた単語音声認識方式,” 信学技報，NLC2004-11 pp.1-6，Nov．2004.