単語音声認識のための音声合成装置を用いた誤認識データベースの自動構築

空野 皇司　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

近年 SONY 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」，トヨタ自動車の「パートナーロボット」のように，ロボットは「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．人間とコミュニケーションできるロボットが実現できれば，業務，医療，福祉介護，交通案内，娯楽など多くの分野で活躍できると考えられる．人間との円滑なコミュニケーションには音声による会話が望ましい．そのため，人間の言葉を正確に聞き取る音声認識の技術が必要となる．その際，ロボットが日常的に人間に接するには，特定の話者だけでなく，あらゆる人間の声を認識できるようにし，実用性を高める必要がある．

既存の音声認識装置では，不特定多数の話者に対しての音声認識率が低く，また環境（マイク，雑音など）によって音声認識率が変動する．その精度向上のための手法として，誤認識データベースを用いた一致度補正方式の研究[1]がある．誤認識データベースとは入力語に対して，音声認識の誤りパターンを蓄積したものであり，その誤りパターンに対応した補正を行うことによって認識率の向上ができる．この手法は，音声認識装置から得られた結果を補正するため，音声認識装置に依存しない補正システムである．しかし，この誤認識データベースの構築には多くの人手と時間を要し，より良いデータベースを構築するには多くの被験者に音声実験の協力をしてもらう必要がある．また，将来的な活用を見越して女性，子供，老人など一般社会に対応した多様な被験者を無作為に集めるのが望ましいが，実際には困難である．

そこで本稿では，このような問題点を軽減するため，市販されている音声合成装置を用いて自動的に誤認識データベースを構築し，それを改良することによってより高い認識率をもつ誤認識データベースの構築を目指す．

音声合成を用いて誤認識データベースを構築することによって次のようなメリットが挙げられる．

- 被験者に掛かる多大な労力と時間が省ける．
- 音声合成のパラメータ（音程，抑揚など）を操作することで多様な声質を再現できる．
- 自動的に大規模な誤認識データベースを構築できる．

2. 誤認識データベース

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の対の集合を求める．

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

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	…
向く	ぬく,17.0	むく,11.8	ぼく,8.1	…
する	する,62.2	せる,11.8	するん,2.9	…
右	みぎ,25.1	に,7.4	ふたり,7.4	…
腕	うで,45.5	むれ,3.7	うえで,3.7	…

例えば，表１の誤認識データベースの例では，見出し語の「腕」に対しては，事前実験により，認識装置が過去に45.5%の割合で「うで」と認識語を返してきたことを示す．同様に「むれ」は 3.7%，「うえで」は 3.7%と続く．

音声認識装置が返す認識語は読みではなく，漢字や片仮名に変換されたものが多い．そのため，読みが同じでも表記が異なる場合がある．そこで，誤認識データベースに登録する際に，漢字や片仮名，特殊な記号は全て読みに変換してから登録を行った．

また，本稿における誤認識データベースは，会話で頻繁に使用される２００単語を対象として実験を行い構築した．

3. 一致度補正方式

一致度補正方式とは，３つの音声認識装置が返してきた認識結果を一セットとし，誤認識データベースを参照して補正を行う手法である．具体例を以下にあげて説明する（図 1）．

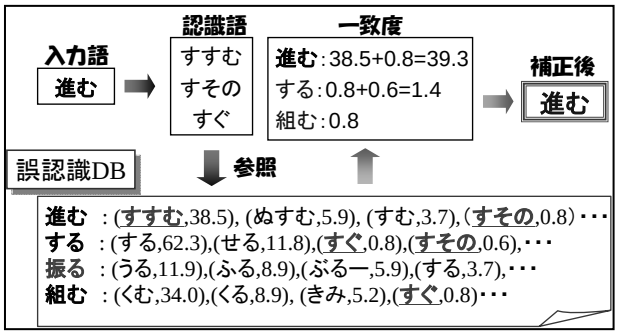

図１：一致度補正方式の例

図１の例では，発話者が「進む」と入力したのに対して，３台の音声認識装置が「すすむ」，「すその」，「すぐ」という認識結果を返している．ここで，誤認識データベースの属性中の「すすむ」，「すその」，「すぐ」を全て参照する．ここでは，見出し語「進む」，「する」，「組む」の中に存在する．認識候補が３つ出てきたため，見出し語の属性に付与されている頻度の総和を比較し，最も高い見出し語を最終的な認識結果とする．この例では，「進む」，「する」，「組む」の属性の頻度の総和がそれぞれ，「38.5+0.8」，「0.8+0.6」，「0.8」なので，頻度の大きい「進む」が最終結果として出力される．

4. 誤認識データベースの自動構築

4.1. 実験環境

4.1.1. 音声認識装置

現在市販されている音声認識装置，及びフリーで配布されている音声認識装置は，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の２種類，また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２種類に分類される．

本稿では，特定話者の連続音声認識装置Ｄ，及び不特定話者の連続音声認識装置Ｊを用いて実験を行った．

音声認識装置Ｄは特定話者音声認識装置であり，使用前に「エンロール」という作業を行う必要がある．これは，話者の音声の特徴を装置に学習させる操作であり，あらかじめ定められたテキストを読み上げることによって，装置に話者のユーザデータが保存される．本稿では不特定話者に対する音声認識率の向上を目的としているため，本装置を使用する際は，エンロールとデータ取得に異なる話者を用いることで，擬似的に不特定話者音声認識を実現する．

音声認識装置Ｊは不特定話者音声認識装置であり，音響モデル，言語モデル，単語辞書から成る．音響モデルとは，あらかじめ大量の学習データから入力音声の各音素の特徴を隠れマルコフモデル(HMM)により表現したものである．言語モデルとは，語彙・文法あるいは言語統計などにより，発声内容を規定したモデルである．また単語辞書とは，各単語の読みを定義したものである．

4.1.2. 音声合成装置

本稿では，音声合成装置Ｌを用いて実験を行った．音声合成装置Ｌは「男性，女性，少年，少女，老人(男)，老人(女)，ロボットA，B，C」の９キャラクターの声を使い分けることができ，また２４段階のスピードに変えることができる．

4.1.3. 入力装置

一般的な音声認識装置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い．しかし，ヘッドセットの着用や，ピンマイクのセットの利用はユーザビリティに優れているとはいえない．そこで，本稿では無線ハンドマイクを利用し，音声認識専用の計算機で入力音声を受信する方式を取ることとした．またそのことで，受信機を増やすことができ，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音声認識装置で解析し，認識語をネットワークを介して１台の計算機で補正を行う．

4.2. 実験手法

4.2.1. 音声認識装置Ｄを用いたシステムの流れ

まず，音声合成装置を用いて音声出力し，無線ハンドマイクで音声を拾う．３人の被験者によってそれぞれエンロールされた３台のマシン毎に認識結果を出力する．音声合成から音声認識までの処理は自動的に動くようにシステム化されている（図 2）．

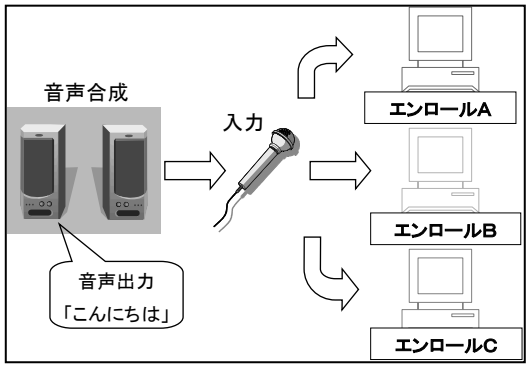

図 2：音声認識装置Ｄを用いたシステムのイメージ

4.2.2. 音声認識装置Ｊを用いたシステムの流れ

音声認識装置Ｊは直接音声出力したものをマイクで拾い，認識結果を返す機能の他に，あらかじめ作成したWAVE ファイルから認識結果を返す機能もある．自動構築の簡易さや効率面から本研究では後者の方法を採用した．

まず，日常会話で使用される２００単語の音声合成のWAVE ファイルを作成する．音響モデルが男性，女性，男女混合にされた３台のマシンに WAVE ファイルを渡し，認識結果を取得する（図 3）．

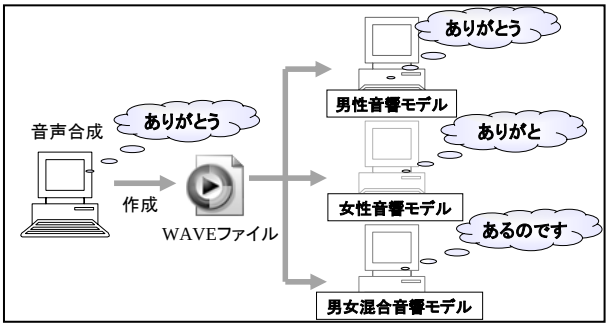

図 3：音声認識装置Ｊを用いたシステムのイメージ

4.3. 提案手法の認識率

4.3.1. 誤認識データベースの作成方法

日常会話で使用される２００単語を音声合成装置Ｌで扱える９キャラクター，５段階のスピードで音声出力し，その認識結果５回分を音声認識装置Ｄ，Ｊそれぞれ取得し，誤認識データベースに登録した．

4.3.2. 結果

この誤認識データベース検証のために，テストデータとして，被験者１６名（男性８名，女性８名）が２００単語を１回読み上げた結果を用意し，作成したデータベースを用いて一致度補正を行った．その結果を以下に示す（図 4）．

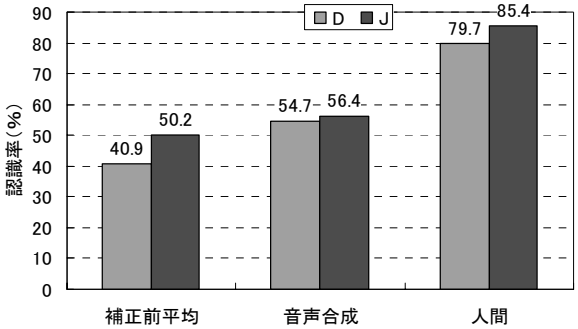

図 4：音声合成での誤認識データベースの認識率

補正前平均とは, それぞれの音声認識装置が返してくる認識結果３つ分の補正なしでの平均認識率のことである. また人間とは, 被験者１５名が３回ずつ入力した誤認識データベースでの認識率である.

4.3.3. 考察

図 4 を見ると, 人間で構築した誤認識データベースと比べると認識率は低い. 表 2 は音声合成による誤認識データベースの一部である. これを見ると見出し語に対して正解語の割合が低いことがわかる. 場合によっては, 一度も正解語が出現していないことがある. 人間の認識結果には正解語が多く出現すると思われる. このため, 低い精度になったと考えられる.

表 2：音声合成での誤認識データベースの一部

見出し語	属性１, 重み	属性２, 重み	属性３, 重み	…
右	２, 19.4	０, 10.5	みぎ,3.0	…
後退	こうがい, 6.1	が,4.9	はん,4.0	…
加速	かかく, 11.7	・, 8.3	かそく, 3.6	…
曲がる	がなる, 9.9	かなり, 7.2	な。, 5.5	…

5. 改良１

5.1. 手法

音声合成の誤認識データベースに正解語を最大の重みで付け加える処理を行った. 具体的な方法としては, 正解語がない場合は重みを最大にして追加し, ある場合は正解語の重みが最大になるように修正した（表 3）.

表 3：正解語の重み最大にした誤認識データベースの一部

見出し語	属性１, 重み	属性２, 重み	属性３, 重み	…
運ぶ	はこぶ, 500	かかげ, 8.0	はこん, 5.5	…
握る	にぎる, 500	に, 11.6	２０,10.1	…
右	みぎ, 500	２, 19.4	０, 10.5	…

5.2. 結果

テストデータとしては, 前章で用いたテストデータを使用し, 一致度補正を行い, 精度を測定した（図 5）.

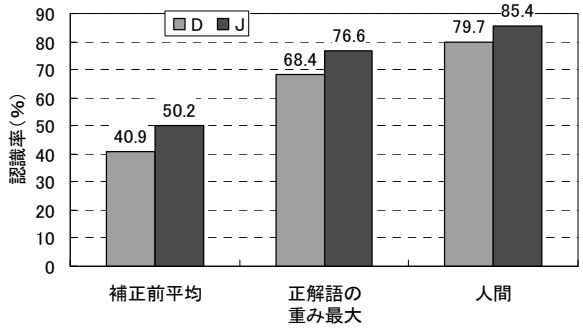

図 5：正解語の重みを最大にしたときの認識率

正解語の重みを最大にすることにより, 音声認識装置Ｄでは約 14%, Ｊでは約 20%の精度向上となった. これは, 前章で述べたように, 正解語が存在しなかったことと, 正解語の重みが低かったことが解消されたためだと考えられる.

6. 改良２

前章で構築した誤認識データベースは, 人間が構築した誤認識データベースと比べると精度は低い. これは音声合成装置の性能がまだまだ人間に追いついていないためである. そこで, 従来の誤認識データベースに人間の認識結果を加える処理を行い, 精度向上を図ろうと考えた.

6.1. 1人分の認識結果を追加

6.1.1. 手法

本研究はなるべく手間をかけずに高い認識率を得ることが一つの目的であるため, 多くの人間の認識結果を加えることはできない. そこで, まず１人の認識結果を追加して認識率の変動を検証した. ここで問題となるのが, 加える認識結果が男性か女性かで精度に違いがあるのかということである. そのため, 男性の認識結果を加えた場合と, 女性の認識結果を加えた場合とでの認識率の比較を行った.

6.1.2. 結果

正解語の重みを最大にした誤認識データベースに男性被験者の認識結果１回分を加えたデータベース３種類と, 女性被験者の認識結果を１回分加えたデータベース３種類の計６種類用意した. 認識結果１回分とは音声認識装置Ｄの場合はエンロールＡ, Ｂ, Ｃの結果３つ分であり, Ｊの場合は男性, 女性, 男女混合の音響モデルの結果３つ分である. この時の結果を以下に示す.

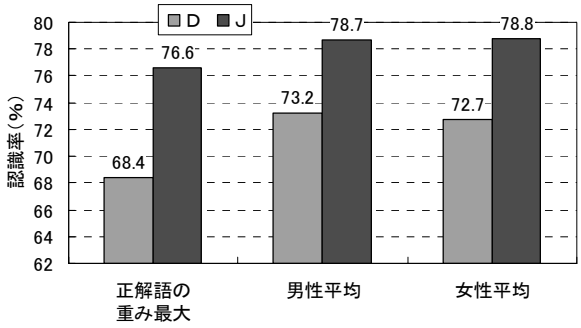

図 6：１人分の認識結果を追加したときの認識率

6.1.3. 考察

次に, この結果を解析した（表 4）.

表 4：男性, 女性それぞれの認識率の上昇量

		男性テストデータ（%）	女性テストデータ（%）
D	男性追加	6.9	3.6
	女性追加	3.3	5.3
J	男性追加	2.8	1.4
	女性追加	1.7	2.6

表 4 より, テストデータと同性の認識結果を加えた方が, より認識率が向上していることが分かる.

6.2. 男女の認識結果を追加

6.2.1. 手法

前節で述べたように男性と女性の認識率の上昇量に違いがある. このことから, 男女両方の認識結果を入れることにより, 男女両方のテストデータの認識率が上がり, 認

識率が向上するのではないかと考えられる.

6.2.2. 結果

テストデータは 4.3.2 節のテストデータを使用して実験を行った. 比較するために, 男性の結果２人分追加した誤認識データベースと女性の結果２人分追加した誤認識データベースも用意し, 精度を測った（図 7）.

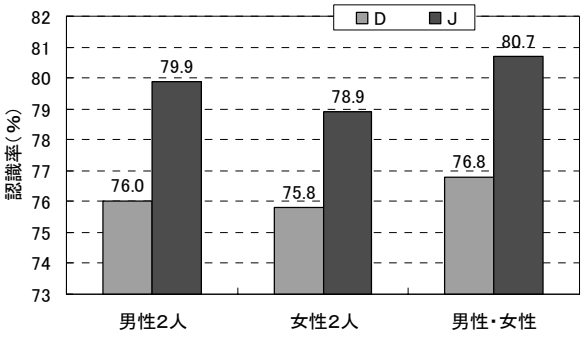

図 7：２人分の認識結果を追加したときの認識率

音声認識装置D, Ｊともに男性, 女性片方だけの結果を加えた誤認識データベースより, 男性と女性両方の結果を加えた誤認識データベースのほうで認識率が高くなった. 音声認識装置D, Ｊともに約 1%の精度向上となった.

6.3. 認識率の推移

人間の認識結果を追加すればするほど, 認識率は向上するのか, もしくはどこかで収束するのかを検証した（図 8）.

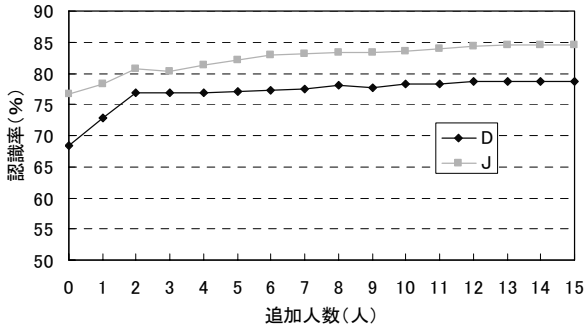

図 8：認識率の推移

図 8 から, 追加する人数を増やすと認識率は上昇していることがわかるが, 追加人数を１人増やしたときの上昇量はわかりにくい. そこで, 上昇量の推移を以下に示す.

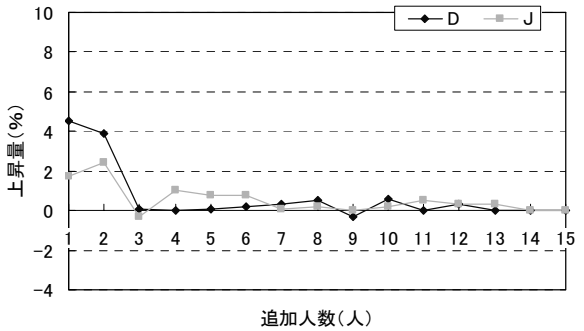

図 9：認識率の上昇量の推移

図 9 について説明すると, 追加人数ｎ人のときの上昇量は図 8 の追加人数ｎ人の時の認識率から追加人数ｎ－1 人の時の認識率を引いた値である. たとえば, 追加人数１人のときの上昇量は図 8 の追加人数１人の時の認識率から追加人数０人の時の認識率を引いた値である.

図 9 を見ると, 追加人数が１～２人のときに最大の上昇量となり, それ以降, 追加人数を増やすにつれて, 認識率の上昇量が減少していることがわかる. 追加人数が３人以上からは, 上昇量は 1%未満になっている. 単語２００語の場合, 人の認識結果を１回取得するのに, 約２０分の時間を要する. 現在は, 単語２００語に限定しているため短時間で済むが, 日常会話をするためには, より多くの語彙数が必要となる. それに伴い人の認識結果を取得する場合にも多大な労力と時間が必要になる. 上昇量と労力の関係から, 追加人数は２人が適当であると思われる.

7. 今後の課題

音声合成装置Ｌでは９キャラクターの声質が使えるが, 人間の声は多種多様であるため, 高い認識率を得るには少ないと思われる. そこで今後の課題として, 他の音声合成装置との組み合わせが必要であると考える. これにより多くの声質の認識結果を誤認識データベースに登録することができ, 精度向上が図れるのではないかと考えられる.

8. おわりに

誤認識データベースの構築には, 人手では多くの時間と労力を要する. そこで, 音声合成装置を用いて誤認識データベースを自動構築し, 人の認識結果を加えて精度向上を行った. 単語補正では, 音声認識装置Ｄで 76.8%, Ｊで 80.6%の認識率を得, その有効性を示した. また, 音声認識装置に関係なく, 補正することにより認識率が向上していることから, 音声認識装置に依存しない補正システムである.

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「人間と生物の賢さの解明と, その応用」における研究の一環として行った.

参考文献

[1] 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.71-76, 2004.