対話型ロボットのための複数の音声認識器を利用した発話理解

Speech understanding with a multiple recognizer for a robot

嶋田 和孝[1*]　遠藤 勉[1]
Kazutaka Shimada[1]　Tsutomu Endo[1]

[1] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
[1] Department of Artificial Intelligence, Kyushu Institute of Technology

Abstract: In this paper, we describe a speech understanding method for a livelihood support robot. The speech understanding method consists of some speech recognizers. We focus on two types of recognizers; a large vocabulary continuous speech recognizer and several domain-specific speech recognizers. We handle these different speech recognizers selectively and integratively. The selective usage is to detect the most suitable output from outputs of all recognizers, that is an utterance verification task. In this task, We use the edit distance between each output for the verification. The integrative usage is to resolve anaphoric expressions in speech inputs. Our method with the multiple recognizer is based on a scoring process using the confidence measure from each speech recognizer and the distance between an anaphoric expression and each antecedent candidate.

1　はじめに

近年の音声認識技術の向上により，音声入力を用いた実用的な対話システムの実現を目指した研究が進められている．しかしながら，実用的な音声対話システムの構築には，音声認識結果の曖昧箇所への対応や音声認識そのものの精度向上が不可欠な状況である．

本論文では，複数の音声認識器を併用することで，特定のタスクやドメインにロバストでかつ自由発話にも柔軟に対応可能な音声理解手法について提案する．我々は現在，介護施設や病院などで暮らしている施設入居者の日常生活や施設スタッフの作業を支援する施設内生活支援ロボットの構築を進めている．本研究はその入力部の 1 つである音声認識部にあたる．提案手法は，大きな分類では，タスク依存の認識器と一般的な言語モデルに基づく認識器から構成される．図 1 に提案システムのイメージを示す．タスク依存の認識器は，主にユーザからの指示を理解する役割を持ち，正確な認識が可能なことが望まれる．一方で，タスク依存の認識器では十分に認識できないような発話に対しては，大語彙認識器を用いて，認識を行う．それぞれの認識結果を選択的にもしくは統合的に扱うことができれば，柔軟でかつロバストな認識器を実現できる．

本論文では，提案手法における，認識器の選択的利用と統合的利用の 2 点について述べる．ここで，選択的利用とは，複数の認識器の結果から最適な 1 つの結果を選ぶことを指し，統合的利用とは，主に大語彙認識の認識結果を利用したタスク依存認識器の認識結果に対する照応解析を指す．

*連絡先： 九州工業大学 知能情報工学研究系
〒 820-8502 福岡県飯塚市川津 680-4
E-mail: shimada@pluto.ai.kyutech.ac.jp

2　選択的利用

複数の特性の異なる認識器を組み合わせて処理を行う場合，どの認識器の結果を最終的に利用するか，という問題が発生する．今回のシステムでは，入力された発話がロボットへの命令発話なのか，それともしれ以外の発話（雑談など）なのかを分別する必要がある．例えば，図 1 (b) で,「それを拾ってくれる？」という入力について，どの認識器の結果を採用するかは大きな問題である．これは発話検証というタスクの範疇に入る [8].

2.1　手法

本手法では，それぞれの認識器の出力の類似性に着目する．タスク依存認識器は，もし入力が想定された発話であれば，高い認識精度が得られるはずである．一方，大語彙認識器は，自身の音響モデル・言語モデルに基づいて，最適と思われる単語列を出力する．この

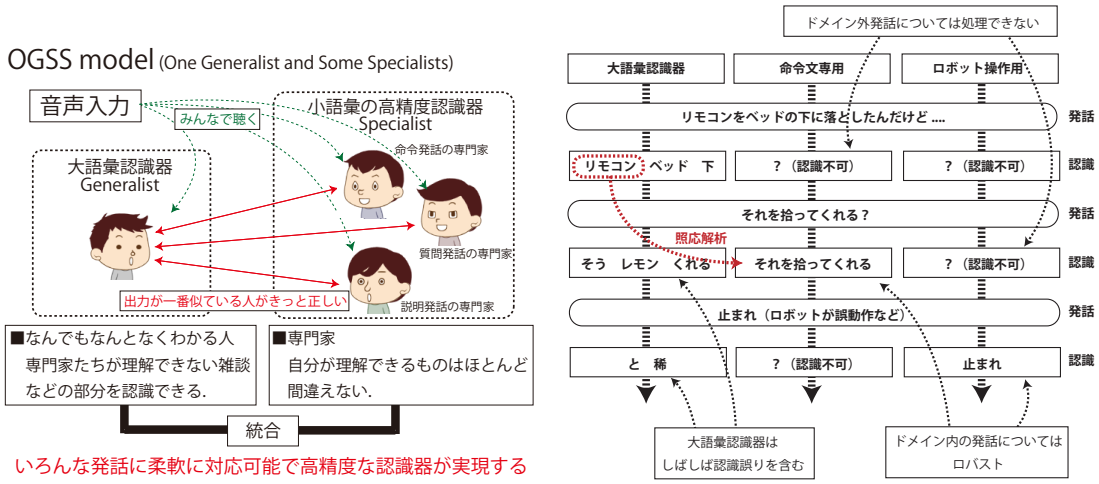

図 1: 提案手法のイメージとその有効性

場合，2 つの認識器の結果は，少なくとも音素レベルでは比較的類似していると考えられる．逆に，想定外の発話の場合，大語彙認識器は命令発話を認識したときと同様に，最適と思われる単語列を出力するが，タスク依存認識器は持ち合わせている語彙と文法を基にあまり適切でない単語列を出力すると考えられる．すなわち，この場合は，2 つの認識結果は，音素レベルでも必ずしも一致しない可能性がある．

そこで，提案手法では，2 つの認識結果の編集距離を算出し，それを分別のための特徴とする．具体的には，それぞれの認識器の出力全体（すなわち発話単位）での編集距離と，各単語ごとの編集距離の 2 つの距離値（ともに音素レベル）を DP マッチングにより算出し，以下のように分別処理を実行する [3].

1. 発話単位の編集距離が閾値未満の場合，編集距離が最小のドメイン依存認識器の結果を採用する．
2. 単語単位の編集距離の平均値で比較する．
   閾値以上：大語彙認識器の結果を採用する．
   閾値未満：最小のドメイン依存認識器の結果を採用する．

2.2 実験結果と考察

実験では，音声認識器として Julius/Julian[2] を用いた．Julius については，音響モデル，言語モデルとも同ソフトウェアに添付されているオリジナルのモデルを利用した．タスク依存認識器としては，以下のような 4 種類の認識器に関する文法・語彙辞書を用意した．

- 患者からの命令発話：机の上のリモコンを取ってくれる？など
- 看護士からの命令発話：この食事を 501 号室に運んで，など

表 1: 分別精度.

認識器	再現率	適合率	F 値
患者命令	0.965	0.873	0.917
看護士命令	0.965	1.000	0.982
制御命令	0.930	1.000	0.964
質問発話	0.975	0.878	0.924
雑談 (LVCSR)	0.765	0.827	0.795
平均	0.920	0.916	0.916

- ロボット制御命令：止まれ，50cm 右に移動，など
- 質問発話：田中さんはどこにいる？など

各々の認識器は 200 単語弱の語彙とそれに基づく 100 パターンほどの文法で構成されている．

実験データとして 100 発話用意し，10 人の被験者で評価した．表 1 に分別精度を表す．ここで分別精度とは，例えば，患者命令発話に対して患者命令用の認識器の結果が正しく選択されたかを指す[1]．提案手法は，シンプルな手法であるが，高い分別精度を得ることができている．表から分かるように，ほとんどの間違いは，雑談との誤判別であり，命令発話内での誤分別は少ない．そこで，命令検出率の観点でも評価する．ここで，命令検出率とは，入力が命令発話（4 つのタスク依存認識の扱う範疇）であった場合の再現率および適合率である．図 2 は，閾値を変化させた場合の命令検出率である．図から分かるように，閾値を厳しくすれば，命令発話に対して 10 回中 7 回程度しか反応しないが，その場合はほぼ間違いなく命令を認識できることになり，提案手法がロバストであることが分かる．

[1] 雑談 (命令以外) の場合は，大語彙認識器の結果が選択されたかどうか．ただし，認識結果の文字列が正しいかは問わない．

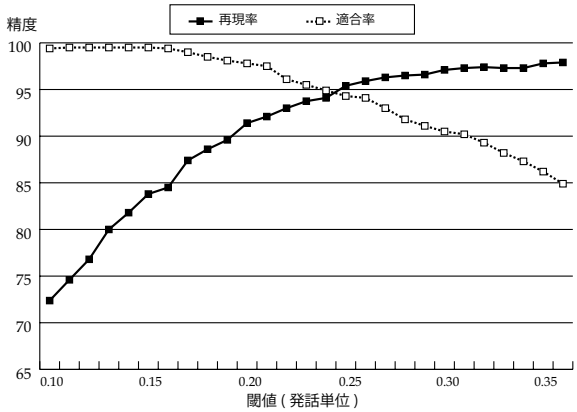

図 2: 命令検出率.

```
<frame case="predicate" semantic="bring_here_V"
       surface="(持って来る|持って来て|取って来る|取って来て)"
       required="obj" requiredSub="loc">
   <case case="goal" semantic="tar" marker="((*) さん|(*) くん)" />
   <case case ="loc" semantic="loc" marker="(*) の (上|中|下|横|手前|奥) の" />
   <case case ="loc" semantic="loc" marker="(*) の (上|中|下|横|手前|奥) にある" />
   <case case="loc" semantic="loc" marker="(ここ|そこ|あそこ) の" />
   <case case="loc" semantic="loc" marker="(ここ|そこ|あそこ) にある" />
   <case case="obj" semantic="bring_N" marker="(*) を" />
   <case case="sour" semantic="loc" marker="(*) から" />
   <case case="sour" semantic="loc" marker="(ここ|そこ|あそこ) から" />
</frame>
```

図 3: 解析辞書の例.

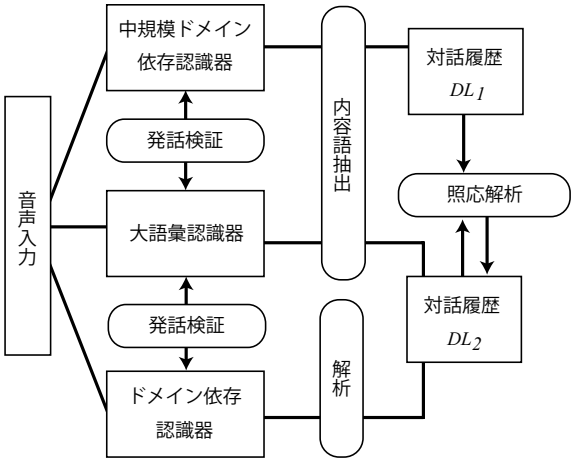

図 4: 照応解析の流れ.

表 2: 照応解析の精度.

手法	累積スコアの場合	最大スコアの場合
精度	68.9%	**71.7%**

関連研究として，Komatani ら [1] は，音声認識器が出力する音響尤度に着目した発話検証の枠組みについて提案している．発話検証に音響尤度の差を利用することは，理論的にも適しており，この手法を我々の手法を統合することは，有効であると考えられる．簡易の実験を行ったところ，分別精度は提案手法と同レベルであった．今後は，詳細なエラー解析をし，先行研究との統合を模索する．また，各認識器を階層化することも認識精度向上には有効であり [6]，今後の課題の一つである．

3 統合的利用

次に，複合認識器の統合的利用について述べる．本論文における統合的利用とは，複数の認識器の結果を利用した照応解析である．

3.1 手法

まず，システムは，編集距離によって得られた音声認識の結果を解析する．具体的には，図 3 のような解析辞書を用意し，それに基づき，格構造のようなフレームを生成する．

次に，この生成されたフレームを走査し，指示詞やゼロ代名詞が存在した場合は，その発話以前に認識された単語を対象とし，先行詞を同定する．先行詞の同定処理では，音声認識器の出力するスコアや照応詞との距離，話者変更などの情報に着目し，スコアリングを行う．具体的には，過去 n 発話中での単語の累積スコアを用いる手法 [5] や各単語ごとにスコアを計算し，最大のスコアをもつ単語を選択する手法 [4] などがある．また，先行詞候補をできるだけ正確に補足するために，対象となる環境でよく使われる語彙のみで構成される中規模な認識器を用いることで，精度向上を図る [4]．処理の流れを図 4 に示す．

3.2 実験結果と考察

実験として，53 個の照応詞を含む 206 発話からなる仮想対話について，被験者 2 名で評価した．その結果を表 2 に示す．2 つの手法を比較すると，最大スコアの精度が 3%程度高い．これは，音声認識の精度と累積値を用いるというスコア計算方法の特徴に起因している．音声認識結果には基本的に誤りが含まれる．特に，挿入誤りが生じた場合，累積スコアを用いると，その誤りによって対話履歴中に存在する多くのノイズに過剰反応してしまい，精度が低下したものと考えられる．

提案手法における照応解析の誤りのほとんどは，音声認識誤りが原因である．例えば，先行詞そのものがそもそも対話履歴に存在しない場合，照応解析は必ず失敗する．すなわち，音声認識誤りを含むことを前提に，照応解析を行うことが重要な課題である．音声文書検索においては，音節同士のマッチングを取ることで，未知語（すなわち音声認識の結果には含まれないもの）の検索を行う手法が提案されている [7]．ここで，

あるクエリにおける音声文書検索の問題を，照応解析における先行詞の探索だと考えることで，音声認識結果に正しい先行詞が単語として含まれていない場合でも，音節的に類似箇所を検出することが可能になる．これは，照応解析の精度向上に繋がると考えられ，現在研究を進めている．

4 おわりに

本論文では，対話型ロボットを対象とした複数の認識における発話理解の手法について述べた．提案手法では，複数認識器の選択的利用として発話検証を，統合的利用の例として照応解析について述べた．発話検証については，シンプルな手法であるが，高い識別精度を得た．照応解析については，十分な精度とはいえず，その精度向上は今後の大きな課題である．

本論文では，音声のみを対象としたが，一般にロボットには様々なセンサーが装備されている．現在我々は，画像と音情報に基づく発話区間推定 [10] や人物識別 [9] も行っており，これらのモダリティから得られる情報は，音声理解においても有用である．例えば，発話中の人物が誰か分かれば，その人物に合わせた語彙情報に基づく認識器を優先して利用することが可能になるだろう．システムのマルチモーダル化は重要な課題の一つである．

また，本システムは，ロボット用の汎用的なコンポーネントの実現を目指しており，各コンポーネントの相互運用性を向上させるために，インターフェースの規格を共通化する作業が進められている [11]．本システムは，将来的には，その枠組みに沿い，RT コンポーネントして，公開する予定である．

謝辞

本研究は，次世代ロボット知能化技術開発プロジェクト（NEDO）における「施設内生活支援ロボット知能の研究開発」の成果の一部である．

参考文献

[1] K. Komatani, Y. Fukubayashi, T. Ogata, and H. G. Okuno. Introducing utterance verification in spoken dialogue system to improve dynamic help generation for novice users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 202–205, 2007.

[2] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. In *Proceedings of Eurospeech*, pp. 1691–1694, 2001.

[3] K. Shimada, S. Horiguchi, and T. Endo. An effective speech understanding method with a multiple speech recognizer based on output selection using edit distance. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC22)*, pp. 350–357, 2008.

[4] K. Shimada, N. Tanamachi, and T. Endo. Combination of 3 types of speech recognizers for anaphora resolution. In *Proceedings of the 24nd Pacific Asia Conference on Language, Information and Computation (PACLIC24)*, 2010.

[5] K. Shimada, A. Uzumaki, M. Kitajima, and T. Endo. Speech understanding in a multiple recognizer with an anaphora resolution process. In *Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (PACLING2009)*, pp. 262–267, 2009.

[6] T. Yokoyama, K. Shimada, and T. Endo. A hierarchical multiple recognizer for robust speech understanding. In *Proceedings of The Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, LNAI 6230, pp. 706–712, 2010.

[7] 西崎, 中川. 音声認識誤りと未知語に頑健な音声文書検索手法. 電子情報通信学会論文誌 D-II, J86-D2(10):1369–1381, 2003.

[8] 河原達也. 音声でスライド画面を操作する. *Bit*, 32(4):436–439, 2000.

[9] 山口, 嶋田, 榎田, 江島, 遠藤. 顔特徴とコンテキスト情報に基づく顔の隠れに頑健な人物識別. 電子情報通信学会, パターン認識・メディア理解研究会 (PRMU), 信学技報, 第 109 巻, pp. 25–30, 2010.

[10] 元吉, 嶋田, 榎田, 江島, 遠藤. 対話型ロボットのための口領域動画像と音情報に基づく発話推定. 情報処理学会 第 71 回全国大会, 5T-3, pp. 429–430, 2009.

[11] 松坂. RT ミドルウェアによるロボットアーキテクチャ - コミュニケーションシステム. 日本ロボット学会誌, 28(5):566–567, 2010.