1O2-3

環境情報を考慮したロボットによる音声命令理解

Comprehension of Spoken Instruction by a Robot Using Environment Information

小堀　嵩博 *1　中村　友昭 *1　長井　隆行 *1　岩橋　直人 *2　船越　孝太郎 *3
Takahiro Kobori　Tomoaki Nakamura　Takayuki Nagai　Naoto Iwahashi　Funakoshi Kotaro

中野　幹生 *3　金子　正秀 *1
Mikio Nakano　Masahide Kaneko

*1電気通信大学
The University of Electro-Communications

*2岡山県立大学
Okayama Prefectural University

*3(株) ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd

We previously proposed robust comprehension method for spoken instruction based on language information. Our proposed model integrated speech recognition, an action identification based on SVM, a slot extraction based on CRF and a co-occurrence relationship between an action type and slots. However, there are instructions that are still difficult to be comprehended only using language information. To comprehend such instructions we introduce environment information, which can be obtained by a robot, to the instruction comprehension. Experiments using utterances for a task in RoboCup@Home competition suggest that effectiveness of using environment information.

1. はじめに

本研究では，音声による命令を理解するロボットの実現を目的とする．音声を利用する利点として，柔軟性と簡潔さがあげられる．タブレットやキーボード等を用い，メニューからロボットに実行してほしい行動をユーザに選択してもらう方法も考えられるが，複雑な行動を指示できるようにするには，インタフェイスが煩雑になる可能性がある．一方，音声による指示は，普段我々が使用している自然言語を利用するため，複雑な命令であっても比較的容易に表現することができる．このようなことから，近年，音声命令の理解が可能なロボットの実現が期待されている．しかし，言い回しの多様性や音声の誤認識への対処など，音声命令理解には様々な課題がある．

本稿では，RoboCup@Home[*1] において実施されているロボットが音声命令を理解し，実行するタスクである GPSR (General Purpose Service Robot) を対象とし，GPSR におけるロバストな音声命令理解モデルを提案する．RoboCup@Home とは，家庭用ロボットの性能の向上を目的とした競技会であり，GPSR はその中のタスクの 1 つである．GPSR では，ユーザがロボットへ音声により命令をし，その命令の理解度と命令の達成度が評価される．このタスクでは，何をどのような言い回しで提示されるか事前には決まっていないため，様々な言い回しに対応する必要がある．さらに，競技の解説者の声や，観客の声などのノイズの影響により，命令音声の誤認識も発生する．このような問題を解決するため，本稿では言語情報と環境情報を統合した音声命令理解モデルを提案する．ここでの命令理解は，音声による命令からロボットがすべき行動タイプの識別と，その行動に必要な情報（スロット）を抽出することと定義する．例えば，「キッチンに行って」というような命令では，行動タイプを “go” と識別し，スロットとしてロボットが行くべき場所である “キッチン” を抽出することで命令理解を行う．提案モデルは，音声認識の曖昧性と，環境情報と命令内容の依存関係を表現したベイジアンネットワークを用いる．ベイジアンネットワークの利点として，様々な情報の依存関係を容易に表現できること，また新たな情報の追加が容易である点が挙げられる．提案モデルにより複数の情報の依存関係を考慮することで，互いに情報を補完することができ，一部の情報が曖昧であったととしても言語理解が可能である．

GPSR の音声命令理解には構文解析を利用する手法が多く提案されている [Schiffer 12, Chen 15, Seib 15]．しかし，構文解析のみを用いた手法では音声認識に誤りがあると解析が難しいことがある．一方，提案モデルでは，様々な情報を統計的に統合しているため音声の誤認識に対しても頑健である．

ロボットに限らない音声命令理解については，統計的手法を用いた手法も多く，最近では深層学習を用いた手法もある [Xu 13]．統計的手法を用いた場合，ある程度音声の誤認識に対応できるが，ロボットの命令理解には適用されておらず，どのような手法が GPSR のような実環境で動作するロボットにおいて有効かは明らかではない．GPSR では，実環境において命令理解を行うため，環境の情報も導入することでより高精度な命令理解が可能となると考えている．

また，周囲の環境情報を用いることで音声認識の言語モデルを変更し，音声認識精度を向上させる手法が提案されている [Roy 05]．この手法は積み木の状況を利用してモデルを変更することで，音声認識の精度を向上させている．しかしながら，GPSR のような様々な物体が存在する実環境で有効に働くかは明らかではない．

2. 音声命令理解モデル

本稿では複数の情報をベイジアンネットワークにより統合した音声命令理解モデルを提案する．前述のように音声命令理解は，ロボットが事前に取得した環境情報 e と，与えられた命令音声 u から，その文字列表記 r と，命令意図 i を推定することである．すなわち，命令理解は，以下の条件付き確率を最大化する i と r を選択することで実現する．

$$\arg\max_{i,r} P(i|e,r)P(r|u) \tag{1}$$

$P(r|u)$ は命令音声 u の文字列表記が r である確率，$P(i|e,r)$ は環境 e の中で，文字列 r の命令意図が i である確率を表している．

ここで，命令意図 i は行動タイプ a とスロット s から構成されると考える．行動タイプとはロボットがすべき行動であり，

連絡先: 小堀嵩博，電気通信大学, 〒 182-8585，東京都調布市調布ヶ丘 1-5-1，t_kobori@radish.ee.uec.ac.jp

*1 RoboCup@Home, http://www.robocupathome.org/.

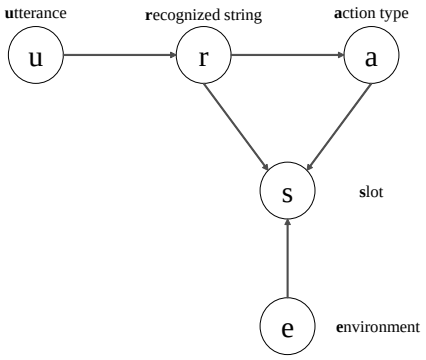

図 1: 音声命令理解のグラフィカルモデル

スロットはその行動を実行するのに必要な情報のことである．すなわち，式 (1) は次式のように変形することができる．

$$\arg\max_{a,s,r} P(a,s|e,r)P(r|u) \tag{2}$$

さらに，$P(a,s|e,r)$ を以下のように 4 つの要素に分解して考える．

$$\begin{aligned} P(a,s|e,r) &= P(a|r)P(s|r,a,o) &(3)\\ &\propto P(a|r)P(s|r)P(s|a)P(s|e) &(4)\end{aligned}$$

ただし，ここではスロット s は r, a, e から独立に影響を受けると仮定し，$P(s|r,a,o) \propto P(s|r)P(s|a)P(s|e)$ とした．$P(a|r)$ は文字列 r が表す行動タイプが a である確率，$P(s|r)$ は文字列 r に含まれるスロットが s となる確率である．さらに，$P(s|a)$ は行動タイプとスロットが共起する確率であり，この確率により行動タイプとスロットの整合性がとれた組み合わせを選択することができる．例えば，行動タイプが “go” となる命令文中には，ロボットが行くべき場所名が含まれる可能性が高い．そのような行動タイプとスロットの共起関係を $P(s|a)$ で表現している．さらに，$P(s|e)$ は環境 e から得られる情報であり，ロボットが環境から得た情報から物体名と場所名の組がスロットとして発生する確率を表している．例えば，コーラがキッチンテーブルの上に存在することをロボットが知っていれば，「○○からコーラを持ってきて」のように場所名を正しく聞き取ることができなくとも，ロボットは命令を正しく理解することができる．式 (4) を用いることで，式 (2) は

$$\begin{aligned} &P(a,s|o,r)P(r|u)\\ &\quad\propto P(a|r)P(s|r)P(s|a)P(s|e)P(r|u) \end{aligned} \tag{5}$$

となる．式 (5) の確率変数の依存関係を表したグラフィカルモデルが図 1 である．

3. 音声命令理解モデルの実装

音声命令理解は，前述の条件付確率を定義し，全てのありえる文字列 r，行動タイプ a，スロット s の組み合わせの中で，式 (5) を最大化する組を選択することである．しかし，その組み合わせは膨大であり，全てを考慮することは現実的ではない．そこで，r の取り得る値は音声 u の認識結果の N-best の文字列とする．また，a と s の取り得る値は，それぞれの r から，行動タイプ識別を行った結果の M-best と，スロット抽出を行った結果の L-best とする．すなわち，$N \times M \times L$ 個の組み合わせの中から式 (5) を最大化する組を，音声命令理解結果とする．

本稿では，行動タイプ識別には Support Vector Machine (SVM) を用い，13 種類の行動タイプを識別する．また，スロット抽出には Conditional Random Field (CRF) を用い，4 種類のスロットを抽出する．これらの行動タイプとスロットの種類は，GPSR で使用される可能性がある行動を人手で分類し決定した．

3.1 音声認識

ここでは，式 (5) 内の $P(r|u)$ の実装について説明する．音声命令の N-best の認識結果から，n 番目の認識結果 r_n とその尤度を得ることができる．しかし，1 位の認識結果の尤度が非常に高い値となることが多く，尤度をそのまま $P(r|u)$ として利用すると，1 位の認識結果のみを利用することになってしまう．誤認識が発生しやすい環境では，2 位以下の認識結果に正解が含まれている場合がある．そこで，下位の認識結果を利用できるように，音声認識スコア $S_{sr}(r_n)$ を尤度が大きいものから $1.0, 0.9, 0.8, \cdots$ と設定する．この音声認識スコアを用い確率 $P(r|u)$ を以下のように定義する．

$$P(r|u) \propto S_{sr}(r_n)^{\alpha} \tag{6}$$

ただし，α はそれぞれのスコアの重要さを表す重みである．

3.2 SVM による行動タイプの識別

本稿では，ペアワイズ法を使用し，多クラスに拡張した SVM により行動タイプ識別を行う．SVM で用いる素性として，認識された命令文を Bag of Words (BoW) 表現へと変換した単語の発生頻度ヒストグラムを使用する．SVM による行動タイプ識別により，命令文 r_n に対する M-best の識別結果 $a_m(r_n)$ とそのスコア $S_{svm}(a_m(r_n))$ を得ることができる．ここで，このスコアを用い確率 $P(a|r)$ を以下のように定義する．

$$P(a|r) \propto S_{svm}(a_m(r_n))^{\beta} \tag{7}$$

ただし，β は行動タイプ識別のスコアに対する重みである．

3.3 CRF によるスロット抽出

スロットは，各形態素に対して IOB2 タグを出力するよう学習した CRF により抽出した．形態素とはそれ以上分解したら意味をあらわさくなる最小の単位である．CRF の素性として，形態素解析によって得られる表層系・原型・品詞の 1-gram と 2-gram を用いる．CRF により，1 つの命令文 r_n から L-best のスロット抽出結果 $s_l(r_n)$ とそのスコア $S_{crf}(s_l(r_n))$ を得ることができる．ここで，このスコアを用い確率 $P(a|r)$ を以下のように定義する．

$$P(s|r) \propto S_{crf}(s_l(r_n))^{\gamma} \tag{8}$$

ただし，γ はスロット抽出のスコアに対する重みである．

3.4 行動タイプとスロットの共起関係

ある行動タイプに必要となるスロットの種類はあらかじめ決まっており，それらは共起する可能性が高いと考えられる．そこで，行動タイプとスロットの組み合わせの整合性を表現した共起スコア $S_f(a_m(o_n), s_l(o_n))$ を導入する．

$$S_f(a,s) = \frac{2 \times R \times P}{R+P} + \phi \tag{9}$$

$$R = \frac{(\text{スロット } s \text{ のうち } a \text{ で必要な数})}{(a \text{ で必要なスロットの数})} \tag{10}$$

$$P = \frac{(\text{スロット } s \text{ のうち } a \text{ で必要な数})}{(\text{スロット } s \text{ の数})} \tag{11}$$

$$\phi = \begin{cases} \xi & (R = 0 \text{ and } P = 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{12}$$

この共起スコアでは，再現率 (R) と精度 (P) の調和平均である F 値の考え方を利用している．R は必要なスロットが抽出されているかを評価し，P はスロットが過剰に抽出されていないかを評価する．すなわち，このスコアは行動タイプに対して適切なスロットが抽出されたかを評価している．ここで，ϕ はスロットが 1 つも抽出されない場合にスコアが 0 とならないようにするためのパラメータである．また，各行動タイプで必要となるスロットは，人手によって定義した．この共起スコアを用い，確率 $P(s|a)$ を以下のように定義する．

$$P(s|a) \propto S_f(a_m(r_n), s_l(r_n))^{\delta} \tag{13}$$

ただし，δ は共起スコアに対する重みである．

3.5 物体情報

本稿では，ロボットが取得可能な環境情報の 1 つとして，物体存確率を用いた．物体存在確率とは，ある場所に物体が存在する確率を表している．環境内をアクティブ探索法 [Murase 00] を用いて物体認識を行うことで，場所名が p となる場所で物体 o が存在するスコアを以下のように計算する．

$$S_{obj}(s) = \begin{cases} \sum_{n=1}^{N_{po}} \frac{AS_i(p,o)}{N_{po}} & (N_{po} \neq 0) \\ \phi_1 & (N_{po} = 0) \\ \phi_2 & (\text{物体情報が必要でない } a) \end{cases} \tag{14}$$

ただし，N_{po} は場所 p で物体 o を観測した回数であり，$AS_i(p,o)$ は o を i 回目 $(i = 1, 2, \cdots, N_{po})$ に認識した際の物体認識スコアである．また，ϕ_* は物体存在スコアが 0 となることを防ぐためのパラメータであり，ϕ_1 は物体がその場所で観測されなかった場合のパラメータ，ϕ_2 は物体存在スコアを用いることができない行動 (目的地に向かうなど) におけるパラメータである．この物体存在スコアを用いて，確率 $P(s|e)$ を以下のように定義する．

$$P(s|e) \propto S_{obj}(s_l(r_n))^{\epsilon} \tag{15}$$

ただし，ϵ は物体存在スコアに対する重みである．

3.6 マルチモーダル音声命令理解

以上より，式 (5) は以下のようになる．

$$\begin{aligned} & P(a, s, r|u, e) && (16)\\ & \propto P(r|u)P(a|r)P(s|r)P(s|a)P(s|e) && (17)\\ & \propto S_{sr}(r)^{\alpha} S_{svm}(a)^{\beta} S_{crf}(s)^{\gamma} S_f(a, s)^{\delta} S_{obj}(s)^{\epsilon} && (18)\\ & \equiv S_{total}(a, s, r|u, e) && (19) \end{aligned}$$

$S_{total}(a, s, r|u, e)$ は，音声認識スコア，行動タイプ識別スコア，スロット抽出スコア，共起スコア，物体存在スコアを統合した統合スコアである．N-best の音声認識と，M-best の行動タイプ識別，L-best のスロット抽出の組み合わせの中から，このスコアが最大となる組を 1 つ決定する．

$$\begin{aligned} \hat{a}, \hat{s}, \hat{r} &= \arg\max_{a,s,r \in C} S_{total}(a, s, r|u, e) && (20)\\ C &= \{a_m(r_n), s_l(r_n), r_n| \\ &\qquad 1 \leq n \leq N, 1 \leq m \leq M, 1 \leq l \leq L\} && (21) \end{aligned}$$

ただし，式 (18) 内の α，β，γ，δ，ϵ はそれぞれのスコアに対する重みであり，吉川らの手法 [吉川 15] を用いて推定した．

表 1: 学習データとテストデータの命令文中に含まれる命令数

	学習データ (1414 文)	テストデータ (155 文)
1 つの行動	339	40
2 つの行動	48	7
3 つの行動	1027	108

3.7 複数の行動が含まれる命令の理解

1 つの命令文に複数の行動が含まれている場合がある．その場合，SVM による行動タイプ識別では，1 つの文に対して 1 つの行動タイプを識別するため，1 つの行動からなる文に分割する必要がある．例えば，GPSR では，「キッチンに行って，コーヒーを探して，それを持ってきて」といった命令がなされる可能性がある．SVM により行動タイプを識別するために，この命令文を「キッチンに行って」，「コーヒーを探して」，「それを持ってきて」といった 1 つの行動が含まれた文へと分割する．本稿では，各文の終端に “end-of-sentence” タグを出力するように学習された CRF により文の分割を行う．CRF は，系列ラベリングに用いられる対数線形モデルであり，3.3 節で説明するスロット抽出と同様の素性を用いて学習した．以上の処理により，文字列 r_n は J_n 個の文字列 r_{nj} $(j = 1, 2, \cdots, J_n)$ に分割される．

行動タイプ識別，スロット抽出は分割された各文に独立して行い，最終的な命令理解結果として，次式を最大化する行動タイプ $a_1, a_2, \cdots, a_{J_n}$ とスロット $s_1, s_2, \cdots, s_{J_n}$ を選択する．

$$\frac{1}{J_n} \sum_j S_{total}(a_j, s_j, r_{nj}|u, e) \tag{22}$$

4. 評価実験

提案モデルが GPSR における命令理解において有効か検証した．RoboCup@Home の環境に近い状況で命令理解を行うために，シミュレータ SIGVerse[*2] によって再現された 2LDK の空間で実験を行った．

4.1 データセット

本稿では実際に 2011 年の GPSR タスクで使用された命令文を自動生成するソフトウェア [*3] を使用し，命令文を収集した．さらに，より様々な表現の命令文を収集するため，RoboCup@Home に参加したことのある学生にアンケートを実施し，命令文を収集した．以下が収集した命令文の一例である．

- テーブルにあるジュースを掴んで，キッチンにいる田中さんに挨拶して，ジュースを渡して
- ジュースを探してきて
- ジュースはどこにあるか見つけて

このような命令文を 1569 文収集した．

4.2 音声命令理解実験

評価に用いる命令文は 4.1 節で説明したデータの中からランダムに 155 文を選択した．この選択した命令文を，研究室に所属する 20 代の男性 2 人がノイズがない静かな環境で読み上げ，その音声で録音をした．ノイズによる誤認識の影響を

*2 SIGVerse, http://www.sigverse.org/wiki/jp/.
*3 Sentence Generator 2011 for the General Purpose Service Robots, http://komeisugiura.jp/software/software_jp.html.

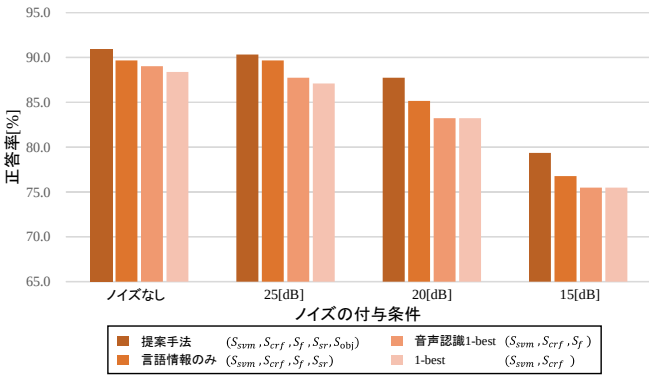

図 2: 各ノイズ条件における命令理解精度

比較するために，この音声に対してホワイトガウスノイズを SN 比 25[dB]，20[dB]，15[dB] の条件で付与した．提案モデルの学習に用いたデータは，残りの 1414 文である．それぞれの命令文に含まれる命令数の内訳は表 1 のとおりである．また，この実験において式 (20) の N，M，L はそれぞれ 10 とした．提案モデルの各実装には，音声認識には Julius[*4] を，形態素解析器には Mecab [Kudo 05] を，行動タイプ識別には LibSVM [Chang 11] を，スロット抽出には CRF++[*5]，重み推定には Liblinear [Fan 08] をそれぞれ使用した．また，シミュレーション環境内のテーブルや棚等に 10 種類の物体を配置した．ロボットは部屋の中をあらかじめ物体認識をしながら移動することで，物体存在確率を計算した．比較対象として，以下の 3 つの手法と比較した．

- 式 (17) から物体存在確率を除いた手法（言語情報のみ）
- 物体存在確率を除き，音声の 1-best のみを利用した手法 (音声認識 1-best)
- 物体存在確率を除き，音声認識・SVM・CRF の 1-best のみを利用した手法 (1-best)

音声命令理解の評価では，行動タイプ識別結果とスロット抽出結果が正解のアノテーションと完全に一致した場合のみを正解とした．

図 2 が音声命令理解の結果であり，155 文中の正答率を示している．この図より，提案モデルがどのノイズ条件においても，正答率が高いことが分かる．さらに提案モデルでは，ノイズがより大きい条件において，より正答率の改善が大きいことが分かる．これは，音声の認識が困難な条件においても，提案モデルにより環境の情報を利用することで命令を正しく理解できていることを示している．例えば，「ディナーテーブルへ移動し，バナナを見つけ，それを持って」といった命令は，他の手法では「バナナ」を「田中」と誤認識しているのに対して，提案手法では，バナナはディナーテーブルに存在するといった情報を利用することで，正しく認識することができている．以上のように，提案モデルが GPSR のような実環境での音声命令理解において，有効であることが示された．

5. まとめ

本稿では，音声認識，SVM による行動タイプの識別，CRF によるスロット抽出，行動タイプとスロットの共起関係，環境情報をベイジアンネットワークを用いて統合した．提案モデルは，複数の情報の依存関係を考慮することで，GPSR における命令理解において，より高い精度で認識することができた．また，環境情報を利用することで，音声だけでは命令理解が困難な条件においても，理解精度が向上することが示された．

提案モデルは，グラフィカルモデルに基づいており，確率分布を定義することで新たな情報を容易に導入することができる．そこで，今後物体情報以外の環境情報をこのモデルに導入することを考えている．また，今回評価に用いたホワイトガウスノイズは実際の家庭環境におけるノイズとは異なるため，より家庭環境に近いリアルなノイズがある環境で実験を行うことを考えている．

*4 汎用大語彙連続音声認識エンジン Julius, `http://julius.osdn.jp`.

*5 CRF++:Yet Another CRF toolkit, `https://taku910.github.io/crfpp`

参考文献

[Chang 11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, p. 27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[Chen 15] Chen, X., Shuai, W., Liu, J., Liu, S., Wang, N., Lu, D., Chen, Y., and Tang, K.: KeJia: The Intelligent Domestic Robot for RoboCup@ Home 2015 (2015), Teamdescription papers: RoboCup@Home League

[Fan 08] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874 (2008)

[Kudo 05] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer, *http://mecab. sourceforge. net/* (2005)

[Murase 00] Murase, H. and Vasudevan, V. V.: Fast visual search using focused color matching−active search, *Systems and Computers in Japan*, Vol. 31, No. 9, pp. 81–88 (2000)

[Roy 05] Roy, D. and Mukherjee, N.: Towards situated speech understanding: Visual context priming of language models, *Computer Speech & Language*, Vol. 19, No. 2, pp. 227–248 (2005)

[Schiffer 12] Schiffer, S., Hoppe, N., and Lakemeyer, G.: Natural language interpretation for an interactive service robot in domestic domains, in *Agents and Artificial Intelligence*, pp. 39–53, Springer (2012)

[Seib 15] Seib, V., Manthe, S., Holzmann, J., Memmesheimer, R., Peters, A., Bonse, M., Polster, F., Rezvan, B., Riewe, K., Roosen, M., et al.: RoboCup 2015-homer@ UniKoblenz (Germany) (2015), Teamdescription papers: RoboCup@Home League

[Xu 13] Xu, P. and Sarikaya, R.: Convolutional neural network based triangular crf for joint intent detection and slot filling, in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 78–83 (2013)

[吉川 15] 吉川敦史, 岩橋直人, 高渕健太, 中村友昭, 長井隆行, 船越孝太郎, 國島丈生：線形 SVM を用いた複数識別器出力の線形結合係数の最適化, 第 17 回 IEEE 広島学生シンポジウム (2015)