学術・技術論文

言語獲得ロボットによる発話理解確率の推定に基づく物体操作対話

杉浦孔明* 岩橋直人* 柏岡秀紀* 中村哲*

Object Manipulation Dialogue by Estimating Utterance Understanding Probability in a Robot Language Acquisition Framework

Komei Sugiura*, Naoto Iwahashi*, Hideki Kashioka* *and* Satoshi Nakamura*

This paper proposes a method that generates motions and utterances in an object manipulation dialogue task. The proposed method integrates belief modules for speech, vision, and motions into a probabilistic framework so that a user's utterances can be understood based on multimodal information. Responses to the utterances are optimized based on an integrated confidence measure function for the integrated belief modules. Bayesian logistic regression is used for the learning of the confidence measure function. The experimental results revealed that the proposed method reduced the failure rate from 12% down to 2.6% while the rejection rate was less than 24%.

Key Words: Object Manipulation Dialogue, Confidence, Robot Language Acquisition, Bayesian Logistic Regression

1. はじめに

高齢化社会の到来とともに，生活環境で人間を支援するロボットへの期待が高まっている．生活支援ロボットにとって，ユーザの命令を理解するコミュニケーション機能は極めて重要であるが，ロボットの対話処理機構は必要なレベルにまったく到達していない [1]．例えば「コップ持ってきて」という命令を聞いて，食器棚に向かうのか，目の前のテーブルに手をのばすのか，など適切な行動を選択することは，現状の対話処理技術にとって難しい問題である．家の中には候補となる「コップ」が多く存在し，食事の準備か片付けのためかなどによって，ユーザに渡すべき対象は異なるためである．

現状のロボット対話処理機構では，動作コマンドの伝達を目的とすることが多いにもかかわらず，動作情報と音声認識は別々に処理されている（例えば文献 [2]）．ユーザの発話の意味はグラウンドされない知識に基づいて解釈されるため，動作が状況にふさわしいかどうかは音声認識時には考慮されない．しかしながらこのような手法では，ユーザの発話の意味が状況に応じて適切に理解されないため，ロボットが予期しない動作を行ってしまう危険性がある．

本研究では，この危険性を減少させることを目的とする．具体的なタスクとして，**物体操作対話タスク**を扱う．物体操作対話タスクとは，ユーザが発話によりロボットにオブジェクトを操作させるタスクを指す．物体操作対話タスクにおいて，ユーザの発話の意味が適切に理解されるためには，(1) 言語によるオブジェクト参照，(2) 言語による動作参照，における曖昧性を解消する必要がある．例えば，上述の「コップ持ってきて」という発話では，候補の中からカメラ画像やコンテキストからオブジェクトを推定したうえで，その状況において「持ってくる」という動作を表す関節角・オブジェクト位置の軌道を生成する必要がある．

(1) の曖昧性解消に対しては人工知能や自然言語生成の分野で多くの研究が行われてきた [3] [4]．これらの研究では，仮想的なオブジェクトを用いて，オブジェクトを指示する言語表現の生成が試みられている．これに対し，(2) の曖昧性解消と関連が深い研究としては，動作の言語化を目指す試みが近年注目されてきている [5]〜[7] ．高野らは，運動の分節化を通じてヒューマノイドロボットが獲得した原始シンボルを用いて，運動認識・生成を行っている [7]．Ogata らは，動作系列と記号列の間の多対多対応問題を扱うリカレントニューラルネットに基づく手法を提案している [6]．一方，我々は (1) (2) の曖昧性を解消するアプローチとして，言語獲得フレームワーク LCore を提案している [8] [9].

本論文では，LCore を拡張し，発話理解確率を推定することにより動作・発話生成を行う手法を提案する．提案手法の独自性は，以下の 2 点である．

(1) マルチモーダル入力に基づく発話理解確率（統合確信度）の

原稿受付 2009 年 12 月 20 日
*情報通信研究機構
*National Institute of Information and Communications Technology
■ 本論文は学術性で評価されました.

推定問題に対して，ベイズロジスティック回帰（BLR）[10] を用いる．

(2) 統合確信度を用いて過不足ない自然な確認発話を生成する．LCore におけるユーザとロボットのインタラクションは，発話から動作へつながる一方向的過程であった [11]．これに対し提案手法では，ユーザの発話が曖昧であれば確認発話を生成し，動作の実行前にユーザに許可を求めることが可能である．BLR を用いた発話理解確率の利点は，(1) 成功確率を事後確率として推定可能であること，(2) 少数サンプルで推定が可能であること，が挙げられる [12]．

本稿では，まず本研究で扱うタスク環境について 2 章で述べ，3 章では提案手法の基盤となる LCore について概説する．4 章において，提案手法による統合確信度の学習手法と，統合確信度に基づく発話生成手法について詳細を述べる．手法評価のために行った実験の方法について 5 章で述べ，6 章に実験結果を示したうえで考察を加える．7 章では，物体操作対話に限定せず，より広い視点から関連研究について述べたあと，8 章で本稿のまとめを行う．

2. タスク環境

2.1 物体操作対話タスク

本研究では，ロボットがユーザの発話に従って，テーブル上のオブジェクトを操作するような状況を想定する．ユーザは **Fig. 1** のようにロボットと向かい合って座るものとする．実験に用いるオブジェクトを **Fig. 2** に示す．

オブジェクトの名前，動作，文法などは言語獲得手法 LCore [8] [9] を用いて，あらかじめ学習されているものとする．LCore では，オブジェクトやその動かし方に関する非言語知識と，単語や文法（単語や節の並び）など言語知識がモデル化されている．このモデルのパラメータは，ユーザとのインタラクションから学習サンプルを得ることで推定される．つまり，LCore を用いることにより，言語知識と非言語知識に関するモデルが得られる．言い換えると，「赤い」や「箱」という単語がどのような音韻列で構成されるか，またそれらがどのような視覚的特徴を表現するか，に関しての知識は設計者により事前に与えられたものではない．加えて，「上げる」「近づける」などロボットに行わせる動作も設計者が与えたものではない．設計者があらかじめ用意した主な機能は，画像からのオブジェクト抽出および画像特徴量抽出である．LCore の詳細については 3 章で説明する．

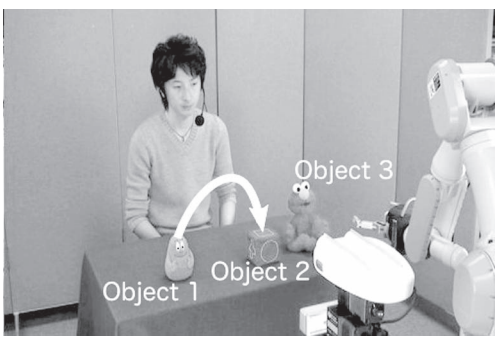


Fig. 1　An example of object manipulation dialogue tasks

Fig. 2　Objects used in experiments

タスクは以下のように進行する．まず，ユーザはオブジェクトを操作するよう音声によりロボットに指示を与える．両者は音声対話により曖昧性を解消し，ロボットがユーザの意図した行動をとればタスク成功とする．例えば，Fig. 1 に示すシーンにおいて，ユーザが「バーバブライト，赤い箱のせて」と発話したとする．このとき，ユーザが意図した行動は，オブジェクト 1（バーバブライト）をオブジェクト 2（赤い箱）にのせる行動である．ロボットはユーザ発話を受けて，必要であれば「小さい赤い箱にのせていいですか？」などの確認発話を行う．最終的に，ロボットが図中で白線で示されるような軌道でオブジェクト 1 を操作すればタスク成功である．

ロボットはユーザ発話を受けて即座に行動してもよいが，その場合はユーザの意図しない行動を行う（行動失敗）というリスクを負うことになる．仮に音声認識誤りによりユーザ発話中の「箱」が「エルモ」と誤認識された場合，「バーバブライト（オブジェクト 1）を赤いエルモ（オブジェクト 3）にのせる」行動が生成されることになる．Fig. 2 に示すようなぬいぐるみが行動失敗により破損することは稀であるが，食器等のオブジェクトを用いる場合はユーザの意図しない行動を生成することは危険である．

本タスクでは，できるだけ少ないターン数の対話を通じて，ユーザの意図する行動を出力できることが望ましい．いま，Fig. 1 の状況において，ユーザが「バーバブライトのせて」と発話したとする．このとき，「上面が平らでないオブジェクトには他のオブジェクトがのせられにくい」という知識があれば，オブジェクト 1 をオブジェクト 3 の上にのせるような行動は起こりにくいと推論できる．この推論に，オブジェクト 1 が「バーバブライト」としてもっともらしいという知識を総合すれば，前述の発話ではオブジェクト 2 を表す間接目的語が省略されたと推論できる．したがって，「青い箱にのせていいですか？」といった確認発話を生成すれば，ユーザの許可を得られる可能性が高い．また，直前の行動でオブジェクト 1 が動かされたあとにユーザが「のせて」と発話したのであれば，直接目的語・間接目的語の両方を推定する必要がある．

ユーザの指示発話は，「大きい赤い箱まわして」のように動詞 1 個と，名詞／形容詞を 0 個以上含むものとする．また，ロボットの生成する発話には機能語（「を」や「に」などの助詞）が含まれるが，ユーザ発話には機能語が含まれないものとする．これは，機能語の音韻と機能語を含む文法が未学習であるため，ユーザ発話に含まれる機能語の音声認識ができないことによる．

ただし，3 章で述べる手法により，発話から直接目的語と間接目的語を含む節が得られるので，これらの節にそれぞれ助詞「を」と「に」を付加することで，よりユーザに分かりやすい確認発話を生成する．また，「エルモの隣にある箱」のように他のオブジェクトとの相対的な位置関係を表すような表現は用いないこととする．

本研究の主旨は，言語と実世界事物の多対多対応という性質に焦点を当てて，対話により行動失敗リスクを低減することである．多対多対応の性質を扱うため，音声認識が完全であったとしても，実世界のオブジェクトが一意に指定できないことがある．例えば，Fig. 1 のオブジェクト 2 を「箱」，「赤いもの」，「小さい赤い箱」などと表現することが可能であるのと同時に，「赤い」という表現は通常オブジェクト 2 およびオブジェクト 3 の両方に当てはまる†．また，動作に関しても，Fig. 1 に示す軌道を「のせる」「置く」など複数のラベルで表現することが可能であるのと同時に，Fig. 1 に示す軌道がやや変形していても「のせる」という表現を用いることができる．

以上のような問題意識から，曖昧性の解消には指さしやグラフィカルユーザインタフェース（GUI）などは用いず音声のみを用いるものとする．テキストベース対話に対する音声対話の利点は，キーボードやディスプレイ等を必要としないハンズフリーなインタラクションが可能な点である．例えば，キーボード等でテキスト入力するようなシステムでは，ユーザが両手にオブジェクトを持っているような状況下で指示を与えることができない．

2.2 ロボットシステム

実験に用いたロボットシステムを **Fig. 3** に示す．ロボットシステムは，7 自由度のロボットアーム（三菱重工製 PA-10），4 自由度のロボットハンド（Barrett Technology 製 BarrettHand），マイクロフォン，ステレオカメラ（Point Grey Research 製 Bumblebee 2），視線表出ユニット（Directed Perception 製 PTU-46-70 にロボットヘッドを取付）からなる．視線表出ユニットは，ユーザまたはオブジェクトにロボットの視線を向けることで，簡単な内部状態（オブジェクトトラッキング中，ユーザ発話受け入れ可能，など）を表出する．ユーザは，ハンドに取り付けられた触覚センサを叩くことで教示信号を与えることができる．

音声データは 16 [kHz]，16 [bit] でディジタル化され，各フレームごとに，25 次元の特徴量ベクトルが計算される．特徴量として，メル周波数ケプストラム係数（MFCC）12 次元，Δ MFCC12 次元，対数パワー一次元を用いた．フレーム長は 20 [ms]，シフト長は 10 [ms] とした．

オブジェクトの画像特徴および二次元座標（カメラ座標系）は，固定されたステレオカメラを用いて得る．なお，オブジェクトの抽出およびトラッキングは，色およびステレオカメラから得られる距離に基づくヒューリスティックな手法により行う．カメラのフレームレートは 30 [frame/s] であり，解像度は 320×240 である．画像特徴量として，色三次元（L*a*b*），形状三次元（オブジェクト領域 f_{area}，四角らしさ f_{sq}，縦横比 f_{whr}）を用いる．これらは，画像上の縦（鉛直方向）h，幅 w，ピクセル数 N_{obj} から，$f_{\mathrm{area}} = wh$，$f_{\mathrm{sq}} = N_{\mathrm{obj}}/wh$，$f_{\mathrm{whr}} = w/h$ と定義する．新規動作の学習や動作認識を行う際には，オブジェクトの座標時系列が利用される．

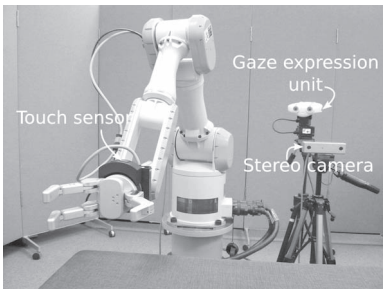


Fig. 3 Robotic platform used in the experiments

3. LCore における発話理解

LCore [8] では，マルチモーダル入力から学習されたユーザモデルを用いてユーザの発話を理解する．本論文では，音声・画像・動作などの各モダリティに対応するユーザモデルを**信念モジュール**と呼ぶ．また，(1) 音声，(2) 動作，(3) 視覚，(4) 動作-オブジェクト関係，(5) 行動コンテキスト，の五つの信念モジュールを統合したユーザモデルを**共有信念** Ψ と呼ぶ．

3.1 LCore における動作生成

LCore による動作生成では，文献 [13] で提案した参照点に依存した隠れマルコフモデル（HMM）に基づく手法を用いる．この手法では，物体操作軌道が**トラジェクタ**（動かされるオブジェクト）と**参照オブジェクト**との相対軌道としてモデル化される．参照オブジェクトとは，動作の基準となるオブジェクトのことを指し，トラジェクタそのもの，あるいは**ランドマーク**（トラジェクタの動きの基準となるオブジェクト）から選択される．いま Fig. 1 において，ユーザがロボットに対して，オブジェクト 1（バーバブライト）をオブジェクト 2（赤い箱）にのせるように指示したとする．この場合，トラジェクタはオブジェクト 1，参照オブジェクトはオブジェクト 2 である．

3.2 共有信念モデルに基づく発話理解

ユーザの発話 s は，以下の概念構造 z と対応づけて解釈される．

$$z = [(\alpha_1, W_{\alpha_1}), (\alpha_2, W_{\alpha_2}), (\alpha_3, W_{\alpha_3})]$$
$$\alpha_i \in \{T, L, M\} \qquad i = 1, 2, 3$$

ここに，α_i は文節の属性を表し，トラジェクタ（T），ランドマーク（L），動作（M）のいずれかをとることとする．よって，W_T, W_L, W_M は，それぞれトラジェクタを表す文節，ランドマークを表す文節，動作を表す文節を意味する．例えば，Fig. 1 に示すシーンにおいて，ユーザが「バーバブライト，赤い箱のせて」と発話したとする．このとき，正しく分割された文節は以下のようになる．

† カラーのカメラ画像では Fig. 1 中のオブジェクト 2 および 3 は赤色で，オブジェクト 1 は水色に近い青色である．LCore では，オブジェクトの属性を「色=青」のように設定するような離散表現を行っていない．実際には画像認識における外乱を前提として，オブジェクト 1 も「赤い」と表現される可能性を考慮するような知識表現がなされている．

[(T, [バーバブライト]), (L, [赤い，箱]), (M, [のせて])]

ただし本手法では，s に含まれる動詞の活用形はすべて命令形であり，音声認識時に助詞を扱わないこととする．また，ランドマークを必要としない動作概念では，$z = [(T, W_T), (M, W_M)]$ である．

文節および単語の順序の規則は，統計的言語モデル G でモデル化される．G は，属性 α_i の系列の出現頻度と，文節内での単語インデックス列の出現頻度からなる．

いま，シーン O において発話 s が与えられたとしよう．O は，カメラ画像中の全オブジェクトの画像特徴量および位置を表す．O において可能な行動の集合 A は以下により与えられる．

$$A = \{(i_t, i_r, C_V^{\langle j \rangle}) \,|\, i_t = 1, ..., O_N, i_r = 1, ..., R_N, j = 1, ..., V_N\} \triangleq \{a_k \,|\, k = 1, 2, ..., |A|\}, \quad (1)$$

ここに，トラジェクタのインデックスを i_t，参照オブジェクトのインデックスを i_r，O 中のオブジェクトの数を O_N，動作を表す単語数を V_N，j 番目の動作モデル $C_V^{\langle j \rangle}$ に対して可能な参照オブジェクトの数を R_N とする．したがって，物体操作対話タスクでは，s に対し正しい行動 a_k を選択することが求められる．以下，本稿では a_k のように動作インデックスと関連オブジェクトインデックスで表される離散的な表現（記号レベル）を「行動」と称し，軌道レベルの物体操作を表すために「動作」という表現を用いる．

各信念モジュールを以下のように定義する．

- 音声信念 B_S
  B_S は，文法 G の下での，発話 s に対する z の条件付き確率の対数として表す．
- 視覚信念 B_I
  B_I は，オブジェクト i の視覚特徴量 $\mathbf{x}_I^{\langle i \rangle}$ に対する確率モデル（ガウス分布）の対数尤度である．
- 動作信念 B_M
  B_M は，トラジェクタ i_t の位置 $\mathbf{x}_p^{\langle i_t \rangle}$ が与えられたうえでの $\hat{\mathcal{Y}}_k$ に対する動作モデルの対数尤度で表される．この動作モデルには参照点に依存した HMM により表現される．ここに，$\hat{\mathcal{Y}}_k$ は a_k に対する最尤軌道を表す．動作モデル λ は，$\mathbf{x}_p^{\langle i_t \rangle}$ および，ランドマークの位置 $\mathbf{x}_p^{\langle i_r \rangle}$，動作インデックス $C_V^{\langle j \rangle}$ から文献 [13] の手法により得られる．
- 動作–オブジェクト関係信念 B_R
  B_R は，オブジェクト (i, j) の視覚特徴量に対する確率モデル（ガウス分布）の対数尤度である．
- 行動コンテキスト信念 B_H
  B_H は，コンテキスト $\mathbf{q}^{\langle i \rangle} = (q_1^{\langle i \rangle}, q_2^{\langle i \rangle})$ のもとでの，指示対象としてのオブジェクト i の適切さ（スコア）を表す．ここに，$q_1^{\langle i \rangle}, q_2^{\langle i \rangle}$ をそれぞれ，オブジェクト i が「把持されている」，「直前に操作された」状態を表す真偽値である．B_H は以下で定義される．

$$B_H(i, \mathbf{q}^{\langle i \rangle}; h_c) = \begin{cases} 10 & (q_1^{\langle i \rangle} = 1) \\ h_c & (\mathbf{q}^{\langle i \rangle} = (0, 1)) \\ 0 & (\mathbf{q}^{\langle i \rangle} = (0, 0)) \end{cases} \quad (2)$$

B_H により，指示語や日本語に多い目的語の省略をモデル化できる．B_H のパラメータ h_c（非負の実数）は，Minimum Classification Error（MCE）学習 [14] に基づいて学習される．

以上より，共有信念関数 Ψ を，各信念モジュールの重み付き和として定義する．

$$\begin{aligned} \Psi(s, a_k, O, \mathbf{q}^{\langle i_t \rangle}) = \\ \max_z \Big\{ & \gamma_1 \log P(s|z) P(z; G) && [B_S] \\ & + \gamma_2 \left(\log P(\mathbf{x}_I^{\langle i_t \rangle} | W_T) + \log P(\mathbf{x}_I^{\langle i_r \rangle} | W_L) \right) && [B_I] \\ & + \gamma_3 \log P(\hat{\mathcal{Y}}_k | \mathbf{x}_p^{\langle i_t \rangle}, \mathbf{x}_p^{\langle i_r \rangle}, C_V^{\langle j \rangle}) && [B_M] \\ & + \gamma_4 \log P(\mathbf{x}_I^{\langle i_t \rangle}, \mathbf{x}_I^{\langle i_r \rangle} | C_V^{\langle j \rangle}) && [B_R] \\ & + \gamma_5 \left(B_H(i_t, \mathbf{q}^{\langle i_t \rangle}) + B_H(i_r, \mathbf{q}^{\langle i_r \rangle}) \right) \Big\}, && [B_H] \end{aligned} \quad (3)$$

ここに，$\mathbf{x}_p^{\langle i \rangle}$ はオブジェクト i の位置，$\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_5)$ は，各信念に対する重み（非負の実数）を表す．$\boldsymbol{\gamma}$ の学習には，MCE 学習を用いる．Ψ により，発話 s と行動 a_k の対応の適切さを評価することができる．

4. 発話理解確信度の推定に基づく応答生成

4.1 統合確信度による発話理解確率のモデル化

前章の共有信念関数を用いると，コンテキスト q，シーン O，発話 s が与えられたときの最適行動 $\hat{a}$ は以下で得られる．

$$\hat{a} = \underset{k}{\operatorname{argmax}} \Psi(s, a_k, O, \mathbf{q}) \quad (4)$$

行動 a_j と，最適行動 $\hat{a} (k \neq j)$ のマージンを以下の関数 d により定義する．

$$d(s, \hat{a}, O, \mathbf{q}) = \Psi(s, \hat{a}, O, \mathbf{q}) - \max_{j \neq k} \Psi(s, a_j, O, \mathbf{q}) \quad (5)$$

いま，最大値の次に大きい値を与える行動を a_l とする．式 (5) より，最適行動 $\hat{a}$ に対するマージンは $\hat{a}$ と a_l の共有信念関数の値の差であることが分かる．よって，$\hat{a}$ に対するマージンが 0 に近ければ，発話 s は $\hat{a}$ と a_l を指示する発話として同程度に適した表現であるといえる．逆に，マージンが大きい場合には，$\hat{a}$ のほうが s の指示する行動として適している．したがってマージン関数は，行動 $\hat{a}$ を指示する発話としての s の曖昧性の尺度として用いることができる．

ここで，マージンを用いて $\hat{a}$ に対する確信度を得ることを考える．音声対話システムでは，認識結果に対する確信度を導入することにより，発話を棄却するか否かを制御する研究が行われている [15]．なお，マージンは第一候補と第二候補の共有信念関数値の差のみから得られるが，第三，第四，.. の候補について指標に含める方法も考えられる．しかし，式 (3) におけるパラメータ γ の学習において，マージンの最大化を基準として MCE 学習が行われているため，第一候補と第二候補を扱うことが合理的である．

提案手法では，**統合確信度**関数 $f(d)$ をロジスティックシグモイド関数を用いて以下のように定義する．

$$f(d;\mathbf{w}) = \frac{1}{1+\exp(-(w_1 d + w_0))} \quad (6)$$

ここに，パラメータ $\mathbf{w} = (w_0, w_1)$ である．この $f(d)$ により，d のもとで発話が正しく理解される確率をモデル化する．

4.2 統合確信度関数の学習

マージンと正解ラベルを学習サンプルとして，ロジスティック回帰により $f(d;\mathbf{w})$ のパラメータ $\mathbf{w}$ を推定することを考える．学習サンプル集合を入力 d_i と教師信号 u_i の組として以下のように与える．

$$\mathbb{T}^{\langle N\rangle} = \{(d_i, u_i) | i = 1, ..., N\}, \quad (7)$$

ただし，u_i は 0（不正解）または 1（正解）の 2 値であるとする．実験では，u_i は音声（「はい」「いいえ」など），または触覚センサなどを通じて入力される．

いま，入力 d_i を与えたときの出力 $f(d_i)$ を，入力 d_i のもとで教師信号 u_i が 1 である確率の推定値であると考える．本手法では，BLR [10] を用いて $\mathbf{w}$ の推定を行う．以下のように，w_j $(j = 0, 1)$ の事前分布として，平均 m_j，分散 τ_j のガウス分布を用いる．

$$P(w_j | m_j, \tau_j) = \mathcal{N}(m_j, \tau_j) = \frac{1}{\sqrt{2\pi\tau_j}} \exp \frac{-(w_j - m_j)^2}{2\tau_j}$$

4.3 期待効用最大化に基づく応答最適化

ユーザの発話 s に対してロボットが行った行動 $\hat{a}$ が，ユーザがロボットに行わせたい行動 a^* と異なることは，安全性の観点から望ましくない．このような危険を回避するための手法の一つは，ユーザ発話が曖昧な場合に確認発話を行うことである．このとき，どれほど曖昧性が含まれる場合に確認発話を行うかという意志決定問題を解く必要がある．

この意志決定問題に対し，統合確信度を用いることを考える．この場合，発話 s に対する最適行動 $\hat{a}$ の統合確信度が小さければ，ユーザに $\hat{a}$ を行うか否かを確認する発話をすればよい．別の手法として，マージンに閾値を設定し，確認発話を行うかどうかを決定することも考えられる．この場合，各ユーザに対する γ が異なるため，ユーザごとにマージン閾値を変更する必要がある．特定のユーザに対するマージン閾値が別のユーザに対して有効であるとは限らないので，設計上の負担は大きい．これに対し統合確信度を用いる場合は，統合確信度に対する閾値を別のユーザに適用することができる．これは，各ユーザに対してマージンと確率の写像が学習されているためである．以下では，統合確信度を用いて最適な意志決定を行わせる手法について述べる．

いま応答として，動作応答 b_1 と確認発話応答 b_2 があるとする．統合確信度 $f(d)$ は，マージン d のもとで発話が正しく理解される確率をモデル化するものであった．このとき応答 $b_i (i = 1, 2)$ に対する効用 R_i の期待値 $\mathbb{E}[R_i]$ を以下のように推定することができる．

$$\mathbb{E}[R_i] = r_{i1} f(d) + r_{i2}(1 - f(d)) \quad (8)$$

[†]音声認識結果そのものを聞き返しても曖昧性解消効果は低い．

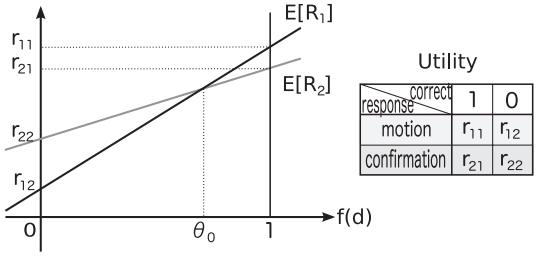


Fig. 4 The relationship between the ICM value and the expected utility

ただし，r_{i1}, r_{i2} はそれぞれ，$\hat{a} = a^*$，$\hat{a} \neq a^*$ のときの応答 b_i に対する効用である．

ここで，$r_{12} < r_{22} < r_{21} < r_{11}$ であるとする．この大小関係は，「動作 $\hat{a}$ を実行して間違った場合の効用が最も低く，動作 $\hat{a}$ を実行して成功した場合の効用が最も高い」ということを表す．$f(d)$ を確率表現することにより，$\mathbb{E}[R_i]$ は $f(d)$ の線形関数になる．このとき，等式 $\mathbb{E}[R_1] = \mathbb{E}[R_2]$ は，$0 < \theta_0 < 1$ なる解 θ_0 を持つ．つまり，θ_0 を閾値として最適応答のインデックス $\hat{i} = \mathrm{argmax}_i \mathbb{E}[R_i]$ が選択できる．

次に，b_2 が最適応答である場合，共有信念として学習されたユーザモデルを言語表現の生成に用いることを考える．例えば食器が複数ある状況では，「四角くて白い皿」のように最も曖昧性が減少し，かつ冗長でない表現でオブジェクトを表現できることが望ましい[†]．提案手法では，ユーザの発話に対しマージンを最大化する単語を加えることで曖昧性を減少させる．

ここで，音響モデルの尤度を含まない共有信念を $\psi(s, a_k, O, \mathbf{q}^{\langle it\rangle}, z)$ で表すことにする．ψ と Ψ の違いは，ψ には音響モデルの尤度が含まないこと，z に対して最大化されていないこと，の 2 点である．このとき，z が与えられたうえでのマージン d_z を以下のように定義する．

$$d_z(s, a_j, O, \mathbf{q}, z) = \psi(s, a_j, O, \mathbf{q}, z) - \max_{k \neq j} \psi(s, a_k, O, \mathbf{q}, z)$$

挿入単語集合 $\mathbf{c}' = \{c'_m \,|\, m = 1, ..., M\}$ が，文節 W（W_T または W_L）に挿入されるとしよう．ここで，W は長さ $|W|$ の単語列 $c_1 c_2 \cdots c_{|W|}$ であるとする．このとき，最適挿入単語集合 $\hat{\mathbf{c}}' = \{\hat{c}'_m \,|\, m = 1, ..., M\}$ と最適挿入位置集合 $\hat{\mathbf{p}} = \{\hat{p}_m \,|\, m = 1, ..., M\}$ は以下で与えられる．

$$(\hat{\mathbf{c}}', \hat{\mathbf{p}}) = \underset{c'_m \notin W, \mathbf{p}}{\mathrm{argmax}}\, d_z(s, a_j, O, \mathbf{q}, z) \quad (9)$$

よって挿入後の文節 W' は以下のようになる．

$$W' = c_1 \cdots c_{\hat{p}_1 - 1} \hat{c}'_1 c_{\hat{p}_1} \cdots c_{\hat{p}_2 - 1} \hat{c}'_2 c_{\hat{p}_2} \cdots c_{|W|} \quad (10)$$

この操作を W_T，W_L に対して行い，最終的に以下の概念構造 z' を得る．

$$z' = [(T, W'_T), (L, W'_L), (M, W_M)]. \quad (11)$$

以上をまとめて，提案手法のアルゴリズムを示す．

Input $\langle O, \mathbf{q}, s\rangle$ をシーン O，コンテキスト $\mathbf{q}$，発話 s からなる入力集合とする．

1. 行動候補集合 A（式（1）参照）のすべての要素について実行予定軌道を生成し，$\Psi(s, a_k, O, \mathbf{q})$ を求める．
2. 最適行動 $\hat{a}$（式（4）参照）に対し，$f(d(s, \hat{a}, O, \mathbf{q})) \geq \theta_0$ ならば $\hat{a}$ を動作応答として終了．$f(d(s, \hat{a}, O, \mathbf{q})) < \theta_0$ ならば **3** へ．
3. 確認発話がターゲットとする行動集合 A' を $A' = A$ で初期化．
4. 単語挿入数 M を $M = 0$ と初期化．ターゲット行動を $a_j = \underset{a_j \in A'}{\operatorname{argmax}} f(d(s, a_j, O, \mathbf{q}))$ とする．
5. $M \leftarrow M + 1$ とし，式（11）に従って z' を生成する．
6. 更新されたマージン d' について $f(d') \geq \theta_0$ ならば **7** へ．$f(d') < \theta_0$ ならば **6**（**a**）へ．
6（**a**） z' に追加可能な単語が存在すれば **5** へ．存在しなければ **9** へ．
7. a_j について確認発話を行う．z' をもとに音声を合成して発話を行う．ただし，W'_T または W'_L が元の W_T または W_L と等しければ発話に含めない．
8. ユーザの応答が肯定であれば，a_j を動作応答として終了．否定であれば，A' から a_j を除いて **8**（**a**）へ．
8（**a**） A' が空集合であれば **9** へ．それ以外は **4** へ．
9. 発話 s を棄却して終了.「わかりません」という発話を出力する．

5. 実 験 設 定

提案手法の評価のために，（1）統合確信度関数の学習，（2）確信度に基づく意志決定，の 2 種類の実験を行う．あらかじめ, [13] [16] で提案した手法により，23 単語（名詞 8 語，形容詞 8 語，動詞 7 語）を学習させた．**Table 1** に [16] の手法で学習させた γ の値を示す．統合確信度学習中の新規語彙の追加は本論文の主旨ではないので，新規語彙の追加は行わせないこととした．

生活支援ロボットへの物体操作対話の導入を考えると，少ないインタラクションから学習可能な手法が望ましい．特に実ロボットとの対話では，ハードウェア故障やオブジェクトを破損させるリスクがあるため，大量のインタラクションを想定することは難しい．そのため実験（1）では，サンプル数が少ない場合のテストセット尤度により提案手法の有効性を検証する．提案手法を比較検討するためのベースラインとして，最尤推定によるロジスティック回帰手法（以下 MLE と略記）を選択した．BLR と MLE の入力はマージンとし，出力は（発話理解）確率とする．サンプルを増加させた場合の両者の性能が同程度になることは理論的に明らかである [17] が，実際には MLE より BLR が物体操作対話タスクに適することを示す．

MLE では，統合確信度関数のパラメータ $\mathbf{w}$ の学習に，Fisher のスコアリングアルゴリズム [18] を用いる．これは，トレーニングセット尤度を最大化するパラメータ推定法である．パラメータの最尤推定値は，Fisher 情報行列を用いた繰り返しアルゴリズムにより求められる．また，Fisher のスコアリングアルゴリズムにおけるパラメータ更新回数を 20 とした．

Table 1 Parameters of the shared belief function used in the experiment

Weight	γ_1	γ_2	γ_3	γ_4	γ_5
Value	1.00(fixed)	0.75	1.03	0.56	1.88

実験（1）において，トレーニングおよびテストデータは以下のように収集した．共有信念関数の学習と同様の実験環境で，被験者にロボットにオブジェクトを操作させるための発話を行わせ，カメラ画像と音声を 100 セット収録した．得られた画像・音声セットに，ユーザが意図した行動を正解としてラベル付けした．本手法の特徴である，動作や画像等の情報を発話理解に利用する点について性能を評価するために，ユーザにはできるだけ省略を用いるような発話を行わせた．収録データのチャンスレベルは平均 2.34%であり，収録した音声に含まれる単語数は平均 2.54 語であった．収録データのうち半数の 50 個をトレーニングセット，残りの 50 個をテストセットとした．

統合確信度関数のパラメータ $\mathbf{w}$ に関するハイパーパラメータの設定においては，平均 m_0，m_1 が標準ロジスティックシグモイド関数の値，すなわち $m_0 = 0$，$m_1 = 1$ になるように設定した．分散に関しては，$\tau_0 = \tau_1 = 100$ と設定した．

実験（2）の目的は，提案手法による行動失敗率の減少について検証することである．実験（2）では，被験者と提案手法を実装したロボットを対話させる．本実験では，実験（1）のトレーニングセットを用いて学習された確信度関数のパラメータを固定して用いる．被験者とロボットの対話は以下のようにして行う．まず，テストセットからサンプルを一つ選択し，オブジェクト配置を再現する．次に対応する音声を入力し，提案手法により応答を生成させる．確認発話応答に対しては，被験者に肯定または否定の応答を行わせる．ロボットによる動作または発話棄却により終了する一連のインタラクションをエピソードと定義する．動作を行ったエピソードにおける，正解行動以外の行動が実行されたエピソードの割合を行動失敗率として評価する．

6. 実験結果および考察

6.1 結果（1）：統合確信度関数の学習

統合確信度関数の学習に関する定性的結果を **Fig. 5** に示す．図におけるそれぞれの曲線は，Fig. 5 はトレーニングサンプル数を変えたときの回帰結果である．(a)〜(d) は，それぞれトレーニングサンプル数 10, 20, 30, 50 のときの結果に対応する．

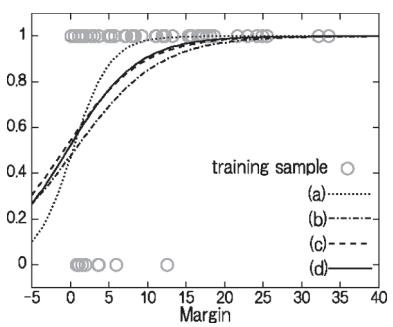


Fig. 5 The forms of $f(d; \mathbf{w})$ trained by (a) 10, (b) 20, (c) 30, and (d) 50 samples

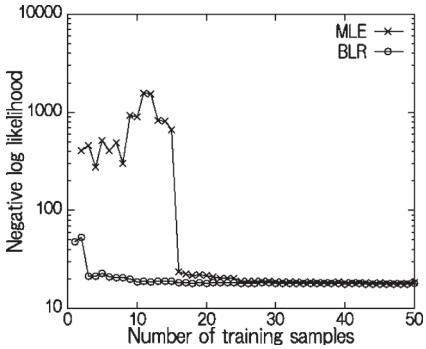


Fig. 6 Average negative test-set log likelihood of the ICM function

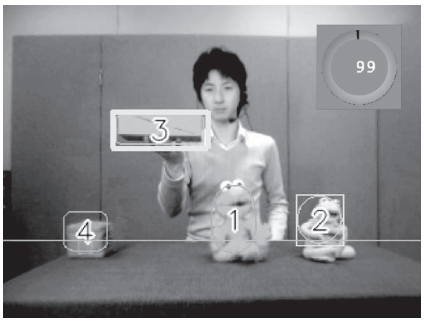
次に，定量的結果について検討する．**Fig. 6** に，ベースライン（MLE）と提案手法（BLR）の比較結果を示す．図には，テストセットが与えられたうえでの統合確信度関数の対数尤度 $\mathcal{L}$ をプロットした．ただし，両者を視覚的に比較しやすくするために，負の対数尤度を対数スケールで示してある．図に示した $\mathcal{L}$ の値は 10 回の実験における平均値であり，各実験は 100 個のサンプルを 50 個ずつトレーニングセットとテストセットにランダムに振り分けて行った．図から，サンプル数が 15 以下の場合に MLE と BLR のテストセット尤度の差が顕著であることが分かる．つまり，BLR の性能は MLE に比べて同等以上であり，特にサンプル数が少ない場合に BLR の性能が高いことが示唆される．物体操作対話タスクでは，一つの学習サンプルを得るコスト（ユーザを用いるコスト，オブジェクト配置を物理的に変更するコスト，ハードウェア故障リスク等）が大きいため，少ないインタラクションから学習できることは重要である．したがって，物体操作対話タスクには，サンプル数が少ない場合であっても高い性能を示す BLR がより適していると考えられる．

6.2 結果（2）：確信度に基づく意志決定

はじめに，定性的な結果について述べる．**Fig. 7**，**Fig. 8** にユーザ（U）とロボット（R）の対話例を示す．これらはともに $\theta_0 = 0.7$ と設定したときに得られたものである．各図において右上の数値は統合確信度を表す．

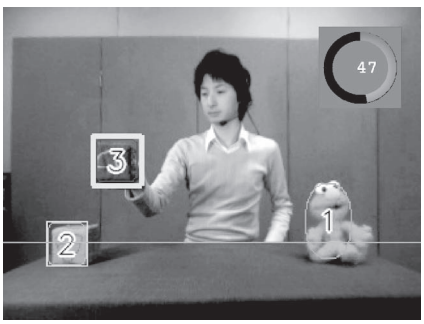
Fig. 7 では，ユーザはトラジェクタおよびランドマークについて発話しなかったものの，ロボットは正解行動（「オブジェクト 2 をオブジェクト 3 に載せる」）を出力した．この理由は以下のように考えられる．Fig. 7 に示すシーンでは，正解行動の軌道に対する尤度（動作信念モジュールから得られる尤度）は比較的小さい．具体的には，正解行動における軌道の尤度は，行動候補 60 通りのうち 24 位であった．しかしながら，動作-オブジェクト関係の信念モジュールと，コンテキスト信念モジュールから得られるスコアを加えることで，正解行動に対するスコアが行動候補のなかで最大になった．さらに，正解行動の確信度は $f(d) = 0.998 > \theta_0$ であったので，動作を実行する効用が確認発話をする効用を上回り，動作を実行する意志決定が行われた．

†カラー画像ではオブジェクト 2 は緑，オブジェクト 3 は青色である．



[Situation: Object 2 was manipulated most recently]
U: のせて． / *Place-on.*
R: (The robot places Object 2 on Object 3.)

Fig. 7 Dialogue example (1). Motion execution without a confirmation utterance. The correct action is to place Object 2 (Kermit) closer to Object 3 (large red box)



[Situation: Object 2 was manipulated most recently]
U: ハコ エルモ ちかづけて． / *Move-closer box Elmo.*
R: ミドリハコをちかづけて？ / *Move-closer green box?*
U: いいえ． / No.
R: アオイハコをちかづけて？ / *Move-closer blue box?*
U: はい． / Yes.
R: (The robot moves Object 3 closer to Object 1.)

Fig. 8 Dialogue example (2). Motion execution with a confirmation utterance. The correct action is to move Object 3 (blue box) closer to Object 1 (Elmo)

Fig. 8 では，最適行動の確信度は $f(d) = 0.478 < \theta_0$ であった．よって確認発話が最適応答であり，「アオイ ハコ」という言語表現が生成された．この言語表現は，オブジェクト 2 と 3 の視覚的特徴のなかで最も異なる属性†について述べており，ユーザにとって理解しやすい．ランドマークについては確認発話を行わなくても確信度に影響はないため，確認を省略していると考えられる．

次に，提案手法の処理時間について述べる．提案手法の質問応答は，ユーザが負担に感じない時間内で行われることが望ましい．$\theta_0 = 0.999$ の場合，質問文選択（5 節のアルゴリズムの 4 から 6）に要する時間は平均 1.73 秒であった．なお，実験に用いた計算機は，Dell Precision 490（2.0 [GHz] の Intel Xeon CPU E5335，メモリ 4 [GB]）であった．以上から，本実験設定上では処理時間は問題ない範囲であったといえる．大量の語彙を学習させるなど設定が複雑になった場合は，画像の尤度を

Table 2 Evaluation of decision-making based on the ICM value

θ_0	0 (speech-only)	0 (baseline)	0.7	0.999
P_f[%]	83.4	12.0	10.4	2.6
P_r[%]	0	0	4.0	24.0
P_c[%]	0	0	12.0	48.0
T_c	-	-	1.17	1.25

用いて不必要な語彙を除くことによって，探索を効率化する必要がある．

Table 2 に，確信度に基づく意志決定手法の定量的結果を示す．表において各項目は以下を表す．

$$\text{行動失敗率：} P_f = N_f/(N_s + N_f)$$
$$\text{棄却率：} P_r = N_r/N_a$$
$$\text{確認発話率：} P_c = N_c/N_a$$

ここに，N_a，N_s，N_f，N_c，N_r はそれぞれ，全エピソード数，正解の行動が行われたエピソード数，行動が失敗したエピソード数，確認発話がなされたエピソード数，発話が棄却されたエピソード数を表す．なお，$N_a = N_s + N_f + N_r$ である．また，T_c は平均確認発話数であり，確認発話が行われたエピソードにおける確認発話の平均回数を表す．

本研究は行動失敗の低減を目的としているため，行動失敗率を直接の指標とする．なお，把持失敗やオブジェクト同士の衝突などに関する失敗については考慮しない．棄却率はできるだけ低いほうが望ましいが，棄却による損失は行動失敗の損失に比べて無視可能なものと考え，補助的な指標とする．同様に，ユーザの負担の観点からは T_c が小さいほうがよいものの，行動失敗の損失に比べて無視可能なものと考え，補助的な指標とする．

Table 2 において，$\theta_0 = 0$（speech-only）は，動作や視覚などの情報を用いず，音声のみで発話理解を行った場合の結果である．音声認識には誤りが含まれないものとしたため，W_M には誤認識がない．一方，トラジェクタとランドマークはランダムに選択されるものとした．表より，音声のみを用いると行動失敗率が 80%以上になることから，たとえ音声認識が完全であったとしても，本タスクを音声のみで解くことは難しいといえる．

Table 2 において，$\theta_0 = 0$（baseline）の条件は，本手法を用いない場合の結果を意味する．これは従来の LCore の結果と等価であるので，これをベースラインとする．このとき，ユーザの発話に対して常に動作応答が行われる．ベースラインにおける行動失敗率は 12.0%（6/50）であった．

表より，本手法を用いる場合（$\theta_0 \neq 0$）には，行動失敗率が 12.0%より小さくなっている．例えば $\theta_0 = 0.999$ の場合には，行動失敗率は 2.6%（1/38）であった．このことから，提案手法は baseline に比べ行動失敗率を低減させることができたといえる．

次に，確認発話率について述べる．確認発話率は $\theta_0 = 0.7$ の場合に 12%であり，実用上問題ないと考えられる．$\theta_0 = 0.999$ では確認発話率は 48%と低くないが，すべてのエピソードで確認を行うような応答選択は行われていない．$\theta_0 = 0.999$ では，動作実行に対して確認発話を行う損失を相対的に小さく設定していると解釈できる．このような設定の下でも，曖昧性が少ない発話では確認発話なしで動作を実行するような，合理的な応答選択が行われていると考えられる．なお，Table 2 より平均確認発話数は 1.3 以下であった．

最後に棄却率について検討する．Table 2 より，θ_0 の増加に伴って行動失敗率が低下する一方，棄却率は上昇していることが分かる．ユーザの発話が棄却されるエピソードは，(1) θ_0 を超える効用を与える確認発話を生成できないと判断された場合と，(2) 確認発話に用いられた表現をユーザが理解できなかった場合に分けられる．(1) の例は，(学習させた) 言語表現のみではオブジェクトを同定できないシーンにおける発話が挙げられる．特に，本実験では「右」や「左」などの位置関係を表す語彙を用いないので，同じオブジェクトが二つあるシーンでは，片方のオブジェクトを同定する言語表現は存在しない．(2) の例は，画像処理における不確実性により，シーンに存在しないオブジェクトの名前を用いた言語表現を生成することが挙げられる．

6.3 考察

6.3.1 ヒューマンロボットインタラクション中の学習

提案手法では，ユーザとのインタラクションを通した学習が可能である．しかし，ユーザが発話してロボットが行動するようなインタラクション中に学習を行わせると，定量評価に多大なコストが発生する．これは，様々な初期設定下で結果を検討するためには，その都度今までのインタラクションに含まれていないようにオブジェクトを再配置しなければならないためである．これに対し，本実験のようにあらかじめデータベースを構築すれば，低コストで学習に必要なサンプル数などの定量評価が可能である．

6.3.2 θ_0 の設定

実験（2）では，設定すべきパラメータ数の削減のため，式（8）の r_{ij} を設定せずに直接 θ_0 を設定した．これは，本研究の主旨が正確に r_{ij} を設定することではないためである．ただし，原理的には r_{ij} を設定することにより θ_0 が求まる．

提案手法を実際の物体操作対話アプリケーションに適用する場合は，アプリケーションに応じて行動失敗や確認発話がもたらす効用（損失）を定量化し，r_{ij} を設定すべきである．例えば，行動成功の効用を 1,000，行動失敗の効用を 1 などと設定すればよい．4.3 節で述べたように，r_{ij} を設定すれば θ_0 は最適応答を選択する閾値となる．最適な効用の設定は簡単ではないが，設定自体は不自然な作業ではない．例えば医療画像の分類においては，誤分類に対する損失の定量化が行われ，期待損失を最小化する閾値が求められている [17].

6.3.3 サンプル数が少ない場合の MLE の性能

Fig. 6 より，提案手法が MLE に比べ高い性能を有し，学習サンプルが少ない初期のエピソードでもある程度の性能を有することが分かる．特に MLE では，サンプル数が 10 個程度で性能が著しく悪化する．このことは，対話を続けながらオンラインで MLE により学習を行わせた場合，初期のエピソードより中期のエピソードで行動失敗を行う危険性が高くなることを意味する．

Fig. 6 において，MLE のテストセット尤度 $\mathcal{L}$ がサンプル数

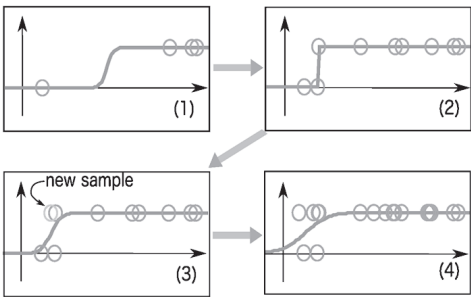


Fig. 9 Four typical phases of learning. The open circles represent training samples. Each curve represents the ICM function trained by the samples

$i = 10$ 付近で最小値（グラフ上は最大値）をとる理由を考察する．いま，

$$\tilde{d} = \min_{u_i=1} d_i - \max_{u_i=0} d_i. \qquad (12)$$

とすると，トレーニングサンプルを無作為に抽出すれば $\tilde{d}$ は単調減少することが分かる．いま，有限回のパラメータ更新により，トレーニングセット i で $\tilde{d}^{\langle i \rangle}, w_1^{\langle i \rangle}, \mathcal{L}^{\langle i \rangle}$ が得られるとする．このとき，$\tilde{d}^{\langle i_1 \rangle} > \tilde{d}^{\langle i_2 \rangle} > 0 > \tilde{d}^{\langle i_3 \rangle}$ の条件下で $\tilde{d}^{\langle i_2 \rangle}$ を 0 に近づけると，ロジスティックシグモイド関数の傾きは $w_1^{\langle i_1 \rangle}, w_1^{\langle i_3 \rangle} < w_1^{\langle i_2 \rangle}$ となる．w_1 が増大すれば，テストセット中の誤分類したサンプルの損失（負の尤度）が増大するので，テストセット尤度は減少する．つまり，このような単調減少する $\tilde{d}$ に対して，テストセット尤度は $\mathcal{L}^{\langle i_1 \rangle}, \mathcal{L}^{\langle i_3 \rangle} > \mathcal{L}^{\langle i_2 \rangle}$ となる．**Fig. 9** に上記の過程の典型的な流れを示す．

(1) 学習の初期フェーズ（$\tilde{d} > 0$）
フェーズ（2）に比べ Fisher のスコアリングアルゴリズムに用いる勾配は小さい．本実験設定では，パラメータ更新が有限回なので（2）に比べて統合確信度関数の傾きは急峻ではない．

(2) $\tilde{d}$ が 0 に近付く（$\tilde{d} \to +0$）
正事例と負事例の境界において，統合確信度関数の傾きはフェーズ（1）より大きい．その結果，誤分類により $\mathcal{L}$ が大きく減少するので，このフェーズで $\mathcal{L}$ は最小値を持つ．

(3) 初めて $\tilde{d} < 0$ なるサンプルが入力されるフェーズ
トレーニングセット尤度を最大化するために，（2）に比べ傾きは緩やかになる．そのため，$\mathcal{L}$ はフェーズ（2）より大きくなる．

(4) 十分なサンプルが入力されたあとのフェーズ．

7. 関 連 研 究

本章では，物体操作対話に限定せず，より広い視点から関連研究について述べる．

本研究では，マルチモーダル入力に基づく発話理解確率推定問題に対してベイズロジスティック回帰を適用した．これに対し音声認識や対話システムの分野では，音声情報のみから得られる確信度に基づいて応答（確認発話や棄却）を決定する手法が提案されてきた（例えば文献 [15] [19]）．音声認識で用いられている確信度については文献 [20] が詳しい．

これらの研究では，主に音声対話におけるエラーハンドリングが扱われることが多い．Komatani らは，ホテル検索用の音声対話システムにおいて確信度により応答選択を行った [21]．また Misu らは，音声認識結果に対する確信度を基準として，タスク達成の効用を最大化する確認発話を生成する手法を提案している [22]．文献 [12] は提案手法と同様，確信度を発話理解確率としてモデル化し，バスの経路を検索する音声対話システムにおけるエラーハンドリングに用いている．文献 [12] と提案手法の差異は，動作を含むマルチモーダル対話に確信度を用いることと，確信度を用いた発話生成を行うことである．

提案手法における統合確信度の機能の一つは，様々なモダリティの情報を統合して，発話理解誤りを低減させることである．つまり，音声認識に起因する発話理解誤りは画像や動作など他のモダリティによってカバーされている．提案手法と異なり動作や画像は含まれないものの，特徴量を統合する目的で確信度が用いられている研究としては，文献 [23] などが挙げられる．文献 [23] では，認識結果の文としての音響的スコア，文に含まれる単語の平均スコア，単語の最小スコアなどの特徴量が統合され，発話の分類器を学習させている．

提案手法における期待効用の最大化については，音声対話システム分野における強化学習による対話戦略学習の研究と関連が深い [24]．Singh らの研究 [24] は，強化学習ベースの対話戦略学習の最初の研究の一つであるが，近年，これを部分観測マルコフ決定過程（POMDP）上の強化学習に拡張した手法が注目されている [25]．提案手法は，マージンから発話理解確率の推定値を求めたうえで，（8）により行動に対する期待効用 $\mathbb{E}[R_i]$ を求めているのに対し，文献 [25] では行動の価値を Value Iteration により求めている．また，リアルタイムではない学習に限られるものの，より柔軟な対話戦略の獲得を目指して重みつき有限状態トランスデューサ（WFST）に基づく手法も提案されている [26]．これらの手法は，（グラウンドしない）知識に基づく対話システムに適用される場合がほとんどである．将来的にはマルチモーダル対話システムに適用されることが期待されるが，動作や画像に関してどのような対話データを収集すべきかについて明らかにされていない．

2 章では，ユーザがオブジェクトを一意に絞りこめるような言語表現を用いた確認発話生成について論じた．Fig. 1 の例では，オブジェクト 2 に対応する言語表現は「箱」「赤い箱」「赤くて四角いもの」など無数にあるが，ロボットは過不足ない自然な表現を用いることが望ましい．

オブジェクトを表す言語表現生成を狙った研究は，人工知能や自然言語生成の分野で広く行われている [3] [4] [27]．山肩らはコップ類の名称とイメージモデルの参照関係における曖昧性に一貫した個人差があることを示している [3]．さらに，音声・言語・画像レベルの情報を統合して，複数のオブジェクトの中からユーザが意図したオブジェクトを選択する手法を提案している．ただし，オブジェクトの属性の種類および属性値の種類は，コップ類に特化して設計者が与えたものであり，提案手法が扱うような実世界のオブジェクトへの応用性は低い．提案手法の文献 [3] に対する主な差異は，1）動作と複数オブジェクトを参照する発話に適用可能，2）設計者により属性が与えられない問題領域に適用可能，3）実世界におけるオブジェクト画像の学

習・認識が可能，の 3 点である．Roy は，ディスプレイ上の複数の長方形のうち一つを指示する言語表現をテキストベースで生成する手法を提案している [4]．文献 [4] で提案された手法では，単語のカテゴリは教師なし学習の枠組みでクラスタリングされるため，設計者が属性を用意する必要はない．ただし，動作とオブジェクトをともに指定する発話を理解する手法ではなく，音声対話による曖昧性解消も考慮されていない．

自然言語生成（NLG）の分野では，コーパスに基づいて機械に言語表現を生成させる試みが行われてきた [27]〜[29]．例えば，TUNA コーパスは，人間が生成する自然言語表現と近い表現を生成する手法を比較評価するために構築されたコーパスである [30]．ただし，システムへの入力は家具の画像と属性値（XML テキスト）であり，属性の種類および属性値の種類は非常に少ない．例えば，サイズ属性は大・小のいずれか，タイプ属性は椅子・ソファ・机・扇風機のいずれか，など事前に定義された属性とその値がシステムに直接与えられる．このような前提は，カメラ画像を処理するような実環境における人間–ロボット対話の前提との隔たりが大きい．他の多くの研究でも，システムへの入力データはテキストや GUI 上の図形で与えられている [27] [28]．このような入力データは，現状では物体操作対話タスクが前提とする「実世界オブジェクトをロボットが操作する」という設定を考慮したものではなく，本研究が扱う問題には適さない．

8. お わ り に

生活支援ロボットが日常環境に導入されるためには，ユーザとの安全・安心なインタラクションを実現する必要がある．本論文では，ユーザの発話の曖昧性を定量化し，タスク達成の効用を最大化する応答を生成する手法を提案した．本手法は，ユーザが曖昧性が少ない発話を行った場合は，状況に応じて最も適切な動作軌道を HMM を用いて生成する．また，曖昧性が大きい発話に対しては，ユーザにとって自然な確認発話を生成することで，不適切な動作を実行前に中止させて行動失敗率を減少させることが可能になった．本研究で構築したシステムの動画は，http://mastarpj.nict.go.jp/˜ksugiura/video_gallery/lcore_dec/ で閲覧可能である．

謝 辞 システム統合を支援していただいた今木理英氏に感謝の意を表する．本研究の一部は，日本学術振興会科学研究費補助金（基盤研究（C）課題番号 20500186）および国立情報学研究所による研究助成を受けて実施されたものである．

参 考 文 献

[1] 徳永健伸，田中穂積：“ロボットにおける言語理解”，日本音響学会誌，vol.63, no.1, pp.35–40, 2007.
[2] P. Dominey, A. Mallet and E. Yoshida: “Real-time cooperative behavior acquisition by a humanoid apprentice,” Proceedings of IEEE/RAS 2007 International Conference on Humanoid Robotics, 2007.
[3] 山肩洋子，河原達也，奥乃博，美濃導彦：“音声対話システムにおける物体指示のための信念ネットワークを用いた曖昧性の解消”，人工知能学会論文誌，vol.19, no.1, pp.47–56, 2004.
[4] D. Roy: “Learning visually grounded words and syntax for a scene description task,” Computer Speech and Language, vol.16, no.3, pp.353–385, 2002.
[5] T. Inamura, I. Toshima, H. Tanie and Y. Nakamura: “Embodied symbol emergence based on mimesis theory,” International Journal of Robotics Research, vol.23, no.4, pp.363–377, 2004.
[6] T. Ogata, M. Murase, J. Tani, K. Komatani and H.G. Okuno: “Two-way translation of compound sentences and arm motions by recurrent neural networks,” Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System, pp.1858–1863, 2007.
[7] 高野渉，中村仁彦：“統計的相関に基づく動作パターンのリアルタイム教師なし分節化と原始シンボルの自律的獲得”，日本ロボット学会誌，vol.27, no.9, pp.1046–1057, 2009.
[8] N. Iwahashi: ‘Robots that learn language: Developmental approach to human-machine conversations,’ Human-Robot Interaction. eds. N. Sanker, et al., pp.95–118, I-Tech Education and Publishing, 2007.
[9] N. Iwahashi, R. Taguchi, K. Sugiura, K. Funakoshi and M. Nakano: “Robots that learn to converse: Developmental approach to situated language processing,” Proceedings of International Symposium on Speech and Language Processing, pp.532–537, 2009.
[10] A. Genkin, D. Lewis and D. Madigan: “Large-scale bayesian logistic regression for text categorization,” Technometrics, vol.49, no.3, pp.291–304, 2007.
[11] 中村慎也，岩橋直人，長井隆行：“実世界における人とロボットの共有信念の推定に基づいた相互適応的な発話生成”，知能と情報，vol.21, no.5, pp.663–682, 2009.
[12] D. Bohus, B. Langner, A. Raux, A. Black, M. Eskenazi and A. Rudnicky: “Online supervised learning of non-understanding recovery policies,” Proceedings of the IEEE/ACL Workshop on Spoken Language Technology, pp.170–173, 2006.
[13] K. Sugiura and N. Iwahashi: “Learning object-manipulation verbs for human-robot communication,” Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction, pp.32–38, 2007.
[14] S. Katagiri, B. Juang and C. Lee: “Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method,” Proceedings of the IEEE, vol.86, no.11, pp.2345–2373, 1998.
[15] T. Kawahara, C. Lee and B. Juang: “Flexible speech understanding based on combined key-phrase detection and verification,” IEEE Transactions on Speech and Audio Processing, vol.6, no.6, pp.558–568, 1998.
[16] N. Iwahashi: “Interactive learning of spoken words and their meanings through an audio-visual interface,” IEICE Transactions on information and systems, vol.91, no.2, p.312, 2008.
[17] C.M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
[18] T. Kurita: “Iterative weighted least squares algorithms for neural networks classifiers,” New generation computing, vol.12, no.4, pp.375–394, 1994.
[19] D. Bohus and A. Rudnicky: “Sorry, I didn’t catch that!-an investigation of non-understanding errors and recovery strategies,” Proceedings of 6th SIGdial Workshop on Discourse and Dialogue, 2005.
[20] H. Jiang: “Confidence measures for speech recognition: A survey,” Speech Communication, vol.45, no.4, pp.455–470, 2005.
[21] K. Komatani and T. Kawahara: “Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output,” Proceedings of the 18th conference on Computational Linguistics, pp.467–473, 2000.
[22] T. Misu and T. Kawahara: “Bayes risk-based optimization of dialogue management for document retrieval system with speech interface,” Proceedings of INTERSPEECH, pp.2705–2708, 2007.
[23] O. Lemon and I. Konstas: “User simulations for context-sensitive speech recognition in spoken dialogue systems,” Pro-

ceedings of EACL 2009, pp.505–513, 2009.
[24] S. Singh, M. Kearns, D. Litman and M. Walker: "Reinforcement learning for spoken dialogue systems," Advances in Neural Information Processing Systems 12 (NIPS), 2000.
[25] J. Williams and S. Young: "Scaling pomdps for spoken dialog management," IEEE Transactions on Audio Speech and Language Processing, vol.15, no.7, pp.2116–2129, 2007.
[26] C. Hori, K. Ohtake, T. Misu, H. Kashioka and S. Nakamura: "Statistical dialog management applied to WFST-based dialog systems," Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4793–4796, 2009.
[27] R. Dale and E. Reiter: "Computational interpretations of the Gricean maxims in the generation of referring expressions," Cognitive Science, vol.19, no.2, pp.233–263, 1995.
[28] P. Jordan and M. Walker: "Learning content selection rules for generating object descriptions in dialogue," Journal of Artificial Intelligence Research, vol.24, no.1, pp.157–194, 2005.
[29] K. Funakoshi, P. Spanger, M. Nakano and T. Tokunaga: "A probabilistic model of referring expressions for complex objects," Proceedings of the 12th European Workshop on Natural Language Generation, pp.191–194, 2009.
[30] I. van der Sluis, A. Gatt and K. van Deemter: "Manual for the TUNA corpus: Referring expressions in two domains," Technical Report AUCS/TR0705, University of Aberdeen, 2006.

杉浦孔明（Komei Sugiura）

2007 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．日本学術振興会特別研究員，ATR 音声言語コミュニケーション研究所 研究員を経て，2009 年より情報通信研究機構 知識創成コミュニケーション研究センター 専攻研究員．ロボットによる言語獲得，ロボットの音声対話機構，および機械学習の研究に興味をもつ．計測自動制御学会，人工知能学会などの会員．（日本ロボット学会正会員）

柏岡秀紀（Hideki Kashioka）

1993 年大阪大学大学院基礎工学研究科博士後期課程修了．博士（工学）．同年国際電気通信基礎技術研究所（ATR）入社．1998 年 ATR 主任研究員．1999 年奈良先端科学技術大学院大学情報科学研究科客員助教授（兼任）．2006 年情報通信研究機構（NICT）専門研究員（兼任）．2006 年 ATR 音声言語処理研究室室長．2009 年 3 月 ATR 退社，2009 年 4 月 NICT 研究マネージャー．主に音声対話，自然言語処理，機械翻訳，音声言語処理の研究に従事．

岩橋直人（Naoto Iwahashi）

1985 年慶應大学理工学部計測工学科卒業．1985 年～1998 年ソニー勤務．1990 年～1993 年 ATR 自動翻訳電話研究所に出向．1998 年～2004 年ソニーコンピュータサイエンス研究所．現在，情報通信研究機構 専攻研究員，ATR メディア情報科学研究所客員研究員．博士（工学）．人とロボットのインタラクションの研究に従事．人工知能学会，日本認知科学会各会員．

中村 哲（Satoshi Nakamura）

1981 年京都工芸繊維大学工芸学部電子情報工学科卒業．同年よりシャープ（株）勤務．京都大学博士（工学）．1994 年～2000 年奈良先端大学助教授．2000 年～2009 年 ATR 音声言語コミュニケーション研究所長，ATR フェロー．現在，情報通信研究機構 知識創成コミュニケーション研究センター 副研究センター長，MASTAR プロジェクト PL．独カールスルーエ大学客員教授，けいはんな連携大学院教授．音声翻訳，音声認識などの音声言語情報処理の研究に従事．電気通信普及財団賞，情報処理学会山下賞，AAMT 長尾賞，ドコモモバイルサイエンス賞，情報処理学会業績賞，日本音響学会技術開発賞受賞．IEEE，電子情報通信学会，情報処理学会，日本音響学会会員．