物体操作タスクにおける発話理解確信度の推定に基づく発話と動作の生成

○杉浦孔明 †,††，岩橋直人 †,††

†(独) 情報通信研究機構　　††(株) 国際電気通信基礎技術研究所

Generation of Utterances and Motions Based on Confidence Measure Estimation for Utterance Understanding in Object Manipulation Tasks

*Komei Sugiura†,††，Naoto Iwahashi†,††

†National Institute of Information and Communications Technology　††ATR

Abstract— This paper proposes a method that generates motions and utterances in an object-manipulation task. Belief modules of speech, vision, motions are integrated in a probabilistic framework for understanding user's utterances. Responses to the utterances are optimized based on a confidence measure function for the integrated belief modules. Fisher's scoring algorithm is used for training the confidence measure function. Experimental results reveals that the proposed method can improve the success rate in the object-manipulation task.

Key Words: motion learning, confidence, logistic regression, spoken dialogue, language acquisition

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とのコミュニケーション機能は極めて重要であるが，現状の対話処理技術は必要なレベルに全く到達していない．

さらに，従来の対話技術には安全性上の観点から大きな問題がある．それは，ユーザの発話の意味が適切に理解されずに，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ことである．

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 (物体操作指示) を対象とする．物体操作指示において，ユーザの発話の意味が適切に理解されるためには，(1) 言語による動作参照，(2) 言語によるオブジェクト参照，における曖昧性を解消する必要がある．

これに対し従来研究では，コンピュータ上のオブジェクトを用いて，オブジェクトを指示する最適な表現を生成して (2) の曖昧性を解消する手法が提案されている [6,9]．また，音声対話の分野では，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研究が行われている [5]．しかしながら，(1)(2) の曖昧性をともに解消する研究は今までにない．一方我々は，実世界にグラウンドした動作のイメージをユーザとロボットが共有する手法 LCore を提案している [1]．よって LCore に確信度に基づく確認発話生成を導入することで，上記の危険性を解消できる可能性がある．

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 (動作あるいは確認発話) を生成する手法 LCore-DEC を提案する．提案手法の独自性は，(1) 音声・視覚・動作などを統合したユーザモデルに対する確信度を，統計的学習手法に基づいて学習すること，(2) 確信度を用いて過不足ない自然な確認発話を生成すること，である．さらに，ユーザの動作指示発話を受けてロボット (Fig. 1 参照) が行動する場合には，動作の確率モデルをもとに，状況に応じて最も適切な動作軌道を計画する．この動作はユーザの教示によりロボットが学習したものであるので，ユーザにとってロボットの動作がイメージしやすい．

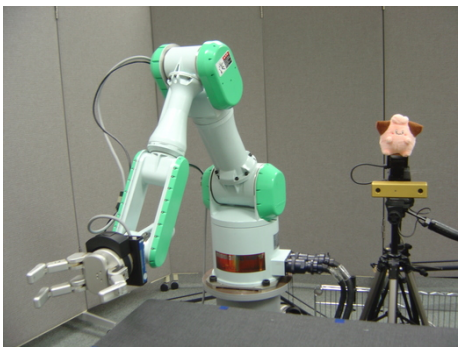


Fig.1 LCore-DEC を実装したロボット

2. 共有信念モデルに基づく発話理解

提案手法では，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．さらにこのユーザモデルを用いて，適切な行動や発話を生成する．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 5 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．信念モジュールと共有信念は，教師あり学習の枠組みにより学習される．教師データは，ユーザとロボットの実世界インタラクションを通じて収集される．

2·1　レキシコン

2·1.1　視覚情報・動作と対応した単語集合

本論文では，音声，視覚，動作と対応付けた単語集合をレキシコンと呼ぶ．レキシコン L は，Fig. 2 に示すような概念インデックスと対応づけられた確率モデルの集合である．各確率モデルは，マイク，カメラなどから得られた入力に統計的学習手法を適用することにより得られる．図において，$C_N^{\langle i \rangle}$ はオブジェクトの視覚的特徴を表す要素であり，$C_V^{\langle j \rangle}$ は動作を表す要素である．また，i, j は要素のインデックスを表す．

以下では，各信念モジュールに対する説明の準備として，音声・視覚・動作信念に関する確率モデルについて説明する．

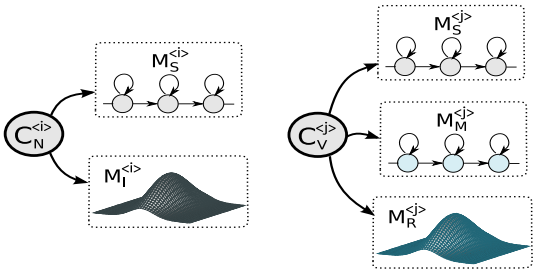


Fig.2 レキシコン：確率モデルと概念インデックスの対応

2·1.2　視覚特徴を表す確率モデル

$C_N^{\langle i \rangle}$ は 2 種類の確率モデルに対応づけられている．すなわち，音声特徴量を表す確率モデル $M_S^{\langle i \rangle}$ と，オブジェクトの画像特徴量ベクトルを表す確率モデル $M_I^{\langle i \rangle}$ である．$M_S^{\langle i \rangle}$ は隠れマルコフモデル (HMM) で表現され，$M_I^{\langle i \rangle}$ は多次元ガウス分布で表現される．$C_N^{\langle i \rangle}$ により，「アカイ」「エルモ」などの音声 (実際には音韻列) と，対応する視覚特徴が対応づけられる．M_I と $M_S^{\langle j \rangle}$ の学習には，[1] に示す方法を用いる．

2·1.3　動作を表す確率モデル

$C_V^{\langle j \rangle}$ は 3 種類の確率モデルと固有座標系タイプ [7] に対応づけられている．すなわち，音声特徴量を表す確率モデル $M_S^{\langle j \rangle}$，動作 (物体操作軌道) を表す確率モデル $M_M^{\langle j \rangle}$，動作-オブジェクト関係を表す確率モデル $M_R^{\langle j \rangle}$ である．$C_N^{\langle i \rangle}$ の場合と同様に，$M_S^{\langle j \rangle}$ は HMM で表現される．$M_M^{\langle j \rangle}$ は，オブジェクトの位置，速度，加速度の時系列を表す確率モデルであり，HMM を用いて表現される．$C_V^{\langle j \rangle}$ により，「まわす」「のせる」などの音声，動作，動作に関連する視覚特徴が対応づけられる．

$M_R^{\langle j \rangle}$ は動作に関連するオブジェクト (群) の画像特徴量を表す確率モデルである．$M_R^{\langle j \rangle}$ は多次元ガウス分布で表現され，分布の推定にはベイズ学習を用いる．$M_R^{\langle j \rangle}$ により，物体に対する行為の可能性 (アフォーダンス) を，画像特徴量に対するモデルの尤度として評価することが可能になる．

$M_M^{\langle j \rangle}$ の学習には，参照点に依存した物体操作の学習手法を用いる [7,8]．これは，動作の基準となる参照点の推定を行ない，参照点に依存した物体操作軌道を表す HMM を学習する手法である．いま，ユーザが Fig. 3 に示す点線に沿ってオブジェクトを動かしたとする．この軌道は，トラジェクタ (動かされるオブジェクト) とランドマーク (動作の基準となるオブジェクト) との相対軌道としてモデル化される．Fig. 3 では，オブジェクト 2 の重心が参照点となる．

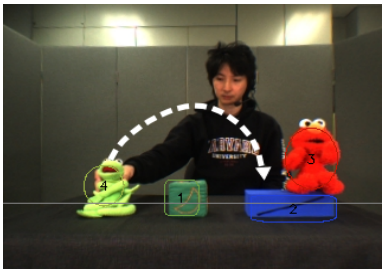


Fig.3 カメラ画像の例．オブジェクト上の数字は，画像から抽出された物体のインデックスを表す．

2·2　文法

本節では，音声信念モジュールに関係するもう一つの確率モデルである文法について述べる．本手法では，発話における文節や単語の並びの規則は，文法 G により表現される．文法 G は，(1) 文節列の出現確率 P_S と，(2) 文節内での単語インデックス列の出現確率 P_W からなる．

ユーザの発話 s は，トラジェクタを表す文節 W_T，ランドマークを表す文節 W_L，動作を表す文節 W_M からなる概念構造 $z = (W_T, W_L, W_M)$ と対応づけて解釈される．ただし本手法では，s に含まれる動詞の活用形は全て命令形であり，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，$z = (W_T, W_M)$ である．

P_S および P_W は，ユーザが与えた教師データから学習される．P_S により，語順や文節が省略される傾向をモデル化できる．くわえて語順をモデル化するので，語順が異なる言語にも対応できる．P_W は，文節中の単語列の出現頻度をバイグラム確率として学習させる．文法 G の学習の詳細については，[8] を参照されたい．

2·3　信念モジュール

シーン O において発話 s が与えられたとする．O において可能な動作の集合 A は以下により与えられる．

$$A = \{(i_t, j, r) \mid i_t = 1, ..., O_N, \\ j = 1, ..., V_N,\ r = 1, ..., R_N\} \quad (1)$$

ここに，トラジェクタのインデックスを i_t，参照点のインデックスを 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，$C_V^{\langle j \rangle}$ に対して可能な参照点の数を R_N とする．従って，発話 s によって物体操作を行わせるタスクは，s に対して $a_k \in A$ を選択するタスクであるといえる．

このとき，各信念モジュールを以下のように定義する．まず，レキシコン L と文法 G をパラメータとして，音声信念 B_S を発話 s に対する z の対数尤度として表す．視覚信念 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 L の対数尤度である．同様に，動作-オブジェクト関係信念 B_R は，オブジェクト (i, j) の視覚特徴量に対する L の対数尤度である．a_k に対する最尤軌道を $\mathcal{Y}_k$ とすると，動作信念 B_M は，トラジェクタ i の位置 $\mathbf{x}_p^{\langle i_t \rangle}$ が与えられたうえでの $\mathcal{Y}_k$ に対する L の対数尤度

で表される．

行動コンテキスト信念 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．$q_1^{\langle i\rangle}, q_2^{\langle i\rangle}$ 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とすると 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 は，Minimum Classification Error(MCE) 学習 [2] に基づいて学習される．

2·4 共有信念関数

共有信念関数 Ψ を，各信念モジュールの重み付け和として以下のように定義する．

$$\begin{aligned} &\Psi(s, a_k, O, \mathbf{q}^{\langle i_t\rangle}; L, G, \boldsymbol{\gamma}) \\ &= \max_{r,z}\Big\{ \ \gamma_1 \log P(s|z; L, G) \\ &+ \gamma_2 \log P(\mathcal{Y}_k|\mathbf{x}_p^{\langle i_t\rangle}, C_V^{\langle j\rangle}, \mathbf{x}_p^{\langle r\rangle}; L) \\ &+ \gamma_3 \Big(\log P(\mathbf{x}_I^{\langle i_t\rangle}|W_T; L) + \log P(\mathbf{x}_I^{\langle i_l\rangle}|W_L; L)\Big) \\ &+ \gamma_4 \log P(\mathbf{x}_I^{\langle i_t\rangle}, \mathbf{x}_I^{\langle i_l\rangle}|C_V^{\langle j\rangle}; L) \\ &+ \gamma_5 \Big(B_H(i_t, \mathbf{q}^{\langle i_t\rangle}; h_c) + B_H(i_l, \mathbf{q}^{\langle i_l\rangle}; h_c)\Big)\Big\} \quad (3) \end{aligned}$$

ここに，$\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_5)$ は，各信念に対する重みを表し，i_l はランドマークのインデックスを表す．$\boldsymbol{\gamma}$ 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．

3. 発話理解確信度の推定に基づく発話と動作の生成

3·1 統合確信度による発話理解確率のモデル化

提案手法は，行動を指示する発話の曖昧性を定量化し，あらかじめ定義された効用を最大化する応答 (動作あるいは確認発話) を生成するものである．本節では，まず共有信念関数に基づく曖昧性の尺度について説明する．

前節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 $\hat{a}_k$ は以下で得られる．

$$\hat{a}_k = \underset{a_k}{\operatorname{argmax}}\, \Psi(s, a_k, O, q; L, G, \boldsymbol{\gamma}) \quad (4)$$

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

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

ただし，パラメータ $L, G, \boldsymbol{\gamma}$ の表記を省略した．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 $\hat{a}_k$ に対するマージンは $\hat{a}_k$ と a_l の共有信念関数の値の差であることがわかる．よって，$\hat{a}_k$ に対するマージンが 0 に近ければ，発話 s は $\hat{a}_k$ 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，$\hat{a}_k$ の方が s の指示する行動として適している．従ってマージン関数は，行動 $\hat{a}_k$ 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 $\hat{a}_k$ に対する確信度を得ることを考える．音声認識の分野では，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 [3]．また，確信度は，動作認識や音声認識において新規動作や未登録語の検出にも用いられている．

提案手法では，統合確信度関数 $f(d)$ を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．$f(d)$ の形から $0 < f(d) < 1$ であり，マージン d が大きいほど $f(d)$ が 1 に近付くことがわかる．

3·2 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 $\{(d_i, u_i)|i = 1, ..., N\}$ として与える．ただし，u_i は 0 または 1 の 2 値であるとする．

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．このとき，学習サンプル集合に対する確信度関数の対数尤度は，以下で与えられる．

$$\mathcal{L} = \sum_{i=1}^{N} \{u_i \log f(d_i) + (1 - u_i) \log(1 - f(d_i))\} \quad (7)$$

ただし，学習サンプル間の独立性を仮定する．

統合確信度関数のパラメータ $\mathbf{w}$ の学習には，Fisher のスコアリングアルゴリズム [4] を用いる．これは，対数尤度 $\mathcal{L}$ 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．

3·3 期待効用最大化に基づく応答の決定

ユーザの発話 s に対してロボットが行った動作応答が，ユーザがロボットに行わせたい行動と異なることは，安全性の観点から望ましくない．これに対し，統合確信度を用いれば，このような危険を回避できる可能性がある．例えば，発話 s に対する最適行動 $\hat{a}_k$ の統合確信度が小さければ，ユーザに $\hat{a}_k$ を行うか否かを確認する発話をすればよい．本節では，応答に対する効用を導入し，これを最大化する応答 (最適応答) に関する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から，応答 $b_i (i = 1, 2)$ に対する期待効用 $\mathbb{E}[R_i]$ および最適応答 $\hat{b}$ を以下のように推定することができる．

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

$$\hat{b} = \underset{i}{\operatorname{argmax}}\, \mathbb{E}[R_i] \quad (9)$$

ただし，r_{i1}, r_{i2} はそれぞれ，$\hat{a}_k$ が正解，不正解であったときの応答 b_i に対する効用である．Fig. 4 右図に効

用をまとめた．

ここで，$r_{12} < r_{22} < r_{21} < r_{11}$ であるとする．$\mathbb{E}[R_i]$ は $f(d)$ の線形関数であるので，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 (Fig. 4 参照)．つまり，θ_0 を閾値として最適応答が選択できる．

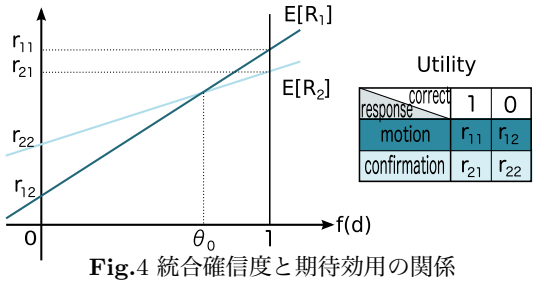


Fig.4 統合確信度と期待効用の関係

次に，確認発話において，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．ただし，加える単語数は，$f(d) \geq \theta_0$ を満たす最小の単語数とする．

以上をまとめて，提案手法 LCore-DEC のアルゴリズムを示す．LCore-DEC は，シーン O，コンテキスト $\mathbf{q}$，発話 s を入力として最適応答を生成する．

1. 動作候補集合 $A = \{a_k | k = 1, 2, ..., |A|\}$ の全ての要素について実行予定軌道を生成し，共有信念関数 $\Psi(s, a_k, O, \mathbf{q})$ を求める．

2. 最適行動 $\hat{a}_k = \underset{a_k}{\operatorname{argmax}} f(d(s, a_k, O, \mathbf{q}))$ に対し，$f(d(s, \hat{a}_k, O, \mathbf{q})) \geq \theta_0$ ならば $\hat{a}_k$ を動作応答として終了．$f(d(s, \hat{a}_k, O, \mathbf{q})) < \theta_0$ ならば **3** へ．

3. 確認動作ターゲット集合 A' を $A' = A$ で初期化．

4. ターゲット動作を $a_j = \underset{a_j \in A'}{\operatorname{argmax}} f(d(s, a_j, O, \mathbf{q}))$ とする．

5. レキシコンの要素から，マージンを最大化する単語 1 語を概念構造 z に追加する．すなわち，$\hat{C}_N^{\langle i \rangle} = \underset{C_N^{\langle i \rangle}}{\operatorname{argmax}} d(s, a_j, O, \mathbf{q})$ なる視覚特徴を表す単語 $\hat{C}_N^{\langle i \rangle}$ を W_L または W_T に加える．ただし，分節中で重複する単語は用いない．

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．更新された z をもとに音声を合成して発話を行う．ただし，W_T と W_L のうち単語が追加されない文節については確認しない．

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．空集合でなければ **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．

4. 実験手法: タスク環境および実験設定

4·1 ハードウェア

実験に用いたロボットシステムを Fig. 1 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)，視線表出ユニットからなる．

オブジェクトに関する画像特徴や座標は，ステレオカメラから得られた画像から抽出される．Fig. 3 に，カメラより得られた画像の例と，それに対応する観測情報の内部表現を示す．カメラのフレームレートを 30[frame/sec] とし，解像度を 320×240 とした．画像特徴量として，色 3 次元 (L*a*b*)，形状 3 次元を用いる．

4·2 信念モジュールと共有信念関数の学習

共有信念関数の学習を行う前に，あらかじめレキシコン L と文法 G を学習させた．学習させた単語の一覧を Table 1 に示す．オブジェクトを指示する単語に対しては，音声・画像特徴量の組を平均 7 セットずつ与えた．また，動作を指示する単語に対しては，音声・軌道の組を 15 セットずつ与えた．ただし，動作を指示する単語についての動作-オブジェクト関係については，共有信念関数の学習時に同時に教示した．文法 G を学習させるための教師データは，被験者にオブジェクトを操作させながら発話を行わせることで収集した．G の学習に用いた教師データの総数は 72 セットである．

次に，動作-オブジェクト関係 M_R，行動コンテキスト信念 B_H，共有信念関数のパラメータ $\boldsymbol{\gamma}$ の学習を行わせた．まず，Fig. 3 に示すように，被験者をロボットとオブジェクト (ぬいぐるみ) が置かれたテーブルをはさんで対面させた．被験者に，オブジェクトをロボットに操作させるための発話を行わせ，常に最適行動を動作応答とする方策によりロボットにオブジェクト操作を行わせた．「ユーザによる発話，ロボットによる動作，ユーザによる正解・不正解の評価」を 1 つのエピソードとして，オンラインで $M_R, B_H, \boldsymbol{\gamma}$ の学習を行わせた．エピソード数は 96 であり，1 エピソードのチャンスレベルは平均 2.37%であった．ユーザの発話に含まれる単語数は平均 3.39 語であった．学習後の $\boldsymbol{\gamma}$ の値を Table 2 に示す．

Table 1 学習させた単語

オブジェクトを指示する単語 C_N			
アカイ	アオイ	ミドリ	キイロイ
オレンジ	マルイ	オオキイ	グローバー
チイサイ	エルモ	カーミット	チュートトロ
プーサン	ハコ	バーバズー	バーバブライト
動作を指示する単語 C_V			
のせて	はなして	ちかづけて	とびこえさせて
あげて	さげて	まわして	

Table 2 学習後の共有信念関数のパラメータ

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

4·3 提案手法の評価実験

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．

実験 (1) の目的は，学習が収束するサンプル数について検討することで，統合確信度の学習に関する提案手法の性能評価を行うことである．また実験 (2) の目的は，提案手法による動作失敗率の減少について検証することである．

実験 (1) において，統合学習の訓練および評価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．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 個を訓練集合，残りの 50 個を評価集合とした．実験 (1) では，共有信念関数における重み γ の更新を行わない．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．

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) の訓練集合を用いて学習され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，評価集合からデータを 1 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させる．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，正解動作以外の動作が実行されたエピソードの割合を動作失敗率として評価する．

5. 実験結果と考察

5·1 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 は訓練サンプル数を変えたときの回帰結果である．(a)〜(d) は，それぞれ訓練サンプル数 10, 15, 20, 25 のときの結果に対応する．図より，サンプル数 20 までに収束していることがわかる．

次に，定量的結果について検討する．Fig. 6 に，テスト集合に対する統合確信度関数の対数尤度 $\mathcal{L}$ を示す．図には，訓練サンプル数に対する $\mathcal{L}$ をプロットした．ただし，図に示した $\mathcal{L}$ は 10 回の実験における平均値である．各実験は，100 個の教師データを 50 個ずつ訓練集合とテスト集合にランダムに振り分けて行った．Fig. 6 より，訓練サンプル数 20 までに学習が収束していることがわかる．

対数尤度 $\mathcal{L}$ がサンプル数 $i = 10$ 付近で最小値をとる理由を考察する．$\tilde{d} = \min_{u_j=1} d_j - \max_{u_k=0} d_k$ とすると，訓練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わかる．いま，有限回のパラメータ更新により，訓練集合 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，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．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．

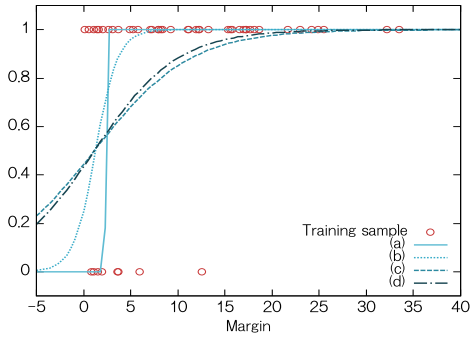


Fig.5 統合確信度関数の学習結果．○は訓練サンプルを表す．(a)〜(d) は，それぞれ訓練サンプル数 10, 15, 20, 25 のときの回帰結果を表す．

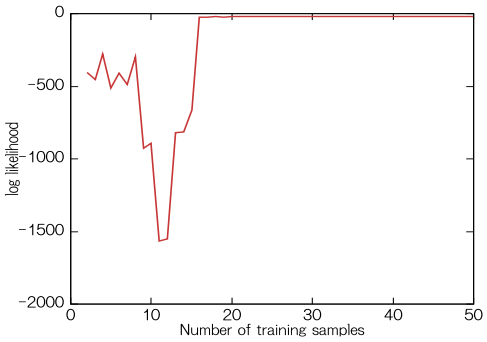


Fig.6 テスト集合に対する統合確信度関数の対数尤度

5·2 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，ロボットの動作は正しいもの (「オブジェクト 2 をオブジェクト 3 に載せる」) であった．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，正解動作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，正解動作軌道の尤度は，動作候補 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，正解動作に対するスコアが動作候補のなかで最大になった．さらに，正解動作の確信度は $f(d) = 0.998 > \theta_0$ 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．

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．

Table 3 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．

- 動作失敗率：動作を行ったエピソードにおける，正解動作以外の動作が実行されたエピソードの割合．把持失敗やオブジェクト同士の衝突などに関する

Table 3 確信度に基づく意志決定手法の評価

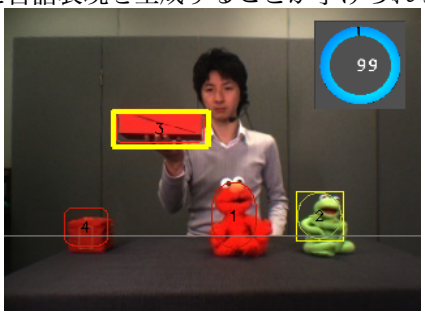
θ_0	0	0.7	0.9	0.99	0.999
実行失敗率 [%]	12.0	10.4	6.5	7.1	2.6
棄却率 [%]	0	4.0	8.0	16.0	24.0
確認発話率 [%]	0	12.0	22.0	28.0	48.0
平均確認発話数	-	1.17	1.27	1.21	1.25

失敗については考慮しない．

- 棄却率：全エピソードの中で，動作が実行されなかったエピソードの割合
- 確認発話率：全エピソードの中で，確認発話がなされたエピソードの割合
- 平均確認発話数：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

Table 3 において，$\theta_0 = 0$ はユーザの発話に対して常に動作を行う方策である．このとき，実行失敗率は 12.0% (6/50) であった．表より，$\theta_0 = 0$ 以外の場合には，実行失敗率が 12.0%より小さくなっている．さらに，θ_0 の増加に伴って実行失敗率が低下する傾向が見られた．例えば $\theta_0 = 0.9$ の場合には，実行失敗率は 6.5% (3/46) であった．Table 3 より，$\theta_0 = 0$ 以外の実験条件において，確認発話率は 50%以下であり，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 3 より，θ_0 の増加に伴って実行失敗率が低下する一方，棄却率は上昇していることがわかる．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わけられる．(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 2 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．

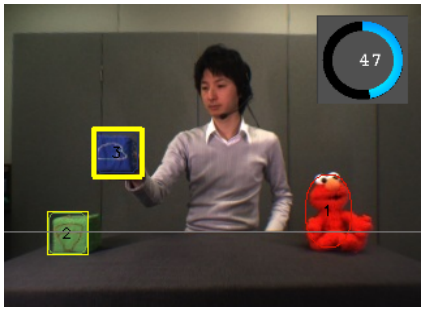


【状況】オブジェクト 2 が直前に操作された
U: のせて．
R: (動作実行：オブジェクト 2 をオブジェクト 3 に載せる)

Fig.7 対話例 (1)．確認発話を行わなかったケース．

6. 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 LCore-DEC 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．この動作はユーザの教示から学習されたものであり，ユーザにとってロボットの動作がイメージしやすいため，予期しない動作を行う不安を解消できる．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実行失敗率を減少させることが可能になった．



【状況】オブジェクト 2 が直前に操作された
U: ハコ エルモ ちかづけて．
R: ミドリハコをちかづけて？
U: いいえ．
R: アオイハコをちかづけて？
U: はい．
R: (動作実行：オブジェクト 3 をオブジェクト 1 に近づける)

Fig.8 対話例 (2)．動作実行前に確認発話を行ったケース．

謝辞

本研究の一部は，立石科学技術振興財団による研究助成を受け実施したものである．

参考文献

[1] Iwahashi, N.: Robots That Learn Language: Developmental Approach to Human-Machine Conversations, *Human-Robot Interaction*, pp. 95–118 (2007).

[2] Katagiri, S. et al.: Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method, *Proc. of the IEEE*, Vol. 86, No. 11, pp. 2345–2373 (1998).

[3] Kawahara et al.: Flexible speech understanding based on combined key-phrase detection and verification, *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 6, pp. 558–568 (1998).

[4] Kurita, T.: Iterative weighted least squares algorithms for neural networks classifiers, *New generation computing*, Vol. 12, No. 4, pp. 375–394 (1994).

[5] Misu, T. et al.: Bayes Risk-based Optimization of Dialogue Management for Document Retrieval System with Speech Interface, *Proc. of INTERSPEECH*, pp. 2705–2708 (2007).

[6] Roy, D.: Learning visually grounded words and syntax for a scene description task, *Computer Speech and Language*, Vol. 16, No. 3, pp. 353–385 (2002).

[7] Sugiura, K. et al.: Learning object-manipulation verbs for human-robot communication, *Proc. of IWMISI 2007*, pp. 32–38 (2007).

[8] 羽岡哲郎ほか: 言語獲得のための参照点に依存した空間的移動の概念の学習, 信学技報, PRMU2000-105, pp. 39–46 (2000).

[9] 山肩洋子ほか: 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).