# 言語獲得ロボットにおける応答戦略最適化のための確信度のベイズ学習

○杉浦孔明, 岩橋直人, 柏岡秀紀, 中村哲 (情報通信研究機構)

## Bayesian Learning of Confidence Measure Function for Response Optimization in Robot Language Acquisition Framework

*Komei Sugiura, Naoto Iwahashi, Hideki Kashioka, and Satoshi Nakamura
National Institute of Information and Communications Technology

**Abstract—** 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:** Bayesian logistic regression, confidence, spoken dialogue system, robot 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グラウンドされない知識に基づいて処理されるため，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，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．

(1) の曖昧性解消に対して多くの研究 (例えば [6, 8]) が行われてきたものの，(1)(2) 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 [4].

本論文では，適応的確信度に基づく動作・発話生成を行う手法 LCore-DEC を提案する．LCore-DEC 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) の推定問題に対して，Bayesian Logistic Regression (BLR) [2] 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．

確信度を用いた対話制御の既存手法 (例えば [1]) に対する本手法の利点は，確信度を成功確率として推定するので，応答の期待効用が計算可能なことである．

## 2. LCore における発話理解

LCore [4] 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 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 $\Psi$ と呼ぶ．

### 2·1 物体操作対話タスク

Fig. 1 に，物体操作対話タスクを行うロボットの例を示す．いま，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．このときのユーザが意図する軌道を白線で表す．本手法では，ロボットの動作生成手法として，[7] で提案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．Fig. 1 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．


**Fig.**1 物体操作対話の例

### 2·2 共有信念モデルに基づく発話理解

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．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．

$W_T$ : バーバブライト， $W_L$ : 赤い 箱， $W_M$ : 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．

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．$O$ において可能な動作の集合 $A$ は以下により与えられる．

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

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 $r$，$O$ 中のオブジェクトの数を $O_N$，動作を表す単語数を $V_N$，動詞 $C_V^{\langle j\rangle}$ に対して可能な参照オブジェクトの数を $R_N$ とする．従って，物体操作対話タスクでは，$s$ に対し正しい $a_k$ を選択することが求められる．

各信念モジュールを以下のように定義する．まず，音声信念 $B_S$ を発話 $s$ に対する $z$ の条件付き確率の対数として表す．視覚信念 $B_I$ 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確率モデルの対数尤度である．$a_k$ に対する最尤軌道を $\hat{\mathcal{Y}}_k$ 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 $\hat{\mathcal{Y}}_k$ 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 $\mathbf{q}^{\langle i\rangle}$ 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．コンテキストの例としては，「オブジェクト $i$ が把持されている」，「直前に操作された」などが挙げられる．

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

$$\begin{aligned}\Psi(s,a_k,O,\mathbf{q}^{\langle i_t\rangle}) = \max_z \Big\{ & \gamma_1 \log P(s|z) \\ & +\gamma_2 \log P(\hat{\mathcal{Y}}_k|\mathbf{x}_p^{\langle i_t\rangle},\mathbf{x}_p^{\langle i_r\rangle},C_V^{\langle j\rangle}) \\ & +\gamma_3 \left(\log P(\mathbf{x}_I^{\langle i_t\rangle}|W_T)+\log P(\mathbf{x}_I^{\langle i_r\rangle}|W_L)\right) \\ & +\gamma_4 \log P(\mathbf{x}_I^{\langle i_t\rangle},\mathbf{x}_I^{\langle i_r\rangle}|C_V^{\langle j\rangle}) \\ & +\gamma_5 \left(B_H(i_t,\mathbf{q}^{\langle i_t\rangle})+B_H(i_r,\mathbf{q}^{\langle i_r\rangle})\right)\Big\}, \end{aligned} \quad (2)$$

ここに，$\mathbf{x}_p^{\langle i\rangle}$ はオブジェクト $i$ の位置，$\gamma=(\gamma_1,...,\gamma_5)$ は，各信念に対する重みを表す．$\gamma$ の学習には，Minimum Classification Error(MCE) 学習を用いる．$\Psi$ 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応答生成

### 3·1 統合確信度による発話理解確率のモデル化

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．

$$\hat{a}_k = \underset{a_k\in A}{\operatorname{argmax}}\, \Psi(s,a_k,O,\mathbf{q}) \quad (3)$$

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．

$$d(s,a_k,O,\mathbf{q}) = \Psi(s,a_k,O,\mathbf{q}) - \max_{j\neq k}\Psi(s,a_j,O,\mathbf{q}) \quad (4)$$

いま，最大値の次に大きい値を与える行動を $a_l$ とする．(4) 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 [5].

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．

$$f(d;\mathbf{w}) = \frac{1}{1+\exp(-(w_1 d+w_0))} \quad (5)$$

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．

### 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．

$$\mathbb{T}^{\langle N\rangle} = \{(d_i,u_i)|i=1,...,N\}, \quad (6)$$

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．本手法では，BLR [2] を用いて $\mathbf{w}$ の推定を行う．$w_j$ $(j=0,1)$ の事前分布として，平均 $m_j$，分散 $\tau_j$ のガウス分布を用いる．

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

### 3·3 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動作応答が，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．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．このとき応答 $b_i (i=1,2)$ に対する期待効用 $\mathbb{E}[R_i]$ を以下のように推定することができる．

$$\mathbb{E}[R_i] = r_{i1}f(d)+r_{i2}(1-f(d)) \quad (7)$$

ただし，$r_{i1},r_{i2}$ はそれぞれ，$\hat{a}_k=a^*$，$\hat{a}_k\neq a^*$ 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．$\mathbb{E}[R_i]$ 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 $\theta_0$ を持つ．つまり，$\theta_0$ を閾値として最適応答 $\hat{b} = \mathrm{argmax}_i \mathbb{E}[R_i]$ が選択できる．

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[1]．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 $\psi(s, a_k, O, \mathbf{q}^{\langle i_l \rangle}, z)$ で表すことにする．$\psi$ と $\Psi$ の違いは，$\psi$ 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 $\mathbf{c}' = \{c'_m \mid m = 1, ..., M\}$ 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 $\hat{\mathbf{c}}' = \{\hat{c}'_m \mid m = 1, ..., M\}$ と最適挿入位置集合 $\hat{\mathbf{p}} = \{\hat{p}_m \mid m = 1, ..., M\}$ は以下で与えられる．

$$(\hat{\mathbf{c}}', \hat{\mathbf{p}}) = \underset{c'_m \notin W, \mathbf{p}}{\mathrm{argmax}} d_z(s, a_j, O, \mathbf{q}, z) \tag{8}$$

よって挿入後の文節 $W'$ は以下のようになる．

$$W' = c_1 \cdots c_{\hat{p}_1 - 1} \hat{c}'_1 c_{\hat{p}_1} \cdots c_{\hat{p}_2 - 1} \hat{c}'_2 c_{\hat{p}_2} \cdots c_{|W|} \tag{9}$$

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．

$$z' = (W'_T, W'_L, W_M). \tag{10}$$

以上をまとめて，提案手法 LCore-DEC 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.** 動作候補集合 $A$ (式 (1) 参照) の全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 $\hat{a}_k$ (式 (3) 参照) 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.** 単語挿入数 $M$ を $M = 0$ と初期化．ターゲット動作を $a_j = \underset{a_j \in A'}{\mathrm{argmax}} f(d(s, a_j, O, \mathbf{q}))$ とする．

**5.** $M \leftarrow M + 1$ とし，式 (10) 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．

[1] 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

## 4. 実験

### 4·1 設定

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．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．あらかじめ，[3, 7] 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．

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ある (統合確信度学習の性能評価)．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，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．ハイパーパラメータを $m_0 = 0, m_1 = 1, \tau_0 = \tau_1 = 100$ と設定した．

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，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．

### 4·2 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. 2 左図に示す．

次に，定量的結果について検討する．Fig. 2 右図に，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 個のサンプルを 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．

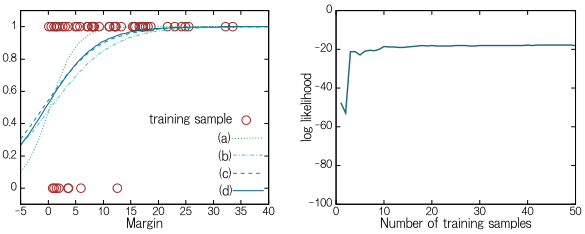



**Fig.**2 左: 統合確信度関数の学習結果．(a)〜(d) は，それぞれ訓練サンプル数 10, 20, 30, 50 のときの回帰結果を表す．右: テスト集合に対する統合確信度関数の対数尤度．

Fig. 2 より，訓練サンプル数 10 までに学習が収束していることがわかる．

### 4·3 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. 3 にユーザ (U) とロボット (R) の対話例を示す．このとき $\theta_0 = 0.7$ と設定した．図において右上の数値は統合確信度を表す．

Fig. 3 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 1 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．

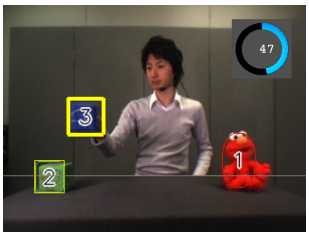
- 動作失敗率：動作を行ったエピソードにおける，正解動作以外の動作が実行されたエピソードの割合
- 棄却率：全エピソードの中で，動作が実行されなかったエピソードの割合
- 確認発話率：全エピソードの中で，確認発話がなされたエピソードの割合
- 平均確認発話数：確認発話が行われたエピソードにおける確認発話の平均回数

Table 1 において，$\theta_0 = 0$ の条件は本手法を用いない場合を意味する．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．$\theta_0 = 0$ における動作失敗率は 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，動作失敗率が 12.0%より小さくなっている．例えば $\theta_0 = 0.999$ の場合には，動作失敗率は 2.6% (1/38) であった．また，Table 1 より，$\theta_0 \neq 0$ では，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 1 より，$\theta_0$ の増加に伴って動作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)$\theta_0$ 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．

**Table** 1 確信度に基づく応答生成手法の評価

| $\theta_0$ | 0 (baseline) | 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 |




【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.**3 対話例．動作実行前に確認発話を行ったケース．

## 5. 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．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．

### 謝辞

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] Bohus, D. et al.: Online supervised learning of non-understanding recovery policies, *Proc. of the IEEE/ACL Workshop on Spoken Language Technology*, pp. 170–173 (2006).
[2] Genkin, A. et al.: Large-scale bayesian logistic regression for text categorization, *Technometrics*, Vol. 49, No. 3, pp. 291–304 (2007).
[3] Iwahashi, N.: Interactive Learning of Spoken Words and Their Meanings through an Audio-Visual Interface, *IEICE Trans. information and systems*, Vol. 91, No. 2, p. 312 (2008).
[4] Iwahashi, N.: Robots That Learn Language: Developmental Approach to Human-Machine Conversations, *Human-Robot Interaction*, pp. 95–118 (2007).
[5] Kawahara, T. 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).
[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*, pp. 32–38 (2007).
[8] 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).