音声言語情報処理 39－14
(2001. 12. 20)

視覚情報を話題の対象とする音声対話システム

山肩 洋子† 河原 達也† 奥乃 博†

† 京都大学 情報学研究科 知能情報学専攻
〒606–8501 京都府京都市左京区吉田本町

E-mail: †{yamakata,kawahara,okuno}@kuis.kyoto-u.ac.jp

あらまし ユーザとの音声対話により実世界中でオブジェクトを探索するロボットの実現を目指す。音声認識や画像認識においては認識誤り、言語情報と視覚情報の対応づけには個人差によるあいまい性が生じる。また、ユーザの信念の誤りによって誤解が生じる可能性もある。そこで本研究では、信念ネットワーク及びユーザモデルを導入し、これらの確率的枠組みに基づいてユーザとの対話をプランニングすることで上記の問題の解決を図る。ユーザの視野外におけるオブジェクト探索タスクで実装を行った結果、ユーザの意図したオブジェクトを同定するまでに必要な対話回数を削減でき、また画像認識結果から音声認識結果を絞り込めることを示した。

キーワード 音声対話システム, 音声言語理解, ユーザモデル, 信念ネットワーク

Spoken Dialogue System for Robot with Computer Vision

Yoko YAMAKATA†, Tatsuya KAWAHARA†, and Hiroshi G. OKUNO†

† Graduate School of Informatics, Kyoto University
Yoshida-hommachi Sakyo-ku, Kyoto, 606–8501, Japan

E-mail: †{yamakata,kawahara,okuno}@kuis.kyoto-u.ac.jp

Abstract A spoken dialogue system is developed with the aim of creating a robot which searches for an object in the real world through interacting with the user. Speech and image recognition errors may occur within the system and differences among individual users may cause errors when translating the speech into a image representation. Misunderstandings may also occur due to false user beliefs. These problems are solved using a dialogue planning mechanism based on the probablistic framework of the belief network and a user model. We design and implement a system which searches for the object that is specified by the user but is not within the user's view. We demonstrate that this system can reduce the number of interactions for identifying on the object, and improves the speech recognition result by using the results of image recognition.

Key words spoken dialogue system, spoken language understanding, user model, belief network

1. はじめに

高度高齢化社会の到来に向け、介助ロボットなどの開発が望まれており、これに用いる音声対話システムの研究も進められている[1]。本研究では介助ロボットに求められる基本的なタスクの一つである、ユーザの発話に従いユーザの意図するオブジェクトを同定するというタスクを行うシステムを考える。具体的には以下のような機構を備える。

- ユーザの発話に対し音声認識・言語理解を行い、ユーザの意図したイメージモデルを特定する
- 目標オブジェクトの探索過程において生じる様々なあいまい性や誤解に対処する
- 言語理解におけるユーザの個人差をユーザモデルで扱い、対話を行うにしたがってユーザに適応するよう学習する

視覚情報のコミュニケーションタスクには、一部共通で一部不完全な対の地図を用いて、二人の人間が協調して地図を完成させる「日本語地図課題[2]」や、音声対話システム同士で出発地点から目的地までの電車の路線図を協調して作成する「DiaLeague[3]」などがあるが、いずれも本研究で扱うような機械対人間との音声対話ではない。また、ロボットに対する動作命令を音声認識・意味理解する「傀儡[4]」においては、ユーザとロボットが視覚情報を共有しているため、本研究で扱うような、オブジェクトや属性に関するあいまい性が少ない。

2. 章ではタスクとこれに対するシステムの処理の流れを述べる。3. 章ではあいまい性・誤解について述べ、4. 章でそれらを解決するためのプランニングについて、5. 章で信念ネットワークを用いた実現手法について述べる。6. 章で実験と評価について述べ、最後に 7. 章でまとめを述べる。

2. ユーザとロボットによる協調的なオブジェクト探索タスク

本研究で取り組むタスクは、例えば「隣の部屋の机の上にある赤いコップを持ってきて」といったユーザの発話をトリガとして、ユーザと音声で対話しつつ協調的にユーザの意図するオブジェクトを同定することである (図 1)。以降、このオブジェクトの存在範囲を対象世界と呼ぶ。対話の時点では対象世界はユーザの視野外であるが、ユーザは対象世界にどのようなオブジェクトがどう配置されているか、各オブジェクトの形や色などの情報を予め獲得しており、これに基づいて信念を形成している。オブジェクト探索はこのユーザの信念より動機付けられる。上記の例では、ユーザの欲求するインスタンスは「コーヒーカップ」であり、「赤色」であり、さらに「机の上」にあったという信念を持っている。ロボットはユーザの発話を音声認識し、言語情報に変換する。ここで、言語情報とは、ユーザの意図した目標オブジェクトの『名称』が「コーヒーカップ」であり、『色』は「赤」であり、『対象世界』は「机の上」であるというような、属性と単語の対である。

次にロボットは『対象世界』をカメラで撮影し、その画像を認識することにより『対象世界』に存在するオブジェクトを認識する。画像認識の結果はイメージモデルとして得られる。イメージモデルとはオブジェクトの視覚的特徴を記述したものであり、テンプレートマッチングにおけるテンプレートに相当し、特徴量で表現される。

上記の言語情報と視覚情報を統合するため、ロボットはユーザの発話から得た言語情報が、どのイメージモデルを指し示しているかを解釈する (図 2)。これは本研究における言語理解に相当し、後述するユーザモデルを用いて実現する。こうして推定したイメージモデルは目標オブジェクトの一つの属性を特定するものであり、「コーヒーカップ、赤」というように一つ以上の属性によって目標オブジェクトを指定する場合は、Dempster-Shafer の統合規則を用いて情報を統合する。その結果、最も尤度の高いオブジェクトをユーザの意図した目標オブジェクトと推定する。

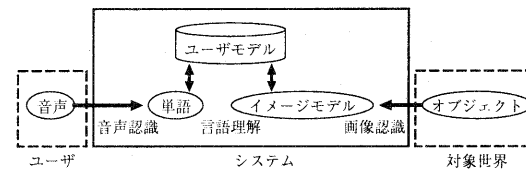


図 2　言語情報と視覚情報の統合

3. 対話におけるあいまい性・誤解の分類

音声対話システムにおけるあいまい性や誤解を扱った研究は多数ある[5]。しかし本研究では視覚情報を話題の対象とするため、発生するあいまい性・誤解の性質が大きく異なる。具体的に本タスクで生じるあいまい性・誤解は次の通りである。(1) 音声認識や画像認識において発生する認識誤り。(2) 言語理解における個人差によるあいまい性。例えば全てのユーザが「コーヒーカップ」と「ティーカップ」を明確に区別しており、その判定がまったく同一であるという仮定は現実的ではない。ある程度共通の理解が仮定できるが、個人あるいは家庭で独自の理解が存在し、その違いを無視してユーザの発話意図を理解することは難しい。(3) ユーザの信念の誤りによる誤解。これはユーザの記憶間違いや、ユーザの知らない間に対象世界が変化し、結果として対話の時点で対象世界の状態とユーザの信念とが食い違うことにより生じる。この場合、システムはユーザの発話を正しく理解しても、対象世界からユーザの意図するオブジェクトを同定できない。ユーザの視野外を話題の対象とする際には、ユーザは信念を更新することができないため、このような問題は避けられない。

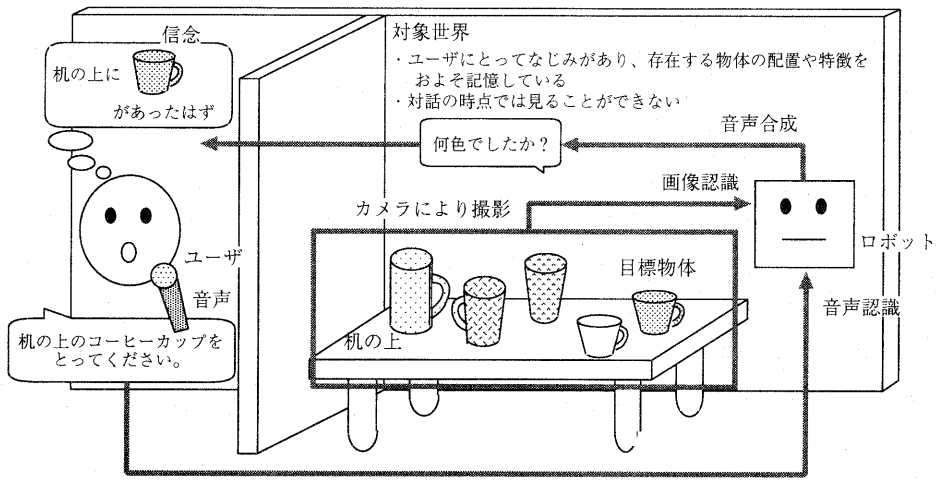


図1　ユーザとロボットによる協調的なオブジェクト探索タスク

4. あいまい性・誤解解消のためのプランニングレベル

前節で挙げたあいまい性・誤解を解決するため、システムを以下のような3種のプランニングレベルに分類する。

（1） システムのみによる解決
（2） ユーザとの確認対話によるあいまい性の解消
（3） ユーザの信念の誤りへの対処

実行するプランニングのレベルは上記の順序に従い、あいまい性や誤解が解決できた時点で終了とする。以下で、具体的な処理内容を述べる。

4.1 システムのみによる解決 (レベル1)

レベル1では以下で述べるようなプランニングを行う。

4.1.1 音声認識・画像認識の相互作用

ユーザの信念が正しいと仮定すると、ユーザの発話にしたがって対象世界を探索すれば指示通りのオブジェクトが見つかるはずである。そこで、音声認識結果を手がかりに、画像認識に用いるイメージモデルを絞り込む。これは例えば「コーヒーカップを探せと言われたのだからコーヒーカップのイメージモデルを用いて対象世界を画像認識する」といったものである。逆に目標オブジェクトの存在する場所を正しく画像認識し、そこにどのような物体が存在するかがわかれば、ユーザの発話内容は限られ、音声認識の候補を絞り込むことができる。これは例えば、目標オブジェクトの存在場所が「机の上」と特定されており、「机の上」には「コーヒーカップ」と「グラス」しかない場合、それらが発話される事前確率を高く設定する。音声・画像の両認識結果を関連づけるのが言語理解であり、言語理解の確信度や、絞り込みの手がかりとなる方の認識結果の信頼度に応じて、絞り込みの度合いが調節される。

4.1.2 ユーザモデルを用いた言語理解

ユーザに適応した言語理解を実現するためユーザモデルを用いる。このユーザモデルは一般的なモデルをベースとし、特定のユーザがタスクを遂行していく中でそのユーザに適応するよう学習していく機構を備える。このユーザモデルは発生してしまった誤解を解消するものではないが、ユーザとロボットの言語理解の相違を軽減する。

4.1.3 レベル1からレベル2への移行

システムのみではあいまい性が解消できなかった場合、レベル2へと移行する。具体的には、音声・画像認識の信頼度や発話理解の確信度が低かった場合や、推定されたオブジェクトの尤度が低い、あるいは同程度の尤度を持つオブジェクトが複数存在し、解を一意に決定できない場合である。

4.2 確認対話によるあいまい性の解消 (レベル2)

レベル2では以下のようなプランニングを行う。

4.2.1 確認発話

音声認識の信頼度が低い場合は認識結果を確認する。また、複数の候補が選定された場合は選択要求を行う。

4.2.2 探索条件の追加要求

レベル1で解が一意に決定できない場合や、ユーザモデルが未熟である (レベル1での言語理解の確信度が低い・対話回数が少ない) 場合に、未だ指定されていない属性についてユーザに質問することにより、条件の追加要求を行う。

4.2.3 レベル2からレベル3への移行

あいまい性を解消するため有効であると考えられる質問対話を全て行ったにも関わらず、該当する解を発見できなかった場合、ユーザの信念を疑うレベル3へと移行する。

4.3 ユーザの信念の誤りへの対処 (レベル3)

レベル3では次のようなプランニングを行う。

4.3.1 代替案の提示

ユーザの提示した条件との尤度が最も高いオブジェクト

を解の候補とするが、その際、ユーザモデルのミスマッチの可能性を考え、レベル2のときとは異なる言語表現を用いてユーザに提示する。

4.3.2 妥協点の推定

代替案に対するユーザの返答から、目標オブジェクトに対する属性の優先順位を推定し、妥協点を探る。例えばユーザの「赤いコーヒーカップを探して」という発話に対し、ロボットが「青のコーヒーカップでもいいですか?」と代替案を提示した際、ユーザがこれを受け入れなければユーザは目標オブジェクトが「赤」であることに自信を持っていると判断し、『色』属性の重みをあげてオブジェクトの再探索を行う。

5. 信念ネットワークによる言語理解

以上で述べた理解のシステムを信念ネットワークを用いて実現する。以下にその詳細を述べる。

5.1 単語とイメージモデルとの関連度

目標オブジェクトを指し示す属性は以下のものとする。

- 名称(「コーヒーカップ」「ワイングラス」等)
- 色(「赤い」「ブルーの」等)
- 模様(「無地の」「花柄の」等)
- 形状(「丸い」「四角い」「大きい」「小さい」等)

本稿ではこのうち、『名称』に属する単語の言語理解におけるあいまい性を主に扱う。『名称』に含まれる語彙はシソーラスを用いて「コップ類(カップ、グラス、ジョッキ…)」「皿類(丸皿、角皿、グラタン皿…)」のようにカテゴリわけしている。語彙は各カテゴリに対し20語程度である。イメージモデルは図3のような線画のイラストで表されており、言語情報と同じカテゴリにわけられている。ユーザモデルはカテゴリごとに、図4のようにイメージモデルと単語の対応づけ(関連度)のテーブルからなる。一般化ユーザモデルは、各項目について多数の被験者に対してアンケートを取り、その平均をとることにより作成する。

図3 「コップ類」におけるイメージモデル

単語 \ イメージモデル	モデル1	モデル2	モデル3
カップ	0.4	0.4	0.4
グラス	0.9	0.2	0.9
ジョッキ	0.4	0.9	0.4
湯飲み	0.4	0.1	0.9

図4 「コップ類」におけるユーザモデル

5.2 認識の信頼度・ユーザモデルにおける関連度

音声認識結果には単語単位の信頼度が付与されている。これはN-bestの解それぞれに対して認識器によって付与されているスコアから算出した事後確率である。単語L_lについての音声認識の信頼度を$CMS_l(\sum_n CMS_n = 1)$とする。画像認識はハンド・シミュレーションにより、各オブジェクトに対してイメージモデル単位の尤度を付与し、これを画像認識の信頼度とする。モデルM_mとオブジェクトO_oとの画像認識の信頼度は$CMI_{mo}(0 \leq CMI_{mo} \leq 1)$とする。単語$L_l$とイメージモデル$M_m$との関連度は$w_{lm}(0 < w_{lm} < 1)$で表されており、表4のクロステーブルの各欄の値に相当する。

5.3 オブジェクトの尤度

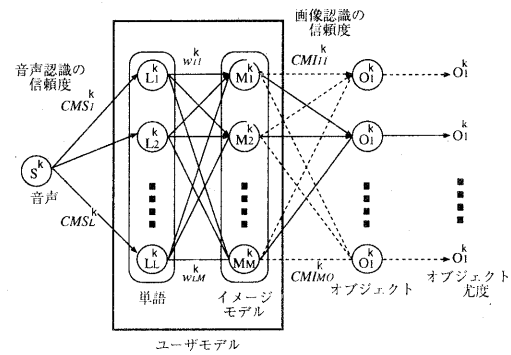


図5 属性kに対する各オブジェクトの尤度

ある属性kに関する信念ネットワークは図5のようなものである。ここでS^kはユーザの発話のうち属性kを表す単語を発声した部分の音声、L_l^kはその認識結果である単語、M_m^kはイメージモデル、O_o^kは探索空間中でロボットが識別することのできたオブジェクトを示す。以降、属性単位の尤度算出手法について述べる。

まず音声S^kが単語L_l^kである確率$Bel(L_l^k|S^k)$は、

$$Bel(L_l^k|S^k) = CMS_l^k$$

であり、モデルM_m^kである確率$Bel(M_m^k|S^k)$は、

$$Bel(M_m^k|S^k) = \frac{\sum_p w_{pm}^k Bel(L_p^k|S^k)}{\sum_q \sum_p w_{pq}^k Bel(L_p^k|S^k)}$$

である。発話S^kがオブジェクトO_oを指し示す尤度$Bel(O_o|S^k)$は以下の通り。

$$Bel(O_o|S^k) = \sum_p CMI_{po}^k \cdot Bel(M_p^k|S^k)$$

最後に複数の属性を統合してオブジェクトの尤度を算出するために、Dempster-Shaferの統合規則を用いる。統合後の尤度を$Bel(O_o)$とする。

5.4 ユーザモデルの学習

目標オブジェクトと断定したオブジェクトについて、複数の属性を統合して最終的に得られた尤度が真の尤度と考え、各属性ごとにその属性のみで算出した尤度との差をユーザモデルに伝播する (図 6)。音声認識の信頼度 CMS_l が高い単語と、画像認識の信頼度 CMI_o が高いイメージモデルは他の認識結果より確からしいので、それらの関連度 w_{lm} をより強く学習させたい。そこで次のような学習を行うこととする。ここで α は学習率である。

$$\Delta w_{lm}^{k} = \alpha(Bel(\mathrm{O}_o) - Bel(\mathrm{O}_o|S^k))\frac{CMS_l^k}{\sum_p CMS_p^{k\,2}}\frac{CMI_{mo}^k}{\sum_q CMI_{qo}^{k\,2}}$$

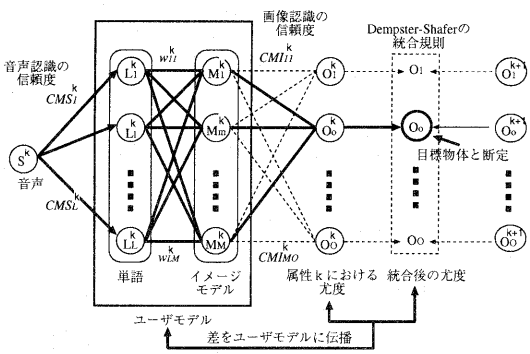


図 6 ユーザモデルの学習

5.5 音声・画像の相互作用

モデル M_m に対応するオブジェクトが存在する確率 $Bel(\mathrm{M}_m)$ を次のように算出する。

$$Bel(\mathrm{M}_m) = \max_o CMI_{mo}$$

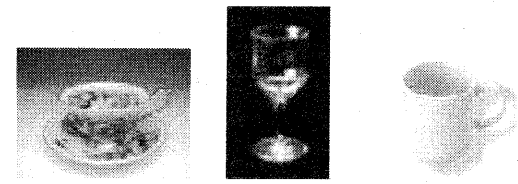
単語 L_l に対応するオブジェクトが存在する期待値 $Bel(\mathrm{L}_l)$ は、

$$Bel(\mathrm{L}_l) = \frac{\sum_p w_{lp} Bel(\mathrm{M}_m)}{\sum_p w_{lp}}$$

となり、これより単語 L_l の真の音声認識の信頼度 $Bel(L_l^k|S^k, O_1, ..., O_O)$ を、

$$Bel(L_l^k|S^k, O_1, ..., O_O) = \frac{Bel(\mathrm{L}_l)\cdot CMS_l}{\sum_p Bel(\mathrm{L}_p)\cdot CMS_p}$$

として算出し、認識結果 L_l をユーザに確認するかどうかの判断に用いる。また、$Bel(\mathrm{L}_l)$ を単語 L_l の事前確率と考え、認識対象語彙の N-gram モデルに利用すれば、ユーザの発話内容を予測した音声認識が実現でき、認識率の向上が期待できる。



オブジェクト 1　オブジェクト 2　オブジェクト 3

図 7 対象世界内に存在するオブジェクト

5.6 発話理解の確信度

ユーザの発話から意図するオブジェクトのイメージモデルを十分高い尤度で選定することができれば、ユーザの発話は理解できていると考えられる。そこで発話理解の確信度を $\max_m Bel(\mathrm{M}_m^k|S^k)$ とし、レベル 1 からレベル 2 への移行における判断基準とする。

6. 実　　験

以上のシステムを実装し、実験を行った。音声認識には Julian3.2 を用いた。各種パラメータについては次の通りである。まず解のあいまい性の判定においては、発話理解の確信度と解の尤度が 0.3 以上、1 位と 2 位の候補の尤度の差が 1 位の尤度の 10%以上開いていることが必要であるとし、その場合に限り 1 位の候補を解と断定した。また音声認識結果は、0.8 以上で確認せずに受理、0.6 以上でユーザに確認、0.6 以下は棄却とした。

対象世界内には図 7 のような 3 つのオブジェクトを用意した。一般的な呼び名に従えば、オブジェクト 1 は黄色で花柄のティーカップ、オブジェクト 2 は無色のワイングラス、オブジェクト 3 は白のマグカップである。ここで、あるユーザはオブジェクト 1 を「コーヒーカップ」と呼ぶというシナリオに則って、「コーヒーカップを取ってください」という発話により、ユーザの意図するオブジェクトであるオブジェクト 1 を探索する実験を行った。

一般化ユーザモデルに従えば、「コーヒーカップ」に対して尤度が最大となるのはオブジェクト 3 であるため、一回目の探索ではオブジェクト 3 が最大尤度となった。しかし尤度が低く発話理解の確信度が十分でないので、システムはユーザに対し「色」条件の追加要求を行った。これに対しユーザは「黄色」と応答することにより、オブジェクト 1 が他の候補に比べて高い尤度を示したため、解のあいまい性がなくなりオブジェクト 1 が選定された。システムはこの結果を元にユーザモデルを学習した。

学習率を 1 としたときの学習回数と各オブジェクトの尤度推移を図 8 に示す。形状が似たオブジェクトであるオブジェクト 1 とオブジェクト 3 の尤度は大きく変化して正しい尤度順序に移行しているが、形状がまったく異なるオブジェクト 2 についてはそれほど変化しない。次に、学習率を変化させたときの学習回数とオブジェクト 1 の尤度の関係を図 9 に示す。今回設定したパラメータにおいては、学習率が 1.2 で

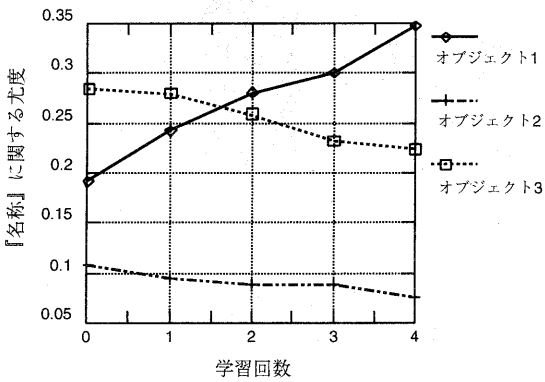


図 8　各オブジェクトの尤度の推移

は探索回数 3 回目、1 では 4 回目、0.8 と 0.6 では 5 回目以降、条件の追加要求を行わないという結果が得られた。ユーザモデルの学習を行わなければ正しい解を導くために常に条件の追加が必要なため、ユーザモデルにより対話回数を削減することができたといえる。

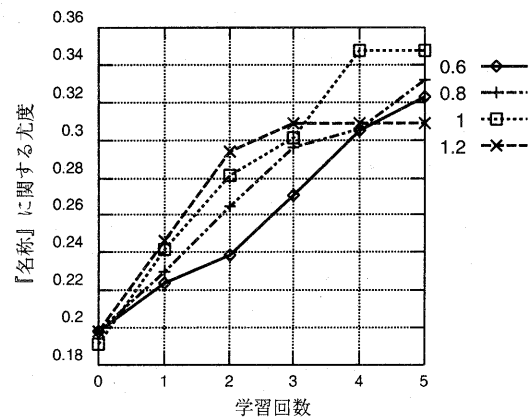


図 9　学習率における変化

音声・画像の相互作用の効果を調べるため、実験中に生じたある認識結果 (コーヒーカップ [0.39]、カップ [0.08]、コップ [0.228]、ティーカップ [0.302]) に対して、一般化ユーザモデルと学習後のユーザモデルによりそれぞれ絞り込みを行った。学習率が 1 のときの学習回数と音声認識の信頼度の関係を図 10 に示す。これより、一般化ユーザモデル (学習回数 0) では絞り込みが誤って行われ、「コーヒーカップ」の信頼度が減少するが、学習が進むにしたがい上昇していくことがわかる。これは画像認識の結果を用いて音声認識の結果を絞り込むには、ユーザモデルが必要であることを意味している。

本稿では言語理解における個々人の差異のみに注目し、状況や文脈による変化には触れなかった。しかし例えば「コーヒーカップではなくてコップをとって」という発話に対しては「コーヒーカップ」と「コップ」との違いに注目すべきで

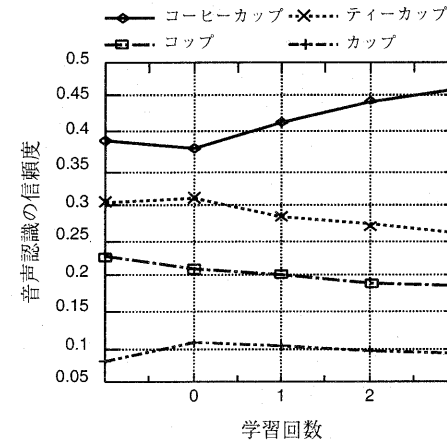


図 10　ユーザモデルの学習による音声認識信頼度更新の遷移

あるし、対象世界に互いに類似したものが多く存在する場合と、そもそも該当するものが一つしかない場合とではユーザの発話に対する解釈も異なるべきである。このような状況に依存したユーザモデルも考慮しなければならない。

7. ま と め

本稿ではユーザの発話の意図する属性・オブジェクトを正しく理解することを目標とし、信念ネットワーク及びユーザモデルを導入し、柔軟でユーザに適応的な言語理解を行うシステムを設計した。これを、ユーザの視野外に存在する、ユーザの意図したオブジェクトを、ユーザとの音声のみの対話により協調的に同定するタスクで実装した。実験では一般的な理解と違う言語理解を行うユーザを想定し、そのシナリオにおいてもシステムが正しくユーザの意図するオブジェクトを探索できることを検証した。状況や文脈に依存した言語理解を行うシステムを実現することが今後の課題である。

文　献

[1] 小梨貴司, 鈴木基之, 牧野正三. 介助ロボット用音声対話システム. 電子情報通信学会技術研究報告, SP2001-78, 2001.

[2] 青野元子, 市川熹, 小磯花絵, 佐藤伸二, 仲真紀子, 土屋俊, 八木健司, 渡部直也, 石崎雅人, 岡田美智男, 鈴木浩之, 中野有紀子, 野中啓子. 地図課題コーパス (中間報告). 情報処理学会研究報告, 94-SLP-3-5, 1994.

[3] 橋田浩一, 伝康晴. 人間との対話による対話システムの評価について. 人工知能学会研究会資料, SIG-SLUD-9701-3, 1997.

[4] 田中穂積. 言語理解とロボットの行動制御—音声認識から音声理解へ—. 電子情報通信学会技術研究報告, SP2000-81,NLC2000-33, 2000.

[5] Taro WATANABE. Masahiro ARAKI and Shuji DOSHITA. Evaluating dialogue strategies under communication errors using computer-to-computer simulation. In *IEICE TRANS.INF & SYST.*, IEICE TRANS. INF. & SYST., Vol.E81-D No.9 p.1025, 1998.