社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

階層的強化学習を適用したPOMDPによるカーナビゲーションシステムの音声対話制御

岸本　康秀†　　滝口　哲也††　　有木　康雄††

† 神戸大学大学院工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学自然科学系先端融合研究環　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †kishimoto@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　本稿では，カーナビゲーションシステム（以下カーナビ）における音声インターフェースに対して、部分観測マルコフ決定過程（POMDP）を用いる。この手法は不確定な情報に対しても対話を制御することが出来，雑音状況下で誤認識が起こった場合でも，自然な対話の中で回復することが可能となる．また，本研究では POMDP に階層的強化学習を適用することにより，従来の POMDP よりも大きなタスクを扱うことが可能となった．本稿では，シミュレーション実験を行い，提案手法の有効性を示す．
キーワード　音声対話システム，POMDP，階層的強化学習

Spoken Dialogue Manager in Car Navigation System Using Partially Observable Markov Decision Processes with Hierarchical Reinforcement Learning

Yasuhide KISHIMOTO†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of Engineering, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
†† Organization of Advanced Science and Technology, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †kishimoto@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　In this paper, we propose a dialogue manager in a car navigation systems using Partially Observable Markov Decision Processes(POMDP) that can treat ambiguous information. Even when it occurs speech recognition errors are caused by car indoor noises, it can manage the dialogue. we also propose a variation of the classic POMDP by incorporating hierarchical reinforcement learning. It can deal with large task than traditional system. The results confirms that the proposed method outperforms a handcrafted dialogue manager.
Key words　spoken dialogue system, Partially Observable Markov Decision Processes, hierarchical reinforcement learning

1. はじめに

音声対話システムにおける対話管理は，マルコフ決定過程（MDP）という確率的なモデルによって優れた対話戦略を得ることが出来た．しかし，MDP では環境の状態観測は完全であることが仮定されており，現在までのところでは，MDP は雑音がある状況下や曖昧な発話が行われた場合、うまく機能しない．現実の世界では，人は雑音状況下で発話し，発話が曖昧なことも頻繁に起こりうる [1]．この点から，現在の音声対話システムの最も重要な課題の１つとして，頑健性が上げられる．

近年，音声認識をベースとしたインターフェースを備えたカーナビゲーションシステムの実用化が進んでいる．ドライバーの負荷を軽減し，運転の安全性を向上させる

ため，カーナビの操作方法の一つとして音声認識が取り入れられて久しい．しかし，車室内雑音、発話変形、発話誤りにより誤認識が起こり，その結果として，誤動作が生じ，自動車内で音声認識を利用するユーザが増えない理由となっている [2] [3]．また，現在のインターフェースが未熟であるため，誤認識からの回復も困難となっている．

本稿では，不確定な情報に対しても対話を制御できるようにするために，カーナビの音声対話に部分観測マルコフ決定過程（POMDP）を用いる．この手法では，複数の状態仮説に対して信念という確率分布を保持しており，雑音下で誤認識が起こった場合でも，自然な対話の中で回復することが可能となる．この手法はもともとロボットの制御などの分野で適用されてきたが、対話の制御では、POMDP の状態数の増加に伴い，強化学習の計算量が増大するため今まで大きなタスクを扱うことが困難であった [4]．

本稿では，POMDP の強化学習に階層的強化学習を適用する．階層的強化学習では，強化学習の問題を階層的に分解し，各部分問題に対して局所的な政策を学習してから，それらを統合することによって大域的な政策を学習する [5] [6]．評価実験として，シミュレーション実験を行い、有効性を示す。

以降の 2 章では POMDP について述べ，3 章では強化学習について述べる．4 章で，評価実験について報告し，最後に 5 章で，結論と今後の課題について述べる．

2. POMDP

一般的に POMDP は $\{S,\ A,\ T,\ O,\ Z,\ R\}$ で表される．$s \in S$ は状態を表す（システムとユーザの状態）．$a \in A$ はシステム側のアクションを表す．また T はアクション a によって，状態 s が s' へ変わる状態遷移確率 $P(s'|s,a)$ の集合である．$o \in O$ はユーザや環境から観測される観測値を表し，Z はアクション a によって状態が s' に遷移した後，観測値 o' が観測される観測値出力確率 $P(o'|s',a)$ の集合である．$r(s,a) \in R$ は, 状態 s でアクション a を行った時の期待報酬を表す.

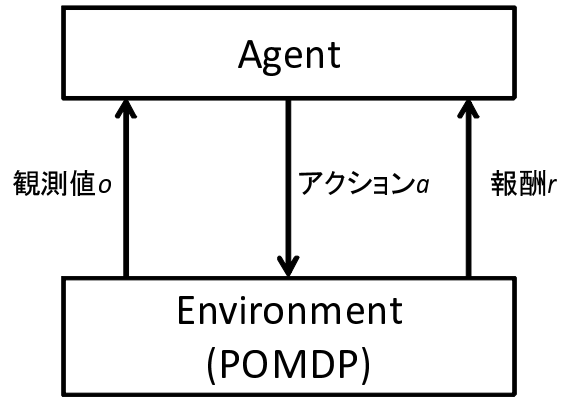

図 1　POMDP のサイクル

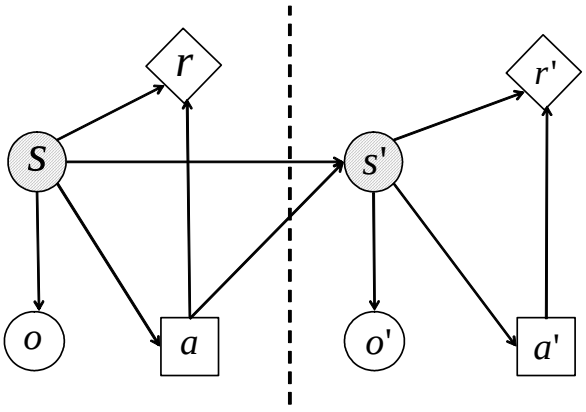

図 2　DBN を用いた POMDP の影響図

ここで、POMDP のサイクルについて図 1 を用いて述べる．まず，システムがアクション a を実行すると，内部状態 s が変化する．この内部状態の変化により，観測値 o が観測される．そして，アクション a と状態 s で規定されている報酬 r を得る。これを対話が終了するまで繰り返す。それぞれの変数の関係をダイナミックベイジアンネットワークで表すと図 2 のようになる。

POMDP では状態を直接観測出来ないので，確率分布として扱う．その分布を $b(s)$ とする．この分布 $b(s)$ が既知の時，アクションによって次の時刻の分布 $b'(s')$ は次式で表される．

$$b'(s') = k \cdot P(o' \mid s', a) \sum_{s \in S} P(s' \mid s, a) b(s) \qquad (1)$$

ここで k は $b'(s')$ の総和を 1 にするための正規化係数である．これを用いると，システムが時刻 t までに得る割引報酬は次式で表される．ここで Υ（< 1）は割引率を表す．割引率は，将来の報酬が現在においてどれだけの価値があるかを決定する．

$$V_t = \sum_{\tau=1}^{t} \Upsilon^{(\tau-1)} \sum_{s} b_\tau(s) r(s, a_\tau) \qquad (2)$$

POMDP の学習では，(2) 式を最大にするような方策を強化学習により求める．方策は，将来獲得できる報酬を最大にするアクション a を時間に独立に信念分布 b のみから選択出来る．

MDP と違い，POMDP の方策は連続な多次元変数の関数になり，最適な価値関数は図 3 の太線部のように，信念空間において凸型の多面体関数 (piecewise linear and convex) で表される．凸型となる原因については，直観的に以下のように説明できる．信念状態空間の中央部付近の領域は，エージェントが現時点の状態認識について区別のつかない状況を示しており，そこではあまり適切な行動選択が出来ないため，価値関数が下に凸になる．最適な価値関数はいくつかのベクトル集合 v_i で表される．そのベクトルは行動 $a(i) \in A_m$ と関係し，$v_i(s)$ は状態 s

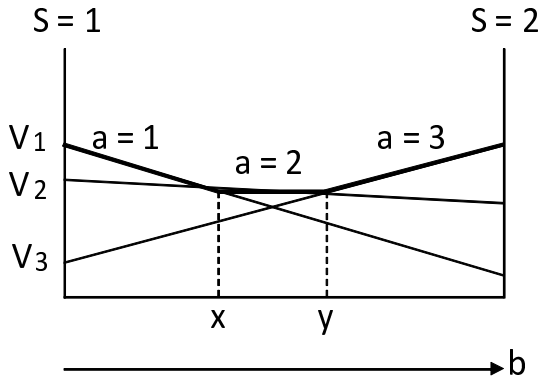

図 3　価 値 関 数

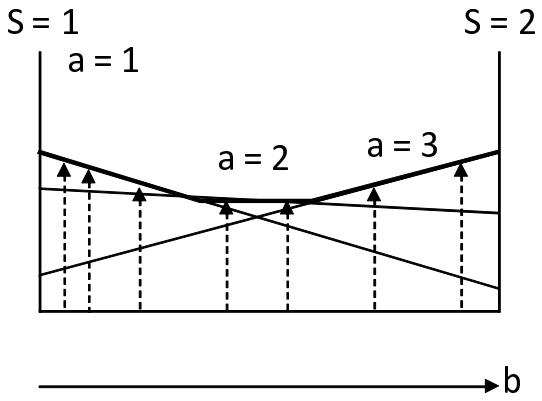

図 5　PBVI

でアクション $a(i)$ を取った時の期待報酬を表している．価値関数を構成する上で完全なベクトル集合を与えられた時，最適な価値関数と、それに相当する方策は次式で表される．

$$V^{\pi^*}(\mathrm{b}) = \max_i \{v_i.b\} \tag{3}$$

$$\pi^*(\mathrm{b}) = a(\operatorname*{argmax}_i \{v_i.b\}) \tag{4}$$

図 3 のように $|S| = 2$ の例で説明する．この時，価値関数は 3 つのベクトルで構成され，太線で表されている．この 3 つのベクトルは信念空間を 3 つの領域に分割しており，そのそれぞれに最適な行動と対応している．図 3 の例では，$b < x$ の時，V_1 が最適価値関数となり，アクション $a = 1$ が選択される．同様に，$x < b < y$ の時には V_2 が最適価値関数であり，$a = 2$ が選択される．

厳密な最適価値関数は価値反復という手法を用いることで求めることが出来る [7]．価値反復は図 4 のように全ての状態と行動の遷移を考慮する動的計画法の 1 つである．動的計画法では，まず終端における部分問題 $V_t^*(s_t)$ の部分問題を解き，再帰的に $V_{t-1}^*(s_{t-1})$，$\cdots$，$V_1^*(s_1)$，$V_0^*(s_0)$ の部分問題を解くことにより価値関数を求める．しかし，動的計画法では全ての状態と行動の遷移を考慮するため，状態空間または行動空間が大きい場合は，厳密な価値関数を求めることは計算量的に実行不可能である．

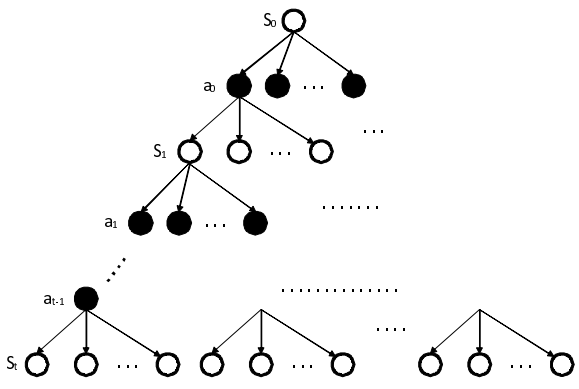

図 4　動的計画法

そこで本稿では，Point-based value iteration(PBVI) [8] [9] という手法を用いる．この手法は，全ての到達しうる状態に対して最適な方策を構築する代わりに，図 5 のように状態空間のいくつかの代表点を用意し，その点でのみ最適な方策を探索し，その方策に応じて状態空間全体を覆うベクトルを作成する．これらのベクトルを用いることで，状態空間上における任意の状態に対して，近似解となる方策を構築している．このように有限個の代表点に関してベクトルを作成することにより，状態数の組み合わせ爆発を回避する．

3. 強 化 学 習

強化学習とは，試行錯誤を通じて環境に適応する学習制御の枠組みである．エージェントに目標のみを与えておけば，エージェント自身が試行錯誤を繰り返し，そこに至るまでの行動を学習してゆく．そのため，以下のような利点がある．

- 目標までの行動を人間が知らなくて良い

エージェントは目標までの行動を試行錯誤によって見つけ出すため，人間が目標までの行動を知っておく必要はない．

- タスク遂行のためのプログラミング強化学習で自動化することにより，設計者の負担の軽減が期待できる．
- 人間以上の行動を見つけ出す可能性がある

未知空間では試行錯誤を繰り返し，目標までの行動を見つけ出す強化学習は，人間以上の最適な行動を見つけ出す可能性がある．

しかし，強化学習の問題点として，状態・行動数が多ければ多いほど，試行錯誤回数が指数関数的に増大するため，学習収束までにかかる時間も増大してしまう [10].

3.1　階層的強化学習

状態空間が巨大で複雑なとき，大域的な最適政策を一度に求めることが困難な場合がある．階層的強化学習では，複雑なタスクを階層的に分解し，まず各部分問題に

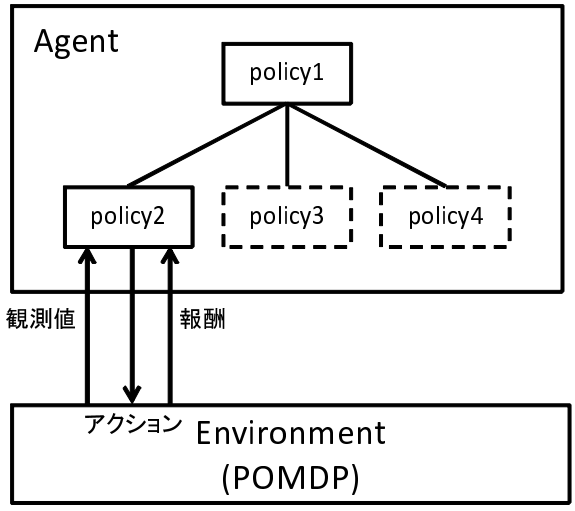

図 6　方策の組み合わせ

対して局所的な政策を学習してから，それらを統合することによって大域的な政策を学習する．巨大で複雑なタスクを，規模の小さい簡単な問題にうまく分解することが出来れば，複雑なタスクを効率よく解くことが出来る．

本稿では，小さなタスクに分解し，PBVI の手法を用いて，１つ１つのタスクを最適化し，それらを統合した．

3.2　階層的強化学習を適用した POMDP

カーナビでの音声操作において，入力における誤り（誤認識）が問題となることは周知である．この要因は，車室内雑音、発話変形、発話誤りの多さ等に起因する．その結果として，誤動作が生じ，自動車内で音声認識を利用するユーザが増えない原因となっている．

そこで本稿では，対話制御部に階層的強化学習を適用した POMDP を導入した．システムはユーザのゴールを確率分布により求めているので，たとえ誤認識を起こした場合でも，回復が可能である．

また，従来の音声認識システムでは，あらかじめ定められた定型句からなる音声コマンドしか受理できず，ユーザが前もって音声コマンドを正確に記憶していることを前提としており，音声入力に不慣れなユーザが直観的に音声だけでカーナビを操作することは困難である．そこで，POMDP において様々な発話を想定したユーザモデルを作成することにより、ユーザの多様な発話にも頑健に対応することが出来，直観的に使うことが可能となる．

本稿では，エアコンの調整，オーディオ操作，近隣の店舗検索を扱うタスクを設定した．対話の制御において，POMDP の状態数の増加に伴い，強化学習の計算量が増大し、大きなタスクを扱うことが困難であったが，今回図 6 のように複数の方策を組み合わせることにより，従来より大きなタスクを扱うことが可能となった．この時，policy1 のようにどのサブタスクを行うかを選択する層と，エアコンの調整など具体的なユーザのゴールを求めるサブタスクを行う層に分かれる．

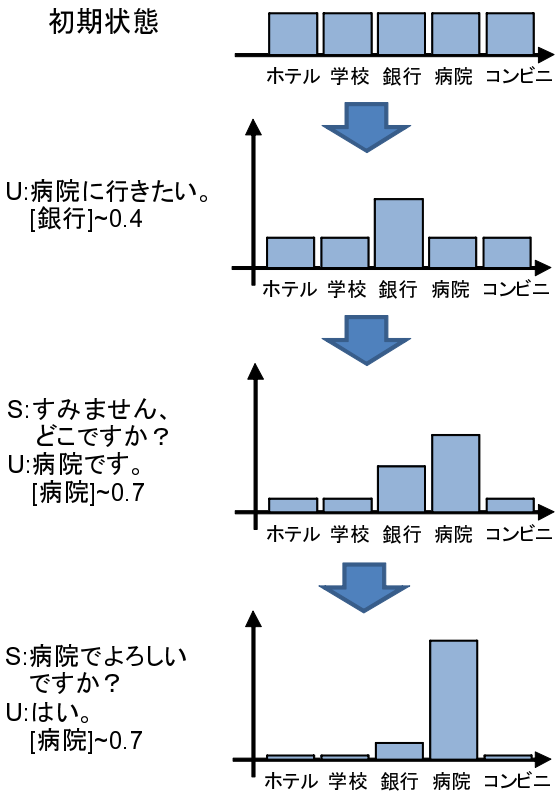

図 7　対 話 例 1

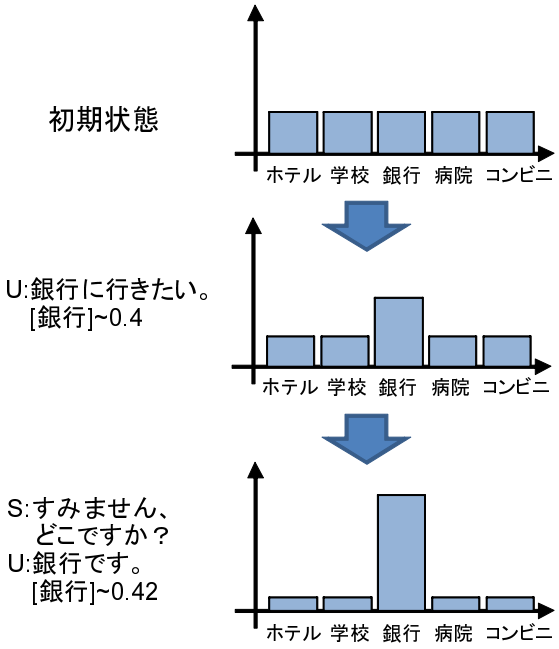

図 8　対 話 例 2

システムのアクションはユーザの発話について，聞き直す，確認を取る，ユーザのゴールを決定するの３種類がある．この時の報酬の設定は，ユーザのゴールを正しく求めることが出来れば大きな正の報酬を与え，誤ったゴールを求めた場合には大きな負の報酬を与える．対話が早く収束するために，聞き直す，確認を取るというア

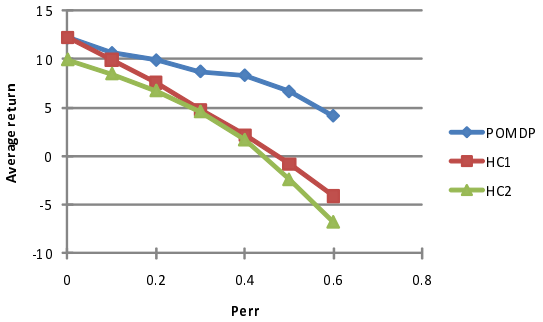

図 9　平 均 報 酬

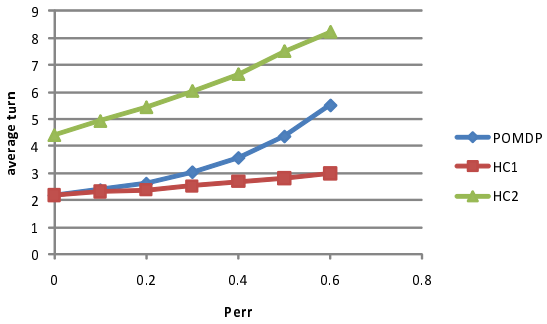

図 10　平均ターン数

クションを取った場合にも小さな負の報酬を与える．

具体的な対話例を図 7，図 8 に示す．対話例 1 では，1 ターン目に誤認識を起こしているが，信頼度が低く，もう 1 度聞きなおすことによって，正しいゴールを求めている．対話例 2 では，認識結果は正しいが，信頼度が低い．この時，従来のシステムでは閾値を越えない場合，信頼度の低い発話を何回繰り返しても認識結果は却下されてしまうが，POMDP の場合では信頼度が低い発話も信念の更新に使われ，たとえ信頼度が低い発話を繰り返したとしても，ユーザのゴールを正しく求めることが出来る．

4. 評 価 実 験

従来のカーナビと POMDP との比較実験を行うために，カーナビの音声認識システムを想定した方策（HC1）を作成した．この方策は，ユーザの発話がシステムのアクションに対して適切でない場合にはもう 1 度聞き直すというアクションを取るが，それ以外の時にはユーザの発話に対して確認を取らずにユーザのゴールを決定する．また，もう 1 つの比較方策として，方策（HC2）を作成した．この方策は，ユーザの発話に対して毎回確認を取ってユーザのゴールを決定する．

この 2 つの方策と POMDP の方策を用い，ユーザシミュレーションによる試行を 1000 回繰り返し、その平均をとって，1 回の試行で得られる平均的な報酬とした．試行ごとにユーザシミュレーションのゴールをランダムに設定し，対話が終了するまでを 1 回の試行とする。また，対話が終了するまでにかかる平均のターン数と正答率を算出した．

図 9 にコンセプト誤り率 Perr（100*{1-(正解発話数)/(総発話数)}）を 0.0 から 0.6 まで変えながらシミュレーション実験を行った結果を示す．結果から，コンセプト誤りが無いときには POMDP と HC1 は等しい報酬を得ているが，コンセプト誤りが増加すると POMDP の方策の方が HC1 より多くの報酬を得ていることが分かる．これは音声認識誤りがない時には，同等の性能を持っているが，認識誤りが増えると POMDP の方が HC1 に比べて頑健に動作していることを示している．それは，

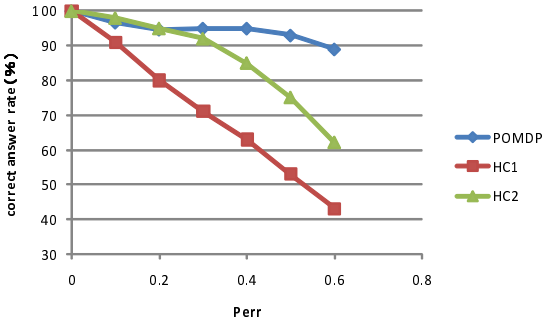

図 11　正　答　率

POMDP が常にユーザのゴールについて複数の仮説を保持しており，誤認識が起こった場合でもユーザのゴールを求めることが出来るからである．

また図 10 は，対話が始まって，終了するまでの平均ターン数を示している．コンセプト誤りが無い時には，POMDP と HC1 の平均ターン数は等しいが，コンセプト誤りが増加すると POMDP の方が平均ターン数が多くなることが分かる．また、HC2 はユーザの発話に対して毎回確認を取ることから，コンセプト誤りが無い時でも他の方策よりターン数が多く，コンセプト誤りが増えるにつれターン数も多くなっていることが分かる．また図 11 はそれぞれの方策の正答率を表している．コンセプト誤りが増えるにつれ，それぞれの方策の正答率は下がっているが，最も HC1 の正答率が低くなっている．また，HC2 と比べても POMDP の方が正答率が高いことが分かる．

これらの結果より，POMDP は音声認識誤りが多い環境において，従来のカーナビよりも頑健に動作すると考えられる．

5. お わ り に

本稿では，カーナビの対話制御部に POMDP を導入して，シミュレーション実験を行った．その結果，POMDP の方がより多くの報酬が得られた．このことから，従来のカーナビの音声認識よりも，POMDP を用いた対話制御部の方が，誤認識に対して頑健であると考えられる．

また，本稿では POMDP に階層的強化学習の考えを導入した．これにより，従来では最適化が難しかった複雑なタスクを階層的に分解し，各部分問題に対して局所的な政策を学習してから，それらを統合することによって大域的な政策を学習することが出来た．今後の課題として，よりタスク拡大を図るとともに，誤認識に強く，直観的に使える音声対話システムの実現に取り組んでいく．

文　献

[1] Nicholas Roy, Joelle Pineau, Sebastian Thrun,“ Spoken Dialogue Management Using Probabilistic Reasoning, ” ACL, pp. 93-100, 2000.

[2] 西村雅史,“ 音声認識ビジネスの現状と将来展望 ”，IPSJ, 2005-SLP-55-3，pp. 13-15 2005.

[3] 神沼充伸,“ 自動車用音声インターフェースへの期待 ”，IPSJ，2006-SLP-63-9-2，pp. 47-48 2006.

[4] Williams,J.D., Young,S.J.,“ Partially observable Markov decision processes for spoken dialog systems ”, Computer Speech and Language 21 (2),231-422.2009

[5] Dietterich, T.G,“ An Overview of MAXQ Hierarchical Reinforcement Learning ”, Lecture Notes in Computer Science, 1864, 26-44, 2000

[6] D.Andre, S.Russell,“ State Abstraction in Programmable reinforcement Learning Agents ”, technical report UCB//CSD-02-1177, Computer Science Division, University of California at Berkeley, 2002

[7] GE Monahan.“ A survey of partially observable Markov decision processes: Theory, models, and algorithms ”, Management Science, 28(1):1-16, 1982.

[8] J Pineau, G Gordon, and S Thrun,“ Point-based value iteration: An anytime algorithm for POMDPs, ” in Proc Int Joint Conference on AI (IJCAI), Acapulco, Mexico, pp. pp1025-1032 2003

[9] MTJ Spaan and N Vlassis,“ Perseus: randomized point-based value iteration for POMDPs, ” Tech. Rep., Universiteit van Amsterdam, 2004

[10] R.S. Sutton, A.G. Barto,“ Reinforcement Learning: An Introduction ”, MIT Press, 1988