1J1-OS13-5

マルチモーダル入力と強化学習による擬人化エージェントの対話制御の検討

Dialog Management using Reinforcement Learning for Multi-modal Spoken Dialog Agent

盧 迪[*1] Lu Di　久保 伸太郎[*1] Kubo Shintaro　深山 覚[*1] Fukayama Satoru　中沢 正幸[*1] Nakazawa Masayuki　西本 卓也[*1] Nishimoto Takuya　嵯峨山 茂樹[*1] Sagayama Shigeki

[*1]東京大学大学院　情報理工学系研究科
The University of Tokyo, Graduate School of Information Science and Technology

To make the dialog between the agent and the user smoother, we propose the use of POMDP for real-time dialog with reinforcement learning. Multi-modal inputs and intermediate speech recognition results during the utterances can be efficiently used with this POMDP framework, which makes barge-in control possible. Two experiments which evaluate the proposed methods were performed. We also implemented the prototype system that utilize the result of reinforcement learning.

1. はじめに

人間と機械の音声コミュニケーションのためには，インタフェースシステムを統合技術と捉えつつ，音声入出力の特質を踏まえ，マルチモーダルインタフェースとしての合理的な設計を行わなくてはならない．嵯峨山 [嵯峨山 1994] によってこのような問題提起がなされ，西本 [西本 1996] はマルチモーダルシステムのための「インタフェースの原則」を提案し，使いやすい音声インタフェースを試行錯誤に頼らず合理的に設計できることを示した．

擬人化音声対話エージェントを用いることの意義もまた「コミュニケーションの効率と質を高める」ことである．つまり，人間は相手の表情から反応を読むことができる．一方が話している間にも頷いたり首をかしげたり，聞き取りにくければ直ちに「え？」と聞き返すことができる [嵯峨山 2004] ．

音声入力の利用における問題のひとつは「インタフェースの透過性」である．人間同士のコミュニケーションの「分かっているのか分かっていないのか反応がある人とは会話がしやすい」という特長を生かすことは，音声インタフェースの有効な利用につながる．

もうひとつの問題は「音声認識の処理速度」である．一般に音声認識アプリケーションは，応答の遅れによって，ユーザに不満を与えたり不安を感じさせたりしている．これに対して，人間の対面コミュニケーションでは，相手が口を開いた瞬間に，あるいは何かを言い終わる前に，言いたいことが相手に伝わってしまうことさえある．話者同士の状況，相手の表情や仕草など，人間はさまざまなモダリティからリアルタイムに情報を得ている．

このような検討の末，以下の仮説に至った [Lu2009]：

仮説「マルチモーダル情報を常に受け取り，意味のある反応をリアルタイムに行う擬人化音声対話エージェントシステムは，効率的なインタラクション実現のために有効である.」

例えば，発話中の割り込みや聞き返しに対する制御，相槌や頷きの生成や応答などは，こうした仮説を支持する提案となり得る．しかしこのような制御モデルの構築は，個別の対話タスクに依存する複雑な問題である．

そこで我々は，効率的で円滑な対話戦略を機械学習によってエージェントに獲得させることを目指している．本報告では，近年普及してきた強化学習に基づく音声対話の枠組みを発展させ，マルチモーダル入力に拡張した汎用的な POMDP を用いて，エージェントの実時間制御を行うモデルを自動学習する手法を提案する．

連絡先: 盧迪，東京大学大学院　情報理工学系研究科，〒 113-8656 東京都文京区本郷 7-3-1，lu@hil.t.u-tokyo.ac.jp

2. 対話制御のモデル化

2.1 本研究の着眼点

従来，音声対話システムの制御には状態遷移機械やスロットフィリングなど決定論的なモデルが多く用いられてきた．現在 Galatea Toolkit [川本 2002] [Galatea] が採用している VoiceXML もこうした枠組みの技術である．

しかし例えば，Galatea ツールキットの音声認識モジュール Julius [Julius] は，高性能な汎用大語彙連続音声認識エンジンであり，数万語彙の連続音声認識を一般の PC 上でほぼ実時間で実行できて，段階的な認識結果を出力することができる．しかし，段階的な認識結果は信頼度が高くないため，決定論的なモデルではうまく扱えない．

一方で近年，音声対話システムの制御に，確率モデルに基づく機械学習を用いる提案がある [Williams2005][Williams2007]. 隠れマルコフモデル (HMM)，部分観測隠れマルコフ過程 (POMDP)，ベイジアンネットワークなどの手法に基づき，強化学習 [Sutton2000] を用いるこれらの手法は，頑健な対話制御を実現している．特に POMDP は，音声認識結果を不完全または部分的な観測として扱えるため，音声対話システムに有効とされる [南 2010][荒木 2010].

こうした機械学習の手法は，時々刻々観測される音声認識エンジンの段階的な認識結果や，顔認識などのマルチモーダル観測情報などを効果的に扱えると期待される．

2.2 強化学習

強化学習とは教師なしの機械学習の一手法である．ある行為を選択したとき，環境から得られる期待報酬を，すべての状態において実際に探索しながら求めることによって，状態と期待報酬を最大化するように行為をマッピングし，最適な戦略を学習していく．本研究においては，エージェントが環境との相互作用から学習して目標を達成することに相当する．

一般的には，強化学習はマルコフ決定過程として定式化される．しかし実環境中の対話システムにおいては，外界のノイ

ズなどの影響があり，観測状態に不完全性や不確実性があるため，部分観測マルコフ決定過程による定式化がなされる．

2.3 部分観察マルコフ決定過程

MDP（Markov Decision Process, マルコフ決定過程）は環境をモデル化するための枠組みである．POMDP（Partially Observable MDP, 部分観測マルコフ決定過程）は MDP を拡張したものである．MDP による対話のモデルでは，状態 s が確定的に観測可能であることを前提にしている．しかし，実環境の中の不完全性や不確実性を考慮するために，すべての状態の信念を確率分布で表わし，システムの行為や不確実な観測を通じて，その信念を更新してゆく過程での意志決定問題として対話制御を捉える必要がある．

POMDP では観測値から状態 s の分布 $b(s)$ を推定する．この状態の分布は一つ前の時刻の分布から求めることができる．現在の分布 $b(s)$ がわかっているものとすると，遷移確率および出力確率から次の時刻の分布 $b'(s')$ は以下の漸化式で表せる．

$$b'(s') = \eta \cdot P(o'|s', a) \sum_{s} P(s'|s, a)b(s) \qquad (1)$$

2.4 リアルタイム対話制御のための POMDP

従来の POMDP の使い方は，多くの場合スロットフィリングの対話タスクを対象としている．そのため，エージェントが学習するとき，発話交替するたびに一回状態を観察して，動作を決める．このような仕組みでは，段階的な観測やマルチモーダル観測を扱いにくいため，割り込みに対応しにくい，

実際の人間の対話を見ると，人間は相手が喋り終わるときに意味を考えはじめるのではない．相手が喋っている途中でも相手の状態を観察して，意味を理解し合理的な返事を考えている．擬人化エージェントのシステムにおいては，このような情報は，例えばユーザの発話の即時音声認識結果と顔の即時認識結果である．

合理性，必然性のあるリアルタイム制御を行うためには，このようなマルチモーダル観測の活用に加えて，タスク知識およびコンテクストの高度な利用が欠かせない．しかし，こうした振る舞いは規則によって記述することが困難である．またタスクによって規則を変えなくてはならない．

さらに，音声認識の途中で得られる情報は断片的かつ不完全である．即時に観察されたユーザの顔認識情報にも，信頼できる情報と信頼できない情報がありえる．このような即時情報の信頼性を考慮したモデルが有用である．

人間同士の対話では，目的や状況がはっきりしていれば，表情や発話の一部分から先回りして行動できる．仮に間違った判断をして間違った返事をしてしまっても，対話が効率的であったかどうかは，対話が終了する時にわかる．このような仕組みを実現するためには，強化学習が適している．

例えばエージェントがユーザに情報を提供する案内型の対話タスクにおいては，学習の目的は，対話を効率的に終了させることと，ユーザにすべての内容をきちんと伝えることである．そして，このような学習問題は，エージェントの利得（報酬の総計）を最大化する対話全体の方策を探索する，最適化問題に帰着される．

3. 実験１：マルチモーダル観測の利用

3.1 実験目的

実験 1 では，システムの前を歩行する人間の顔の向き，歩くスピードなど，人間の状態を多角的に用いて，システムからの対話開始のタイミングを判断するタスクを扱う．

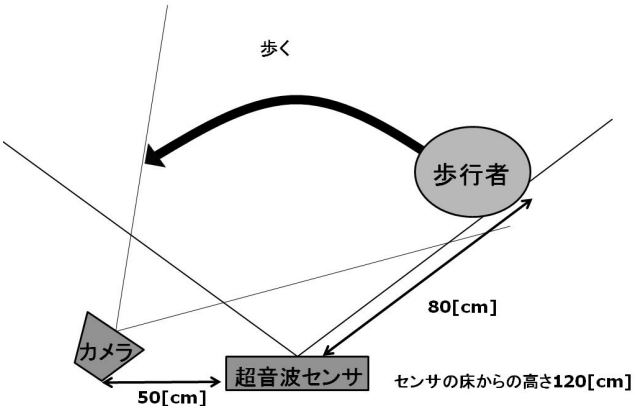

図 1: 実験 1 の概要

円滑な会話の実現のために，視覚情報などのノンバーバル情報の利用が注目されている．実験 1 ではこれがカメラと超音波センサによるユーザ情報の取得と強化学習に基づくアルゴリズムによって実現できることを示す．

カメラでの顔画像認識には OpenCV の顔検出器（Haar 分類器）を用いた．これにより，カメラから一番近いユーザの顔の中心位置の座標，顔の半径が検出可能である．

また，本研究では超音波センサ (スターアイ 2D センサユニット USSM2D-100) を使用した．これは空気中に超音波を照射し，人間など対象物体に当たって跳ね返ってきた反射波をアレイセンサが受信し，信号処理することでわずかな位相差を検出して，前方の物体の位置を距離と方向の 2 次元情報で知ることができる．空気中に超音波を照射する送波素子，反射波を受信する受波素子，2 次元画像化する信号処理 IC，を 70 × 35 [mm] の基板に収めてあり 5 [V] 単一電源で動作する．この超音波センサで歩行者の動作を検出する．

3.2 実験手順

本実験では，エージェントは 0.33 [s] ごとに行動を選択してその行動を行う設定とする．学習の進み方を解析するために 20 エピソードずつ 6 セット (学習の浅い方からセット 1 - セット 6 とする) に分けて合計 120 エピソード行った．

各エピソードで歩行者は「興味ある」,「興味ない」状態を想定してシステムの前を歩き，各エピソードの各ステップにおいて，エージェントが観測した歩行者の状態遷移と，その各状態でエージェントが起こした行動 (「話しかける」,「話しかけない」) に対して，歩行者が満足度（0 - 4）の報酬を与え，強化学習を行う．

3.3 結果と考察

強化学習の Q 関数の遷移の様子を図 2 に示す (上：初期値, 下：学習後)．16 状態× 2 での各 Q 関数の値をグレースケール化しており，白は値が大きいことを，黒は値が小さいことを示している．

各セット (20 回の学習) で与えられた報酬の頻度遷移の様子を図 3 に示す．各セットでの各報酬の出現頻度をグレースケール化しており，白は値が大きいことを，黒は値が小さいことを示している．学習後には報酬 0 の頻度が低くなっており，報酬 4 の頻度が高くなっていることが確認できた．

システムの設計者を被験者にするという設定の中で，ある程度推測に合致した期待通りの結果が得られた．当然ながら著者自身が被験者を務めたことで，提案手法の有効性の検証とし

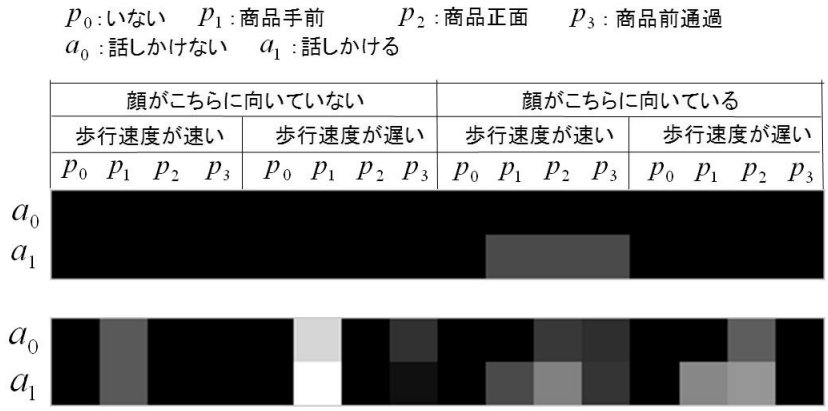

図 2: 実験 1 の結果 (Q 関数の遷移)

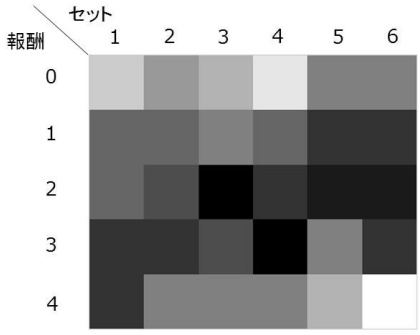

図 3: 実験 1 の結果 (報酬の頻度遷移)

ては課題が残っている．

本実験を通して

- 視覚情報を用いた歩行者への話しかけシステムの実装
- 実装者本人による学習データによって強化学習の過程

の 2 点を確認した．エージェントがカメラ，超音波センサによって歩行者の状態 (歩行者の顔認識，位置検出，歩行速度検出) を観測し,「話しかける」「話しかけない」という行動選択の試行錯誤を繰り返すという強化学習を行うことによって，適切なタイミングで歩行者に話しかけるシステムを実現した．

4. 実験 2：音声認識から段階的な観測の利用

4.1 実験目的

実験 2 のタスクは音声認識の段階的な観測を使ったリアルタイム対話制御である．POMDP に基づいて，信頼できる観察情報と信頼できない観察情報をうまく切り分けて，信頼できる情報だけを使って，人間の割り込み発話にシステムが適切に応答できることを目標とする．

4.2 実験内容：対話エージェントの実時間制御

エージェントがユーザに道案内を行うタスクを取り上げる．例えば屋外に置かれた電子案内板に擬人化エージェントの姿が現れて，立ち寄ったユーザと以下のような会話を行う．

- ユーザ「安田講堂はどこですか？」
- エージェント「まっすぐ行って，2 番目の交差点で，左に曲がって，右側にあります」

この例の中でユーザは「2 番目の」「左に」「右側に」といったキーワードを聞き漏らす可能性がある．その場合ユーザはエージェントに対して「え？」「もう一回？」などと聞き返す発話を行う．このように，エージェントは，相手が聞き返しを行う可能性を常に考慮する．

また，ユーザはエージェントの説明を理解できたときに，時々「うん」「はい」と自然に相槌をする．エージェントが相槌を知覚できた場合は,「自分の言いたいことが伝わっている」という合理的な解釈が可能である．このような前提において，具体的に「エージェントがユーザに長い説明を行い，ユーザは聞き取れなかった指示をエージェントに随時聞き返す状況」に着目する．

音声認識エンジンとして，段階的な認識結果を出力することができる Julian Ver.3.5.3-galatea を-progout オプションで使う．認識処理の第 1 パスで 300ms 周期で単語候補を取り出す．後述する処理では第 1 パスと第 2 パスの結果を区別せずに利用する．認識対象の発話は,「はい」と「もう一回お願いします」を含めて全部 21 種類である．

4.3 POMDP の詳細設計

POMDP の詳細な設計は以下のとおりである．

S をユーザの発話状態とする．A はエージェントの動作である．O はエージェントが観測できるユーザの状態であり，つまり認識できる 21 種類の発話である．そして現在の状態が s である確率を表わす信念を $b(s)$ と表現する．

POMDP の枠組みに基づいて，以下のように定義する．ただし説明を単純化するために，ユーザの発話は「はい」と「もう一回お願いします」の 2 つだけに限定する．

- 動作 A：{最初から言い直さない，最初から言い直す}
- ユーザの状態 S：{はい，もう一回お願いします}
- ユーザの観察情報 O：$\{o_1, o_2 ... o_{20}\}$
- 観測値から推定した状態の信念（確率分布）b (s)
- 観測確率 $p(o|s,a)$：音声認識の認識率を 0.6 にする
- 遷移確率 $p(s'|s,a)$：ユーザの意図はあまり変化しないと仮定して，0.6 にする
- 報酬 R：正しい対応はプラスの報酬，間違って対応はマイナスの報酬
- r(はい, 最初から言い直さない)=5
- r(はい, 最初から言い直す)=-10
- r(もう一回お願いします， 最初から言い直す)=5

- r(もう一回お願いします，最初から言い直さない)=-10

本研究の目標である「常に情報を受け取り意味のある動作を行う」というインタラクションを扱うために，発話交替を単位とした処理ではなく，時間間隔（インターバル）の概念を導入する．つまり，ある時間間隔 t のインターバルごとに行動の決定，状態の決定，報酬の評価を行う．今回は t は 0.5 秒周期とした．

各インターバルにおいて，エージェントはユーザの状態を観測し，ユーザの状態の信念を更新し，探索戦略によってある動作（発話など）を決定し，エージェントからの出力とする．そしてエージェントは報酬を受け取り，以下のような Q 学習が行われる．報酬が収束すれば学習を終了する．

- $Q(s,a)$ を初期化
- 各エピソードに対して繰り返し：
- s を初期化
- エピソードの各ステップに対して繰り返す：
- s 状態で ε グリーディ行動選択で行動 a を選択する
- 行動 a を取る r, s' を観察する
- $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')-Q(s,a)]$
- $s \leftarrow s'$
- s が終端状態ならば繰返しを終了

4.4 実験手順

実装者本人が被験者として，エージェントの道案内の説明を聞いて,「はい」と相槌をしたり,「もう一回お願いします」と割り込んだりする実験を行う．エージェントは 0.5 秒の周期でユーザの状態として音声の段階的な認識結果と最終的な認識結果を観測し，観測状態に応じて各ユーザ状態に関する信念を更新する．もしある発話の信念が 0.8 を超えたら，その発話はユーザの発話状態として選ばれる．強化学習によってユーザ状態に応じた対話の方策が学習され，エージェントの動作が選ばれる．学習率と割引率はそれぞれ 0.5 と 0.8 である．

4.5 結果と考察

学習が進むにつれて，ユーザの発話に対するエージェントの応答が変化した．例えば，ユーザの「はい」の発話を多くの場合に無視するようになった．また，ユーザの「もう一回お願いします」の発話に対して早い段階で（音声認識の第 2 パスの結果が出る前に）エージェントが言い直しを行うような対応が実現できた．これはユーザの発話（状態）とエージェントの行動に対する報酬の設定が適切であったためと考えられる．

現時点では音声認識結果が正しかったか否かのフィードバックは行っていない．今回の実験条件では比較的良好な認識性能が得られており，また POMDP における観測値からの状態推定の仕組みによって，300ms ごとの発話中の単語候補出力が一貫していないことに対しても，頑健な処理が実現されたためと考えられる．

さらに効果的なエージェント制御を行うために，被験者がエピソードごとに行うエージェントの振る舞いの評価，エピソード所用時間，要求された言い直しを十分に行えたかという観点での評価，などを報酬として用いることが有効と考えられる．

5. まとめ

本報告では，近年普及してきた強化学習に基づく音声対話の枠組みを発展させ，マルチモーダル拡張された汎用的リアルタイム POMDP を用いて，エージェントの実時間制御を行うモデルを自動学習する手法を提案した．

今後はマルチモーダル観測と音声認識の段階的観測のアプローチを総合し，POMDP の詳細化に力を入れる予定である．また，擬人化音声対話エージェントへの実装・評価を行う予定である．

参考文献

[嵯峨山 1994] 嵯峨山 茂樹: “なぜ音声認識は使われないか?どうすれば使われるか?,” 情報処理学会研究報告, 94-SLP-1, Vol. 94, No. 40, pp. 23-30, (1994)

[西本 1996] 西本 卓也, 志田 修利, 小林 哲則, 白井 克彦: “マルチモーダル入力環境下における音声の協調的利用 音声作図システム S-tgif の設計と評価,” 電子情報通信学会論文誌 D-II, Vol.J79-D-II, No.12, pp.2176-2183 (1996)

[嵯峨山 2004] 嵯峨山 茂樹, 西本 卓也, 中沢 正幸: “擬人化音声対話エージェント,” 情報処理学会誌, Vol.45, No.10, pp.1044-1049, (2004)

[川本 2002] 川本 真一, 下平 博, 新田 恒雄, 西本 卓也, 中村 哲, 伊藤 克亘, 森島 繁生, 四倉 達夫, 甲斐 充彦, 李 晃伸, 山下洋一, 小林 隆夫, 徳田 恵一, 広瀬 啓吉, 峯松 信明, 山田 篤, 伝 康晴, 宇津呂 武仁, 嵯峨山 茂樹: “カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計,” 情報処理学会論文誌, vol.43, no.7, pp.2249-2263, (2002)

[Galatea] http://sourceforge.jp/projects/galatea/

[Williams2005] Williams, J.D., Poupart, P., Young, S.J., “ Factored partially observable Markov decision processes for dialogue management, ”Proc. Workshop on Knowledge and Reasoning in Practical Dialog Systems, Int. Joint Conf. on Artificial Intelligence (IJCAI), Edinburgh. (2005)

[Williams2007] Jason D. Williams, Steve Young: “ Partially observable Markov decision processes for spoken dialog systems, ”Computer Speech and Language, Volume 21 , Issue 2, pp. 393-422 (2007)

[Sutton2000] Richard S. Sutton, Andrew G. Barto（三上 貞芳, 皆川 雅章 訳）: 強化学習, 森北出版 (2000)

[Lu2009] 盧 迪, 中沢 正幸, 西本 卓也, 嵯峨山 茂樹: “擬人化エージェントとの円滑なマルチモーダル対話のための強化学習を用いた割り込み制御の検討,” HAI シンポジウム 2009 予稿集, 2009

[Julius] http://julius.sourceforge.jp/

[南 2010] 南 泰浩: “強化学習による対話制御,” 日本音響学会講演論文集 3-6-12, Mar 2010.

[荒木 2010] 荒木雅弘: “機械学習による対話制御－その必要性と可能性－,” 日本音響学会講演論文集 3-6-11, Mar 2010.