3B2-08

# ユーザの視線に気づく会話エージェント

## ―アテンションの知覚と制御を利用した会話の円滑化―

## Perceiving and Controlling User's Attention by Conversational Agents

中野 有紀子[*1] Yukiko I. Nakano　岡 兼司[*2] Kenji Oka　佐藤 洋一[*2] Yoichi Sato　西田 豊明[*3] Toyoaki Nishida

[*1] 科学技術振興機構社会技術研究システム RISTEX-JST　[*2] 東京大学生産技術研究所 Institute of Industrial Science, The University of Tokyo

[*3] 京都大学大学院情報学研究科 Graduate School of Informatics, Kyoto University

In face-to-face communication, conversational participants not only display conversational nonverbal signals, but also perceive such signals from the conversational partner. With the goal of improving communicative capability of conversational agents, this paper focuses on attentional behaviors in conversation, and proposes conversational agents that can recognize user's attentional behaviors as well as generate such behaviors to display to the user. First, we review previous studies on attentional behaviors in face-to-face interaction. Based on the discussion, we build an immersive conversational environment IPOC, where a conversational agent embodied in a story-based communication environment. As a component of IPOC, we also implement a conversation management mechanism that maintains user's attentional information in the conversation state model, and exploits the information to determine the next agent's behaviors.

## 1. はじめに

人とコンピュータとのより自然なインタラクションを実現する新しいヒューマンインタフェースとして，会話エージェントが注目されている．会話エージェントとは，人間と同様の身体表現を持ち，音声言語とジェスチャーや表情等の非言語情報を用いて，人とコミュニケーションができるアニメーションキャラクターである．

このような会話エージェントについての従来研究を概観してみると，エージェントによる非言語情報の表出に取り組んだものが多い．つまり，表情やジェスチャーを音声言語にあわせて適切に自動生成することが中心課題となっていた [Cassell, Bickmore et al. 2001]．しかし，インタラクションは双方向的なものであるから，非言語シグナルを表出するのみではなく，相手からの非言語情報を知覚，解釈することもコミュニケーションを遂行する上で不可欠な能力である．すなわち，ユーザの非言語行動を認識し，それをインタラクションに利用する機構がなければ，インタラクティブなシステムとして非常に重要な部分を欠くことになる．

本稿では，特に会話における注視行動に着目し，会話の状況に応じて，エージェントの注視行動を適切に生成すると同時に，ユーザの視線を認識し，それを会話の制御に利用する機構を提案する．さらに，これを没入型会話環境上の会話エージェントに実装する．これにより，ユーザーエージェント間のインタラクションにおいて，非言語情報の双方向的な流れを可能にする仮想会話環境が実現する．

連絡先：中野有紀子，科学技術振興機構　社会技術研究システム，〒105-6218 東京都港区愛宕 2-5-1 愛宕グリーンヒルズ MORI タワー 18F，Tel: 03-5404-2871，Fax: 03-3432-8851，e-mail: nakano@kc.t.u-tokyo.ac.jp

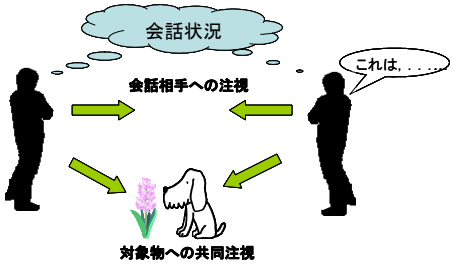


図 1：会話における注視行動

## 2. 会話における注視行動の機能

ここでは，人間同士の会話において視線がどのような役割を担っているのかをまとめ，会話エージェントのデザインの指針を得る．

[Kendon 1967] による挨拶行動の分析では，お互いが相手を認識し，会話を開始するまでの様々な非言語行動が報告されている．会話を始める前には，まず一定以上の距離から相手に視線を向け，小さくうなづいたり手を振ったりしてお互いを認識しあった後に，相手に近づき，距離が十分縮まったら，再度視線を合わせ，会話を開始するといった非言語的なステップが存在する．

一度会話が開始されると，それを維持する過程において，やはり注視行動は重要な役割を担う．図1に示すように，その 1 つは相手に注意を向けることである．一般的には，聞き手から話し手への注視行動は，話し手から聞き手への注視行動よりも多いといわれている [Argyle and Cook 1976]．聞き手が話し手に注意を向けることは，聞き手による会話への参加意思を示し，これ

が話し手へのフィードバックとなる．一方，話し手は，聞き手にたびたび視線を向けることにより，聞き手の注意状況を確認する．

一方，会話中に言及される対象物を共有しながら会話をする場合には，会話相手に対してのみならず，共有された対象物に注意を向けることは(図1)，会話遂行の目的となるタスクへの取り組みを示し [Whittaker 2003]，さらには発話の基盤化(grounding)における理解の証拠として機能する [Nakano, Reinstein et al. 2003]．例えば，地図を見ながら道順を聞く場合には，聞き手が地図に視線を向け，地図の情報を共有していることを示すことは，理解の証拠となる有効な非言語的フィードバックである．

また，ターン交代時には視線の微細なやり取りが重要な役割を担っている．従来研究では，会話が円滑に進んでいない場合には，ターン交代時のアイコンタクトの時間が長くなる [Novick, Hansen et al. 1996]，また，ターンをとる際，話し手は発話の開始時に聞き手から視線をそらす [Duncan 1974] 等の観察結果が得られている．

以上のように，視線は話し手・聞き手両者にとって，会話を円滑に進めるための重要な非言語情報であり，会話参加者は常に相手の視線の動きを観察しながら会話を進めていることが明らかになっている．

## 3. 没入型会話環境 IPOC

前節でまとめた注視行動による会話制御機能を会話エージェントに実装するために，没入型会話環境 IPOC を構築した．図 2 に IPOC とユーザとのインタラクションの様子を示す．本システムでは，パノラマ写真を背景とした会話環境上に会話エージェントが存在し，背景にある建物や対象物について短いストーリーを語ることにより説明する．例えば，日本の古い町並みが背景となる場合には，エージェントはユーザに背景にある建物やそれに関連する歴史的な事柄について会話的に説明を行う．つまり，ユーザは会話エージェントとの没入的なインタラクションを通して，背景世界についての知識を得ることができる．

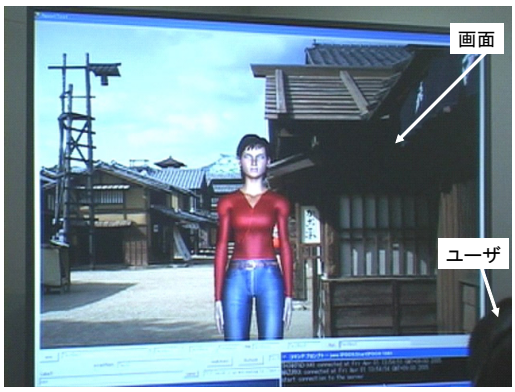


図 2:IPOC とユーザとのインタラクション

### 3.1 システムの概要

図 3 に IPOC のインタラクション制御機構の構成を示す．本機構への入力，つまり入力理解部への入力は，音声認識[Julius]によるユーザの言語的行動と，頭部姿勢推定システムを利用して推定されたユーザの視線方向である．会話制御部では，これらの情報にもとづき，会話の状態を更新し，更新された会話状態に基づき，エージェントによる次の行動を決定する．エージェントの言語的行動には，エージェント動作決定部により，その内容に応じて表情やジェスチャーが自動的に付与される．その結果，非言語情報の注釈が付与された XML 形式のデータが動作スケジューリング部に送られる．動作スケジューリング部では，日立中央研究所で開発された高品質音声合成装置 HitVoice により，合成音声が生成されると同時に，エージェントの動作と同期をとるために，各音素のタイミング情報が出力される．最後に，算出されたタイムスケジュールに従って，音声言語と同期したエージェントアニメーションが出力される．以下に，各構成素について説明する．

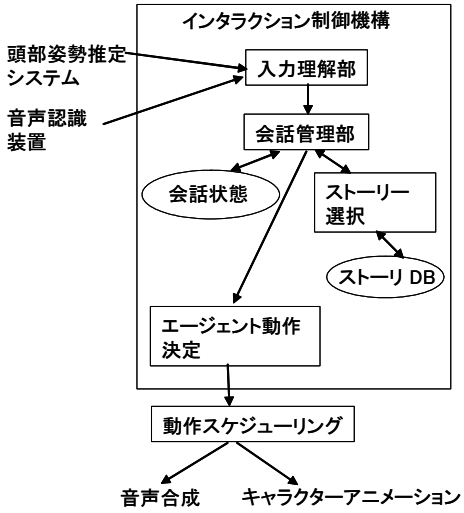


図 3:IPOC インタラクション制御機構

### 3.2 システム構成

#### (1) ユーザの視線認識機構

ユーザの視線推定には，頭部姿勢推定システム [岡 他 2005] を利用した．このシステムでは，パーティクルフィルタにおける仮説の拡散を適応的に制御することにより，ユーザが空間中のある点を注視している場合の推定精度を高く維持すると同時に，ユーザが突発的に動作する場合にも追随性を保つことを可能にしている．このような方式により，本システムは，視線の移動に伴う頭部の微細な動きを精度よく認識できるという特長を持ち，ユーザの視線推定に有用であると判断し，採用した．また，本方式ではユーザは特別な装置を身につける必要はなく，エージェントとの自然なインタラクションを損なわないという利点もある．頭部姿勢推定システムの動作の様子を図 4 に示す．

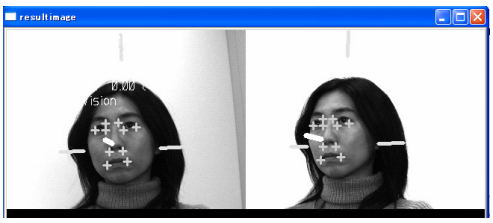


図 4:頭部姿勢推定システム動作例

このシステムを利用して，大まかなユーザの視線推定を行った．IPOC では，ユーザは 70 インチの大画面から約 1.5 メートルの離れた位置でエージェントと会話をする．そこで，大画面を6つの領域に分割し，ユーザがどの領域を注視してるのかを頭部姿勢推定システムから出力される頭部の位置と回転角度から推定した．各観測値には揺らぎがあるので，10 データポイントごとの平均値を毎秒 30 回算出し，その値を視線推定に用いた．また，全ての計算結果を出力するのではなく，ユーザの注視位置が変化した場合のみ，そのイベントが会話管理部に通知される．

### (2) 会話管理部

会話管理部では，会話状態の更新とエージェントの行動決定を行う．ユーザの視線情報を取り入れるために，従来の言語的インタラクションのみを扱う会話管理機構を拡張した点について，以下に述べる．

**(a) 視線情報の管理**：会話状態は Information State [Matheson, Poesio et al. 2000] による会話のモデル化に準じた形式で表現されており，言語的な情報として，現在の話題 ID，発話 ID，発話の開始・終了時間，発話内容，話題の焦点となる対象等が保持されている．さらに，視線情報として，現発話中のユーザの注視行動ログと，発話中のユーザの主要な注視領域（発話区間中に最も長い時間注視していた領域）が計算され，保持されている．このように，ユーザの視線情報を発話単位で管理することにより，言語的なインタラクションに伴う視線情報を会話状態の一部として扱うことができる．

ただし，現在実装している視線認識機構では，10 データポイント（0.3 秒間のデータ）の平均値をとり，さらに発話中の主要な注視領域を算出することによって，比較的安定した注視行動を観測することに最適化したアルゴリズムを採用している．従って，非常に短時間の視点移動は観測できないという欠点も有する．例えば，ターン交代時に話し手が瞬間的に聞き手から視線をそらすといった行動は検知できない．

**(b) ユーザ視線情報を利用した会話状態の更新**：視線情報を会話の状態として管理することにより，会話管理部では，これを会話状態の判断に利用している．例えば，ある発話中にユーザが注視していた対象物が，その発話の焦点となる対象と合致していれば，会話管理部は，ユーザの注視行動が適切であった（話題となっている対象物に対して注意を向けていた）と判断し，その発話が正しく理解されたとみなす（少なくとも，その仮定を支持する証拠が得られたとみなすことができる）．このように，言語的な会話状態に照らして視線情報を解釈することにより，言語情報のみの会話の状態管理よりもロバストな機構が実現する．

**(c) 視線情報を考慮した会話制御**：会話管理部は更新した会話状態に基づき，次のエージェントの行動を決定する．ここでは，視線情報を利用して次の行動を決定するとともに，視線情報の認識誤りによる会話の失敗を防ぐための会話制御を行う．例えば，ストーリーの途中にユーザの視線が他の対象物に移行してしまった場合には，システムは，これをユーザによる非言語的な割り込みとみなし，その対象物についての話題に転換する．この時，「説明の途中ですが，こちらの建物の説明に移りますか？」といった発話を生成することにより，話題を変えることをユーザに確認し，視線の認識誤りによる誤った会話進行を避ける．

### (3) ストーリー選択

会話管理部は，ユーザの発話や注視対象物に関連したストーリーをエージェントに語らせるために，発話内のキーワードや対象物の名称などをクエリとしてストーリー選択部に送る．ストーリー選択部では，クエリに応じたストーリーをストーリーDB から検索し，会話管理部に返す．

### (4) エージェントアクション生成

エージェントによる表情，ジェスチャー，およびリップシンクはエージェント動作決定部で決定されるが，これは，CAST [Nakano, Okamoto et al. 2004] システムを IPOC に組み込むことにより実現されている．CAST は，日本語文を入力すると，語彙・統語情報にもとづき，エージェントの非言語行動を自動的に決定し，さらに音声合成器からタイミング情報を取得して，アニメーションのタイムスケジュールを算出するツールである．また，同時に合成音声ファイルも作成される．CAST を利用することにより，IPOC では，エージェントの発話（ストーリーを構成する各文）に応じた非言語行動を自動的に決定，出力している．

## 4. エージェントとの会話例

以上の機能を実装した会話エージェントとの会話例を図 5 に示す．

ユーザが画面上のエージェントを注視し，ユーザからの視線が，はじめてエージェントに向けられたことをシステムが検知すると，エージェントがユーザに挨拶をし，会話が開始される（[1:S]）．最初の話題が終了した後，[11:S]では，エージェントはユーザに次に何について知りたいかを尋ねると同時に，ユーザの注視点を観測している．その結果，ユーザの視線が画面下方の家屋に向けられていることを認識し，[12:S]において，それに視線を向けながら指差すことにより，ユーザとの間に共同注視を確立しつつ，話題の転換を打診している．さらに，[21:S]では，ストーリーの途中にユーザの視線が画面上方の火の見やぐらに向けられていることを感知し，これをユーザによる非言語的な割り込みとみなすことにより，[22:S]で，話題を途中で変えてもよいか訪ねている．このときにもユーザの注視対象物にエージェントが視線を向けることにより，ユーザ視線へのシステムによるアウェアネスを表現している．さらに，視線認識誤りの可能性を考慮し，[23:H]でユーザの言語的承諾を得た後，火の見やぐらについての話を始めている．

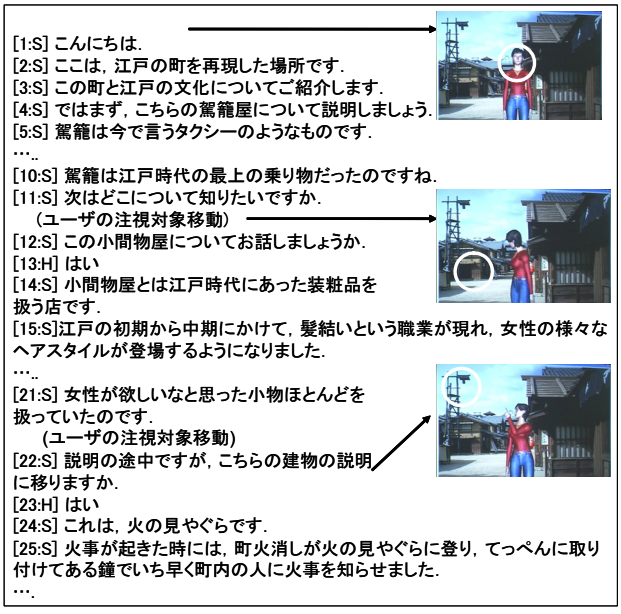


図 5:エージェントとの会話例

## 5. 関連研究

従来のシステムでは，会話環境となる背景に対して会話エージェントが注視行動を行う（生成する）ことは可能であった[Andre, Rist et al. 1999; Traum and Rickel 2002]. 典型的な例は，指差し動作を伴いながら，ある対象物に対してエージェントが視線を向けるといった動作である. しかし，会話中のユーザによる環境（対象物）への注視行動を認識し，それを会話の円滑化に利用する方法は未だほとんど研究されていない.

一方，コミュニケーションロボットの研究では，ユーザとアイコンタクトをとるなど，ユーザ視線に応じて行動を変化させるロボットはいくつか提案されているが，ロボットとユーザとの会話の中で，会話を円滑化させるための機能として視線の情報を利用しようとする取り組みは数少ない [Imai, Ono et al. 2001; Sidner, Lee et al. 2003]. 言語的に意味のあるコミュニケーションを遂行する為に視線の情報を有効に利用する手法や機構の提案は，今後，仮想世界の会話エージェントと実世界のコミュニケーションロボットとの共通の重要な課題になると考えられる.

## 6. まとめ

本研究では，ユーザの視線を認識し，それを会話の制御に利用することにより，ユーザの視線に気づく会話エージェントを実現した. これにより，ユーザがエージェントとの会話に集中し，引き込まれているのか否かを判断したり，ユーザの興味対象を検出することが可能になった. また，視線という非言語的な情報と言語的な情報とを会話管理機構において融合させることにより，よりロバストな会話インタフェースが実現することを示した. 提案手法により，ユーザにとって負荷の少ないインタラクションが実現し，会話エージェントがより有効な情報提供手段になると期待できる. もちろんこれらの仮説は，今後，評価実験により実証されるべきであるが，我々の以前の研究において，視線情報を認識・生成できる会話エージェントとユーザとのインタラクションが，人間同士のインタラクションに非常に類似していることを確認している [Nakano, Reinstein et al. 2003]. この結果は，本稿で提案した方式により，ユーザとエージェントとのより自然なインタラクションが実現することを示唆している.

将来課題も数多く残されている. 視線の動きには安定的な注視行動だけではなく，非常に短時間であるが，重要な非言語シグナルとなるものがある. このような視線移動がどのような対話状態において発生するのかを分析することにより，無視できる視線の動きと短時間であっても意味のある視線の動きとを区別する方法を見つける必要がある. また，本稿では，非言語情報として視線のみを扱ったが，ジェスチャーや体全体の姿勢等，会話において重要な非言語情報は視線以外にも存在する. これら複数のモダリティを統合していくことも今後の課題である. これらの課題を解決していくことにより，会話の円滑化に関して，会話エージェントのコミュニケーション能力をさらに向上させることができると考える.

謝辞
本研究で使用した頭部姿勢推定システムの一部には，オムロン株式会社の OKAO Vision 技術を利用しています.

## 参考文献

Andre, E., T. Rist, et al. (1999). "Employing AI methods to control the behavior of animated interface agents." Applied Artificial Intelligence **13**: 415-448.

Argyle, M. and M. Cook (1976). Gaze and Mutual Gaze. Cambridge, Cambridge University Press.

Cassell, J., T. Bickmore, et al. (2001). "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment." Knowledge-Based Systems **14 (2001):**: 55-64.

Duncan, S. (1974). "On the structure of speaker-auditor interaction during speaking turns." Language in Society **3**: 161-180.

Imai, M., T. Ono, et al. (2001). Physical Relation and Expression: Joint Attention for Human-Robot Interaction. 10th IEEE International Workshop on Robot and Human Communication (RO-MAN2001).

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

Kendon, A. (1967). "Some functions of gaze direction in social interaction." Acta Psychologica **26**: 1-47.

Matheson, C., M. Poesio, et al. (2000). Modelling Grounding and Discourse Obligations Using Update Rules. 1st Annual Meeting of the North American Association for Computational Linguistics (NAACL2000).

Nakano, Y. I., M. Okamoto, et al. (2004). Converting Text into Agent Animations: Assigning Gestures to Text. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Companion Volume, Boston.

Nakano, Y. I., G. Reinstein, et al. (2003). Towards a Model of Face-to-Face Grounding. the 41st Annual Meeting of the Association for Computational Linguistics (ACL03), Sapporo, Japan.

Novick, D. G., B. Hansen, et al. (1996). Coordinating turn-taking with gaze. ICSLP-96, Philadelphia, PA.

岡 兼司, 佐藤 洋一, 中西 泰人, 小池 英樹, "適応的拡散制御を伴うパーティクルフィルタを用いた頭部姿勢推定システム", 電子情報通信学会論文誌 D-II, vol.J88-D-II, no.8, August 2005.（採録決定）

Sidner, C. L., C. Lee, et al. (2003). Engagement Rules for Human-Robot Collaborative Interactions. IEEE International Conference on Systems, Man & Cybernetics (CSMC), Vol. 4.

Traum, D. and J. Rickel (2002). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. the first International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002).

Whittaker, S. (2003). Theories and Methods in Mediated Communication. The Handbook of Discourse Processes. A. Graesser, MIT Press.