社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－SLP－50 (10)
2004／2／7

擬人化音声対話エージェントにおける視線制御方法の検討

○ 中沢正幸† 西本卓也† 嵯峨山茂樹†

†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
{nakazawa, nishi, sagayama}@hil.t.u-tokyo.ac.jp

あらまし 擬人化音声対話エージェントにおいて，対話の流れに応じて適切にアイコンタクトを行うことは，対話相手である人間に自然な印象を与え，エージェントの実在感を高めるうえで重要である。我々は，(1) エージェントは相手に関する情報を得たり相手に合図を送ったりするために能動的に視線を動かしながら対話を行う，(2) エージェントの頭部及び眼球の動きは数理的な視線制御モデルに従う，という仮説に基づいて，対話の流れに応じて擬人化音声対話エージェントの視線の動きを制御する手法について検討している。本報告では，擬人化音声対話エージェントの視線運動に関する基本的な定式化を行う。さらに，マルチモーダル対話データを用いて頭部運動の分析を行い，提案モデルに関する予備的な検討を行う。

キーワード 擬人化エージェント，音声対話，マルチモーダル対話，視線制御，アイコンタクト

Gaze Modeling and Analysis for Spoken Dialog Systems with Anthropomorphic Agents

○ Masayuki Nakazawa† Takuya Nishimoto† Shigeki Sagayama†

†Department of Information Physics and Computing,
The University of Tokyo
7–3–1, Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

Abstract For spoken dialog systems with anthropomorphic agents, it is important to give natural impressions and real presence to human. For this purpose, gaze controls of the agent which are consistent with the spoken dialogs are expected to be effective. Our approach is based on the following hypotheses: (1) An agent performs the dialog concurrently with the intentional controls of the gaze to retrieve the information and to give signals. (2) The movement of the head and eyeballs is based on mathematical models. In this paper, mathematical models of gazing are investigated. We also analyzed the head movement data in RWC multi-modal dialog database.

Keywords anthropomorphic agent, spoken dialog, multi-modal dialog, gaze control, eye contact

1 はじめに

擬人化音声対話エージェントにおいて，対話の流れに応じて適切にアイコンタクトを行うことは，対話相手である人間に自然な印象を与え，エージェントの実在感を高めるうえで重要である。我々は，

1. エージェントは相手に関する情報を得たり相手に合図を送ったりするために能動的に視線を動かしながら対話を行う
2. エージェントの頭部及び眼球の動きは数理的な視線制御モデルに従う

という仮説に基づいて，対話の流れに応じて擬人化音声対話エージェントの視線の動きを制御する手法について検討している。

我々は，「擬人化音声対話エージェント基本ソフトウェア」[1] の基本通信プロトコルに適合し動作する，3D CG で構築された擬人化エージェント (図 1) を開発した。この 3D CG 擬人化エージェントは，全身を持ち，腕・手の動作，表情，唇の形状，視線・頭の動作をそれぞれ独立に制御することができる。我々は，この擬人化エージェントを用い，音声認識・画像認識等の言語によるコミュニケーション能力を加えることで，視線や表情，身振り・手振り，感情などの非言語情報を表現でき，より実在感のある人間型の音声対話環境を構築することを目指している。

この目的のために擬人化エージェントが実現すべきことは，人間と同じような動作を行い，自然な動きを実現することである。特に視線は，広い範囲で捉えると「表情」の一部であり，自然な動きの実現と密接な関係がある。

通常，話し手は最初に聞き手と視線を合わせてから，対話を開始し，話を始めたら聞き手から目をそらす。話の途中で話し手が聞き手を見ることで，聞き手が自分の話をよく聞いているかどうか，あるいは理解しているかどうかを確認できる。また，視線を相手の目に止め，相手に発話権を渡したいという合図を送ることもできる。擬人化エージェントにおいてこれらの視線制御を行うことは，人間同士の対面対話に慣れている人間にとっては機械と対話しているという違和感の解消につながるため，より話しかけやすい対話システムを実現するうえで重要である。

従来，擬人化エージェントの視線を制御する場合には，個々の挙動に対応する頭部や眼球などの座標移動コマンドをあらかじめ用意しておき，対話の流れに応じてコマンドを実行する，といった方法がとられている。

しかし，前述したような挙動を擬人化エージェントを用いて実装する場合には，視線を相手に向けるだけでなく，視線をそらしていく場合にも自然な動作を実現するために，何らかの統一的な視線運動モデルが必要になる。つまり，「相手の顔を凝視する」「視線をそらす」などの行為が，座標移動のコマンドによってもたらされるのではなく，状態の変化に伴うモデルのパラメータの更新に対応しており，実際の頭部や眼球などの座標はモデルから逐次的に計算される，と考えることが望ましい。

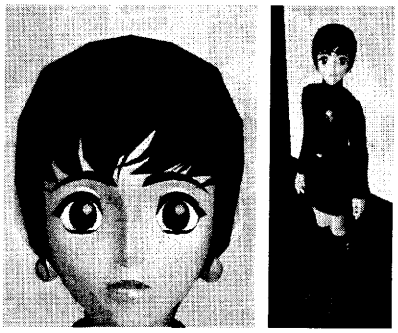

図 1: 3D CG 擬人化エージェント (顔と全身)

また，擬人化エージェントの外見や声に個性をあたえることが可能となってきているが，それに伴った動作において個性が存在しないのは，人間に不自然さを感じさせる恐れがある。我々は，視線制御モデルに，振る舞いにおける個性（個人差）をパラメータとして含めることを目指しており，これにより，外見や声と同時に挙動にも個性をあたえることを実現したいと考えている。

このようなモデルを用いることによって，擬人化エージェントにあたかも現実の人間のように振る舞わせると同時に，エージェントと対話する人間に，擬人化エージェントが意識を持ち能動的に何かに興味を持っているかのように感じさせることが，本研究の目標である。

本報告では，擬人化音声対話エージェントの自然な視線の振る舞いに関する予備的な検討について報告する。まず，視線運動に関する基本的な定式化について述べる。続いて，RWC マルチモーダル対話データベースの頭部運動データの分析結果に基づく検討について述べる。

2 従来の研究

Kendon[9, 10] は，視線活動は「相手に注意を向け，伝達する用意があることを相手に知らせる働き」があることを指摘すると共に，以下の機能があることを述べている。

- 感情表出を中心とする対人的態度の表出 (他者への関与，親和性，好意を示す動きによるもの)
- 自分の働きかけに対するフィードバックを求めることを含む情報収集
- 会話の流れを調整する機能

さらに，視線には対話開始の手がかりとしての機能だけでなく，対話中にも以下の機能があることが指摘されている。

- 発話の終了と継続を，聞き手の凝視によって確かめるモニタ機能 (monitoring)
- 聞き手の視線の動きにより会話が好ましいものかを判断し，より興味を引く話題へ調整する機能 (regulation)
- 会話が好ましいものであるか，拒絶されているかを話し手に伝える無言の表出機能 (expressive)

Argyle[9, 12] によると，視線には，相手に向けた視線の時間，相互凝視 (mutual gaze)，話をしながら相手を見る，相手の発言を聞きながら見る，2-3 秒の瞬間的な視線 (glances)，1 秒程度の相互の視線，視線のたどる軌跡，瞳孔の大きさ，目の表情，相手を見ていないときに向けている視線，瞬目の割合などがあるとし，約 1.8m の距離で，感情を強く喚起しない話題の場合，一方が相手に向ける視線は約 60%という一種の基準を示している。また，アイコンタクトを取ることにより，対話相手との心理距離を縮めることができるが，度を越すと相手を威嚇してしまい，心理的距離を引き離す効果もあることを指摘しており，適切なアイコンタクトの量が存在すると主張している。

このように視線が会話のやり取りを調整することは，Duncan & Fiske[9, 11] らによっても明らかにされ，前田ら [3] や深山ら [4] の追実験によっても確かめられている。

深山ら [4] は，眼球のみから構成された擬人化エージェントを用い，エージェントの視線を生成するための視線パラメータ (凝視量，凝視持続時間，非凝視時視線位置) の値によって定義される「視線移動モデル」を提案している。しかし，非凝視時は，射影されたスクリーン座標系の 2 次元平面上の 4 領域内のどこに属するかということを扱い，視線の運動パターンについては触れていない。

擬人化された音声対話エージェントの研究では，発話の割り込み機能，あいづち機能，視線制御機能を実現したものがある。特に，視線制御においては，Lee ら [7] は，対話における人間の眼球運動を観察し，凝視・非凝視時間を 2 次の多項式により表し，凝視位置の分布を用いてランダムに凝視動作をさせるモデルを提案している。また，Bilve ら [8] は，信念ネットワークにより，各時刻における直前の話者・非話者の凝視状態，話者・非話者の動作状態，発話・非発話状態から話者の次の凝視が発生する確率を推定している。しかしながら，これらの研究では，凝視・非凝視のタイミングと凝視量・非凝視量をモデル化することを主眼としており，その凝視・非凝視位置は，分布を用いてランダムに生成している。

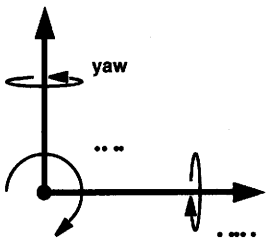

図 2: 頭部の 3 つの運動自由度 (yaw, pitch, roll)

図 3: 頭部動作を伴う視線運動の例 (上下方向，左右方向，上下左右斜め方向の動作)

3 音声対話に伴う視線制御

3.1 擬人化エージェントの視線モデル

我々は，人間型の擬人化エージェントの眼球及び，頭部 (顔方向) の運動自由度を考慮した，数理モデルによる視線運動の自動生成を目的としており，図 2 で示される頭部の運動自由度 (3: yaw, pitch, roll) 及び視線運動の自由度 (2: yaw, pitch)，瞼の自由度 (1: pitch) を動作の対象とする [2]。図 3 に擬人化エージェントの視線運動の状態例を示す。

視線運動のモデル化に際して，次のような簡略化を行った。

(1) 視線の運動パラメータ (yaw, pitch) と頭部の運動パラメータ (yaw, pitch) の比率を 1 とする (例えば正面から 10 度方向のものを見る場合，顔の角度および顔の中の視線の角度をそれぞれ 5 度とする)

(2) 瞼の運動パラメータ (pitch) と視線・頭部の運動パラメータ (pitch) の比率を 1 とする

(3) 擬人化エージェントの発話の開始時及び，終了時には凝視動作 (相手の顔を見つめる動作) を行う

上記，(1), (2) により，自由度を 2 に簡略することができ，スクリーン座標系の平面上における 2 軸 (x 軸，y 軸) のパラメータを用いて視線運動を定義

できる。また，(3) により，凝視の開始を仮定でき，凝視後の動作を考察対象とすることができる。

3.2 視線運動モデルの定式化

本報告では，スクリーン座標系の平面上の 2 軸 (x 軸，y 軸) における点として視点を定義し，視点の移動によって視線運動を表現する。視線運動のモデルとして式 (1,2) を用いる。

$$p = \frac{dx}{dt}, \qquad q = \frac{dy}{dt} \tag{1}$$

$$\frac{dp}{dt} = Bv + C, \qquad \frac{dq}{dt} = Bv + C \tag{2}$$

擬人化エージェントを用いた予備実験において v を 0〜1 の乱数値とし，試行錯誤的に得たパラメータ $B = 0.04, C = 0$ を用いたところ，比較的自然な視線運動を実現できた。しかし，相手の顔を凝視している状態および非凝視状態，自らが発話している状態および非発話状態，などのさまざまな状態を表現し，各パラメータを適切に制御するためには，実際の人間の視線データに基づいてこのモデルを精緻化していく必要がある。

4 対話データの解析

4.1 RWC 研究用マルチモーダル対話 DB

前述したモデルを踏まえて，対話中の人間の頭部運動のデータの解析を行った。

本解析で用いるマルチモーダル対話データベース [5, 6] は，対話過程に現れるマルチモーダル情報の特性を調べることを目的に作成され，光学式モーションキャプチャにより，対話時の音声，動画像及び人体各部位の 3 次元位置情報を収集したものである。

被験者の音声を捉えるマイクと，正面から顔及び上半身画像を捉える 2 台のカメラを内蔵したモニタ (29 インチ，ハーフミラー搭載) を用いて，それぞれ別々の部屋に居て着座した二人が，お互いにモニタに映し出される相手と対話する様子が 10 分間収録されている。身体に貼り付けられた 18 個のマーカー (頭部は 3 個) を追尾する赤外線カメラからの 60 フレーム／秒の 3 次元位置データが，音声波形・動画像と同時に収録されている。

綿貫ら [5, 6] は，複数人のデータから得られた頭部・胴体・手の動きの量の平均が，発話時と非発話時において異なる (有意な差がある) ということを指摘している。

本報告では，まず個人間の振る舞いの違いを考慮せずに視線運動モデルを検討することとし，特定の個人に注目して発話状態と視線運動の関係について詳細に分析する。

4.2 頭部運動からの視線方向の推定

分析対象となる対話データには，視線運動の情報は含まれていないため，次のように視線運動を仮定し，頭部の運動から視線方向を導出して視線運動の分析を行う。

- 視線運動＝頭部の動き＋眼球の動き

対話という過程に限った場合，頭部の動きは，相手を凝視するために頭を対話相手に向ける動作として捉えることができる。また，視線の運動を頭部の動きから求めることにより，不規則な眼球運動に影響されずに，大まかながら安定した分析ができると期待できる。

ここでは視線方向は顔の向きであると定義し，その導出には，頭部につけられた 3 つのマーカー (3 次元空間上の位置) の運動軌跡を用いる。3 つの基準点から座標系を定義し，頭部の 3 次元空間上の位置と向きを決定する。具体的には，一つの点を原点とし，もう一つの点の方向を x 軸，残った点を y 軸，そして，それらと直交する右ネジ方向を z 軸とする。この z 軸方向が，視線の向きとなる。実際には，3 点から決まる 2 つのベクトルの外積により直交ベクトルを求め，そのベクトルの大きさで基準化することにより，視線単位ベクトルとした。分析周期は 30 フレーム／秒とした。

4.3 視線運動分析のための仮説

自然な視線変化の動作を実現するための視線の運動モデルを 2 次式で表現するために，相手の顔の凝視時における発話と非発話の状態を分析する。本分析では特に，「顔凝視時の発話中の視線の動きと，非発話中の視線の動きは異なる」という仮説を立て，視線運動の分析を行う。

収録時の環境は，被験者間からモニターまでの距離が 1.5m である。そのため，それぞれの被験者間の距離は 3m となる。相手の顔を凝視していると考えられる角度を 5 度とすると，相手被験者の顔の中心からのずれは，約 26cm となる。本報告では，視線方向のずれが 5 度以内の場合に，相手の顔を凝視していると判断した。

4.4 視線の変動に関する検討

話者の視線角度（10 分間の時系列データ）を図 4 に示す。ここでの視線角度とは，正面（基準位置）からのずれ角度を示しており，相手の顔を凝視している場合に最小となり，顔から視線がずれるほど大きな値となる。

図 4 の中で，いくつかの箇所に大きな変動が見られるが，これは，あいづち等の頭部の大きな動作を示している。これに対して，おおよそ 10 度以内の比較的小さい変動が顔凝視時の主な視線運動である。

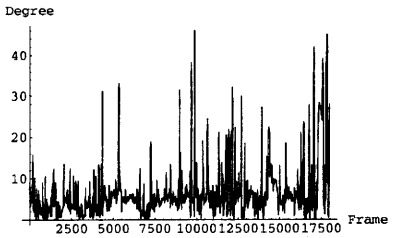

図 4: 対話中の視線角度の時間変化

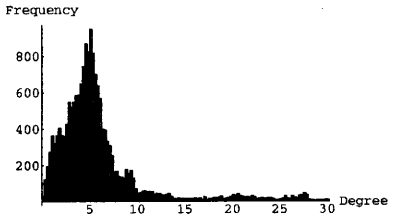

図 5: フレームごとの視線角度のヒストグラム

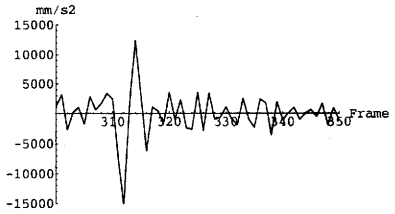

図 6: あいづち時の視線位置の加速度（上下方向）

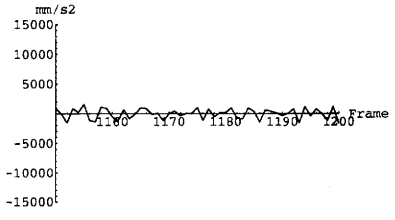

図 7: 顔凝視時の視線位置の加速度（上下方向）

各フレームにおける視線角度のヒストグラムを図 5 に示す．図 5 において比較的小さい運動は 15 度以内に分布しており，5 度付近にピークがある。これは相手の顔の中心からのずれ角度に相当するため，被験者は相手の顔の中心よりも輪郭に多くの時間注目していると考えられる。

あいづち区間における視線位置の上下方向の加速度の例を図 6 に示す（多くのあいづちに関して同様の形状が確認された）。また，顔凝視時における視線位置の上下方向の加速度の例を図 7 に示す。いずれも 50 フレーム（約 1.67 秒）の時間変化である。

4.5 顔凝視時の視線運動の検討

4.3 節での定義による顔凝視時（あいづちによる頭部の大きな運動を含まない）について，話者の各発話区間および非発話区間内における視線角度変化の平均を表したヒストグラムを図 8，図 9 に示す。各発話区間内における視線角度変化平均は，約 1.5 度以内に収まっており，非発話区間内における視線角度変化平均は，約 1 度以内に収まっている。

凝視時の話者の発話区間内，非発話区間内平均における視線の角度変化の分布の母分散が等しいと仮定できないため，Welch 近似での検定を行った。その結果，平均の差が 0 であるという仮説は棄却され（危険率 5%），平均に違いがあることが導かれた。また，視線の角度変化は発話区間において平均 0.368，非発話区間において平均 0.188 であった。

これらの結果より，相手を凝視している時において，発話時は視線の動きが活発（変化量が大きい）であり，逆に，非発話時は視線の動きがにぶい（変化量が小さい）ということができる。

4.6 考察

式 (1,2) のモデルは，視線が基準点付近で微小な動きを行い，基準点から遠ざかるにしたがって速く大きく移動する，といった振る舞いを表現しやすい。一方で，本分析から視線運動モデルの精緻化に関して得られた知見は次の通りである。

1. あいづちは事前の動作を静止させるような負の加速度，およびあいづち動作の正の加速度を示す加速度パターンで表現できる（図 6）。
2. 相手の顔凝視時にも視線には常に上下方向の加速度が加わっている（図 7）。またこのとき，相手の顔の中心よりも輪郭付近を見ている時間が多い（図 5）。
3. 相手の顔凝視時において，自らが発話している時は（あいづちのような動作を伴わない場合でも），非発話時と比較して活発に視線が動く（図 8，図 9）。

これらの知見は，提案モデルにおけるパラメータの制約条件や制御ルールとして実現していく必要がある。

5 まとめ

本報告では，擬人化音声対話エージェントが表現可能な非言語情報の一つである視線に着目し，視線制御モデルの基本的な定式化を行った。また，視

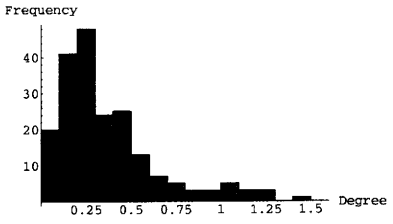

図 8: 視線角度のヒストグラム（顔凝視時・発話区間）

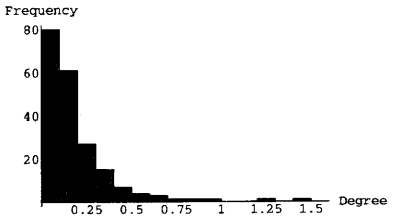

図 9: 視線角度のヒストグラム（顔凝視時・非発話区間）

線運動データの分析を行い，特に相手の顔を凝視している状態における検討を行った。

我々の提案は，擬人化エージェントにおいても人間と同様に，視覚情報が受動的に「見えているもの」ではなく，能動的に「見るもの」[13, 14] であるという前提に基づいている。擬人化エージェントを用いて，たとえ模擬的にであっても，外界の様子を探ったり，相手の表情を読みとったりしながら活動に必要な情報を取得していく様子を表現することで，より自然な人間とエージェントとの対話を実現できると考えている。また，実際に視覚情報を利用して対話を行う場合にも，振る舞いが人間らしく見えることを考慮してエージェントの見掛け上の視線を制御することが求められる。

従来研究 [7] は眼球のみの運動に，[8] は凝視の発生の有無に着目しているが，我々は頭部の動きと眼球の動きをあわせたものとして視線運動を扱っており，より自然な身体動作の実現につながると考えている。一方で，今後の課題としては，眼球に固有の運動に関する検討が必要である。特に，一つの凝視点から次の凝視点にすばやく移動するサッカード [13] については，凝視点の移動において定まった走査経路が存在することや，刺激の中で情報的価値の高い部分に凝視点が集まることなどが報告されている [15]。今後はこのような観点からの検討も行いつつ，モデルの詳細化および擬人化エージェントへの実装・評価を行っていく予定である。

謝辞

本研究の一部は東京大学 21 世紀 COE プログラム「情報科学技術戦略コア」（実世界情報システムプロジェクト）の支援を受けた。また，本研究は，RWC 研究用マルチモーダル対話データベースを使用した。RWC 研究用マルチモーダル対話データベースの利用を許可下さいましたシャープ株式会社三吉主任研究員に深く心から感謝致します。

参考文献

[1] 嵯峨山茂樹，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: "擬人化音声対話エージェントツールキット Galatea," 情処研報 2002-SLP-45-10, pp. 57–64, 2003–02.

[2] 中沢正幸，西本卓也，嵯峨山茂樹: "アイコンタクト機能を持つ擬人化音声対話エージェント," 音学講論, pp. 43–44, 2003–3.

[3] 前田真季子, 堀内靖雄, 市川熹: "自然対話におけるジェスチャーの相互的関係の分析," 情処研報, IPSJ–SIGHI–102–7, pp. 39–46, 2003.

[4] 深山篤, 大野健彦, 武川直樹, 澤木美奈子, 荻田紀博: "擬人化エージェントの印象操作のための視線制御方法," 情処論誌, Vol. 43 No. 12, pp. 3596–3606, 2002.

[5] 綿貫啓子, 関進, 三吉秀夫: "対話過程における身体の動きと音声の関係," 第 14 回人工知能学会全国大会, July, 2000.

[6] Keiko Watanuki, Susumu Seki, and Hideo Miyoshi: "Turn Taking and Multimodal Information in Two-People Dialog," *in Proc. of ICSLP*, 2000.

[7] Sooha Park Lee, Jeremy B. Badler, and Norman I. Badler: "Eyes Alive," *In Proc. of ACM SIGGRAPH*, 2002.

[8] Massimo Bilvi, and Catherine Pelachaud: "Communicative and Statistical Eye Gaze Predictions," *In Proc. of Embodied Conversational Characters as Individuals, AAMAS*, 2003.

[9] 大坊郁夫: しぐさのコミュニケーション -人は親しみをどう伝えあうか-, サイエンス社, 1998.

[10] Kendon, A.: "Some Functions of Gaze-direction in Social Interaction," Aca Psychologica, Vol. 26, pp. 22–63, 1967.

[11] Duncan, S.D., Jr., and Fiske, D.W.: "Face-to-face interaction: Research, Methods, and Theory," Hillsdale, N.J: Lawrence Erlbaum, 1977.

[12] Argyle, M., and Dean, J.: "Eye contact, distance and affiliation," Sociometry, 28, pp. 289–304, 1965.

[13] James J. Gibson: The Ecological Approach to Visual Perception, 1979. (古崎敬, 古崎愛子, 辻敬一郎, 村瀬旻 共訳: 生態学的視覚論 -人の知覚世界を探る-, サイエンス社, 1985.)

[14] 池田光男: 眼はなにを見ているか -視覚系の情報処理-, 平凡社, 1988.

[15] 行場次朗, 向後千春, 高橋信子, 横澤一彦, 斉藤勇: 認知心理学重要研究集 1 -視覚認知-, 誠信書房, 1995.