2C2-01

音声対話における擬人化エージェントの身体動作表現の利用

Nonverbal Behavior Modeling for Spoken Dialogue Systems with Anthropomorphic Agents

西本 卓也 Takuya Nishimoto　中沢 正幸 Masayuki Nakazawa　嵯峨山 茂樹 Shigeki Sagayama

東京大学大学院
The University of Tokyo

For spoken dialog systems with anthropomorphic agents, it is important to give natural impressions and real presence to human. For this purpose, nonverbal behavior such as gaze controls of the agent is expected to be effective. An agent, hypothetically, performs the dialog concurrently with the intentional controls of the gaze to retrieve the information and to give signals. In this paper, internal models of the agent, such as the attention to user and expectation of user's anxiety, are proposed. We also implemented the prototype system based on the proposed models which meets or withdraws the gaze as the response of the user utterances.

1. はじめに

我々は，GUI (Graphical User Interaface) による従来の HMI (Human Machine Interface) の限界を超える手段として，擬人化音声対話エージェントの利用を目指している．特に，我々が開発に参加し現在無償で配布を行なっている擬人化音声対話エージェントのツールキット Galatea [Galatea] [Nishimoto 04] の顔画像合成モジュール (FSM) は，手作業で組み立てられた従来の擬人化エージェントとは異なり，あらかじめ標準ワイヤーフレームモデル中の代表点と正面写真中の対応点を整合させておくことにより，実在の人間の顔写真に基づくリアルな顔画像の合成が可能である．あたかも実在の人間と対話をしているような印象を与えつつ，テキスト音声合成に伴う LipSync，怒り，喜び，悲しみ，驚き，嫌悪，恐れの表情表現，まばたき，眼球や頭部などの詳細な動きを制御できる．

このような擬人化エージェントを用いて音声対話アプリケーションを開発するにあたっては，擬人化エージェント制御の詳細と対話制御を分離して，対話タスクの記述を効率的に行えることが望ましい．そこで我々は Galatea ツールキットの一部として，対話タスクに関する制御を行う VoiceXML 処理系 (DM) の実装を行ってきた [Nishimoto 03]．

一方で，擬人化音声対話エージェントが対話の流れに応じて適切にアイコンタクトなどの身体動作を行うことは，対話相手である人間に自然な印象を与え，擬人化エージェントの実在感を高めるうえで重要である．例えば上松らは Galatea FSM を用いて，頭部，目，まばたきの挙動を手動制御しつつ，擬人化エージェントとしての自然性や音声との同期の効果について検討している [Uematsu 04]．

我々は，音声合成の分野において声帯振動機構に基づいたモデルが合理的な方法で多様な基本周波数 (F_0) パターンを説明できることに着目し，さまざまな対話現象を力学現象に等価変換することで擬人化エージェントの心的状態と挙動を結びつけるモデルの構築を目指している．具体的には，

1. 擬人化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身体や視線を動かしながら対話を行う．

連絡先: 西本卓也，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，電話/Fax:03-5841-6902，nishi@hil.t.u-tokyo.ac.jp

図 1: 擬人化エージェントとユーザの対話の様子

2. 擬人化エージェントの頭部及び眼球などの身体運動は数理的な制御モデルに従う．

という仮説に基づいて，擬人化音声対話エージェントを制御する手法について検討している [Nakazawa 04]．

一般に VoiceXML による対話記述では，ユーザとシステムがどのようなやりとりを行うかを，発話を単位として定義する．しかし，例えば，ユーザが音声入力を完了し，認識結果が得られるまでに，擬人化エージェントがどのような視覚的な振舞いを行うべきか，といった即時的なインタラクションの制御は VoiceXML では困難である．

我々は Galatea において，ユーザとシステムの発話タイミングなどを監視し，まばたき，頭部運動，視線運動などを自律的に行う新たな機能モジュールの実現を目指している．このモジュールは対話タスクの制御を行う VoiceXML 処理系と協調して動作する．音声対話アプリケーションの開発者は，対話タスクの記述のみを行うことによって，ユーザ発話に対する即時的なアイコンタクトなどを実現できる．本報告では，このようなモジュールを実現するために我々が検討している制御モデルとその実装について述べる．なお，本研究の成果の一部は，音声対話技術コンソーシアム [ISTC] での公開を予定している．

2. 対話システムにおける応答タイミング

本研究で用いる音声対話システムは公開中の Galatea for Linux v3.0 をベースに改良中のものであり，音声合成 (SSM) には GalateaTalk を，音声認識 (SRM) には Julian 3.4.1 を

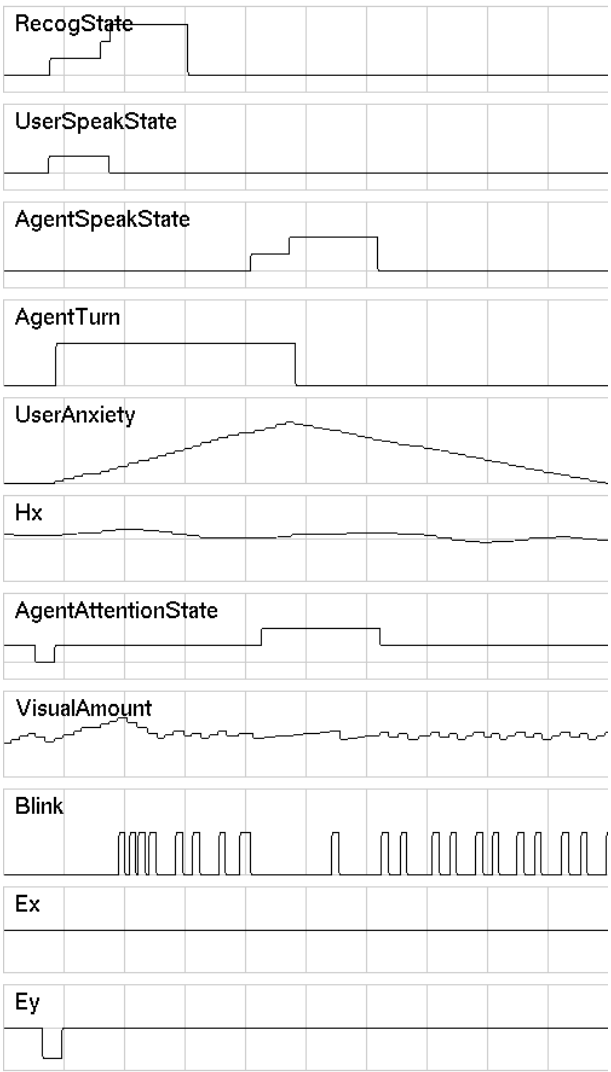

図 2: 音声対話システムにおけるタイミング（例 1：ユーザが発話し，システムが応答する）．横軸は 1 目盛が約 1 秒．RecogState, UserSpeakState, AgentSpeakState, AgentTurn, AgentAttationState については状態遷移を数値の変化で示している．詳細は本文．

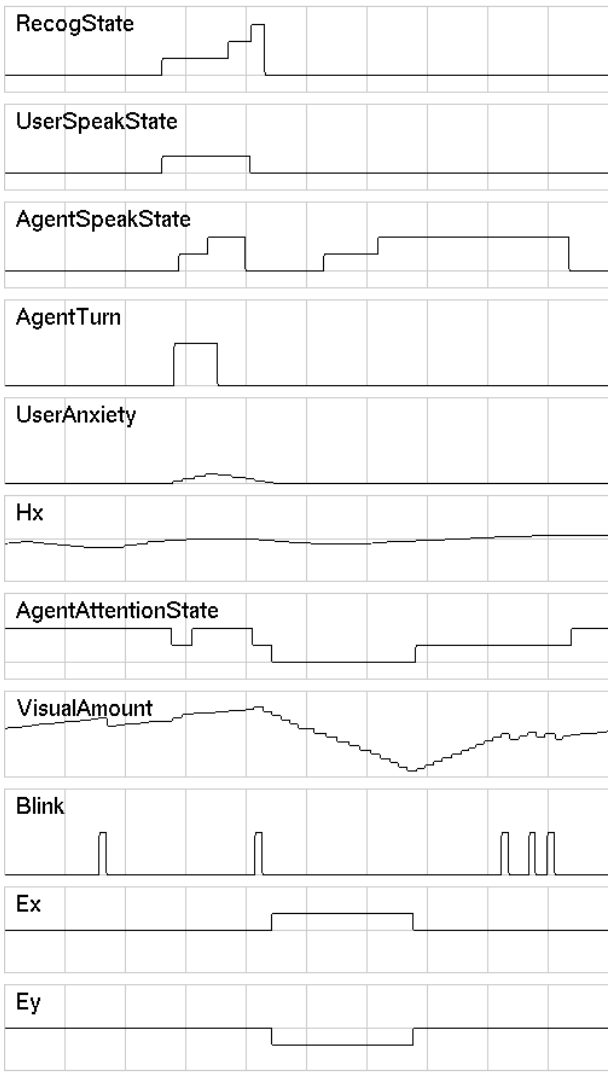

図 3: 音声対話システムにおけるタイミング（例 2：ユーザが発話し，システムが即時に相槌を返した後，さらに応答する）

使用している．図 1 に擬人化エージェントとユーザの対話の様子を示す．

音声対話システムのユーザは，自分の発話に対してシステムからの応答を得るまでに時間がかかり，しかも発話が受理されているかどうかを知ることができないと，システムに対して不安を感じることが多い．このような状況について詳細に検討するために，我々は Galatea における種々のイベントや状態の時間変化を視覚化するツール (GEV) を実装した．

図 2 は，ある対話パターンにおけるユーザの音声入力と認識結果の出力，それに対するシステムの応答発話のタイミングを示している．ユーザ発話イベント (UserSpeakState) において，発話開始から終了 (音声切り出し開始から終了) までは約 1 秒である．また，音声認識イベント (RecogState) においても，発話終了とほぼ同時に第 1 パスおよび第 2 パスの認識結果が得られている（ただし次の音声入力を受理できる状態に戻るまでに 1 秒以上を要している）．

しかし，システム発話イベント (AgentSpeakState) による と，ユーザが発話を開始してからシステム発話コマンドが発行されるまでに約 2.4 秒を要しており，これが対話管理部での処理時間であることがわかる．さらに AgentSpeakState の変化からは，発話コマンドを受理してから実際の発話が行われるまでの発話準備に約 0.6 秒を要していることもわかる．つまりこの例では，ユーザは発話を終えてから約 3 秒間（ユーザの発話開始から約 4 秒間），システムが喋り始めるまで待っている．このような応答の遅延が，音声対話システムのわかりにくさや使いにくさ，あるいはユーザの不安の一因であると考えられる．

システム応答に要する時間を短縮するためには，対話制御や音声合成などの処理を高速にすることも重要であるが，ネットワークを介した情報検索など，応答の遅延が避けられない場合もある．そこで例えば，音声認識や対話処理の結果を待たずに擬人化エージェントが「はい」などの相槌発話を行ったり，うなづきなどの視覚的な表現を行うことが効果的であると考えられる．図 3 は，ユーザの発話開始イベントに対して，システムが内容を理解する前に「はい」という発話を行い，その後，音声認識と対話処理の結果に基づいてユーザ発話への応答を行っている例である．しかし，システムが常にこのような相槌

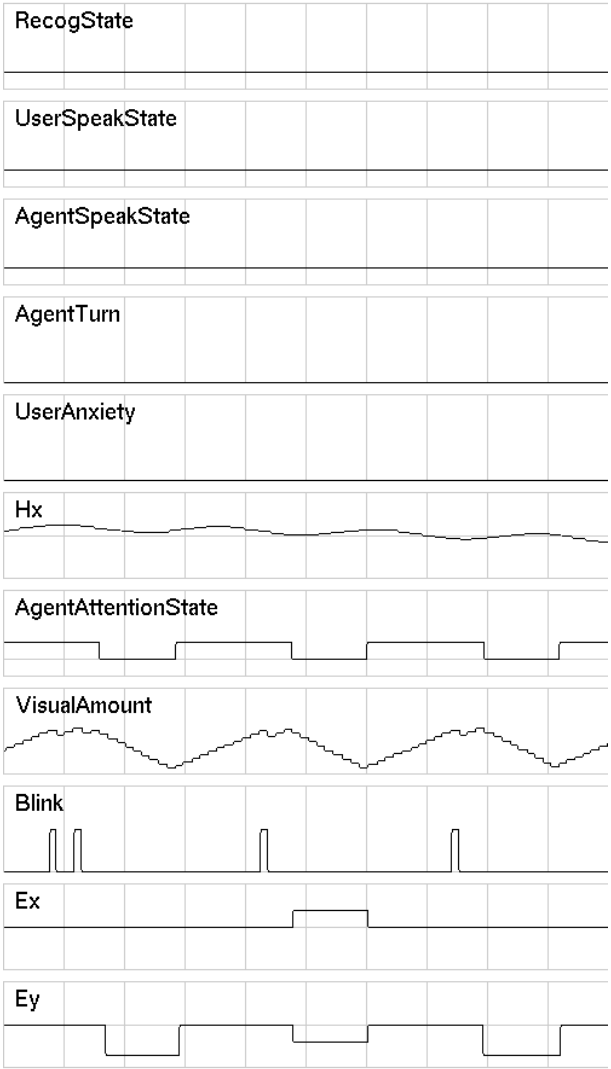

図 4: 音声対話システムにおけるタイミング（例 3：非発話時に，システムが意識非集中と意識集中の状態間を遷移する）

発話を行うことは，スピーカーの音声をマイクロフォンが拾ってしまうような実環境では音声認識性能の低下につながる恐れがあり，またユーザが煩わしさや話しにくさを感じることも予想される．

3. 擬人化エージェントの動作制御モデル

3.1 視線活動による情報提示

我々は，視覚的な情報提示手段として擬人化エージェントの顔画像合成機能のみを用いて，直感的で自然な HMI を実現することを目指している．

対人的コミュニケーションの研究において，視線活動は，相手に注意を向け，何かを伝達する用意があることを相手に知らせる働きを持っているとされる [Kendon 67] [対人行動学研究会 86]．この他にも視線活動にはさまざまな役割や性質が指摘されているが，本研究ではまず，擬人化エージェントにおいて視線活動の表現を効果的に使うことで，ユーザからの働きかけの効果をフィードバックとしてユーザに提示することを目指す．

音声による相槌を使用せず，画像情報のみでフィードバックを行なうことにより，システムからのバージインが頻繁に起こることを回避し，音声認識性能の低下やユーザの煩わしさを防げると考えられる．

3.2 モデルに基づく動作制御

画像入力の手段を持たない擬人化エージェントにおいては，視線活動は仮想的で疑似的なものとなる．そのような仮想的な視線活動は，動作パターンの作り込みや乱数によっても制御できると考えられる．しかし，擬人化エージェントがさまざまな視聴覚センサや身体動作の手段を備えた「ソフトウェア・ロボット」として拡張されていったときには，入出力の複雑な相互作用を統一的に扱う必要がある．また，擬人化エージェントの視線の動きに個性を持たせたり，ユーザの好みなどに適応するインタラクションを実現することも必要である．

そこで，本研究では，いくつかのパラメータの時間変化によって，凝視に関する状態などを制御する．具体的には，システムの内部状態がいくつかの値を保持し，それぞれの値の変化速度や上限値，下限値などのパラメータを規定し，それらの値の制御の結果として視線移動などの身体運動が起こる，というモデルを想定する．

3.3 ユーザの不安を予測するモデル

視線活動による「相手に注意を向けていることの提示」は，「なぜ返事がないのだろう？」という相手の不安を解消するために行なわれると考える．

そこで，下記の仮説に基づいて，システムがユーザの不安を予測し，その予測値を視線活動の制御に用いる．

1. システムが発話権を得ているにもかかわらず，システムの発話が始まらない場合は，ユーザの不安が増大する．
2. ユーザが不安を感じていると予測される場合には，ユーザを凝視するような視線活動を行なう．

人間に例えるならば，何かを言うべきであるが，適切な言葉が浮かばないような状態では，擬人化エージェントはユーザを凝視する，といった振舞いを想定している．

発話権の交替は音声の韻律や発話内容に基づいて適切に判断することが望ましいが，ここでは，発話終了の検出には音声区間の切り出し処理に伴う遅延が大きい，システムとユーザが任意のタイミングでバージイン（割り込み発話）を行なえる，などを想定した簡易モデルとして，ユーザが発話を開始するとシステムに発話権が移り，システムが発話を開始するとユーザに発話権が移る，と定義する．

図 2 および図 3 の例では，システムに発話権がある状態を AgentTurn の値で示している．また，システムに発話権がある状態ではユーザ不安予測値 (UserAnxiety) を増加させ，システムに発話権がない状態では減少させている．図 2 の例ではユーザ不安予測値が高くなる時間が長く続く．これに対して図 3 の例では，ユーザ発話の開始時に相槌発話が行われたため，ユーザ不安予測値の高い状態は短時間で終了している．

3.4 擬人化エージェントの凝視タイミング制御

システムが視線の凝視を用いてユーザに情報を提示するためには，凝視していない状態を適切に生じさせる必要がある．非凝視状態がどのように起きているのかを説明するモデルとして，次のような仮説を考える．

1. 擬人化エージェントは意識の集中に関して，(0) 意識非集中状態，(1) 意識集中状態，(2) 発話中状態，の 3 状態を遷移する（状態番号は AgentAttentionState の値）．

2. 擬人化エージェントは仮想的な視覚記憶バッファ（短期記憶領域）を持つ．
3. 視線の凝視を行なうことで視覚記憶バッファの内容量が増える．発話中状態では内容量の増え方が遅くなる．
4. 視線の非凝視，およびまばたきを行なうことで，視覚記憶バッファ内の内容量が減る（忘却により視覚記憶を失う）．
5. 視覚記憶バッファの容量には上限があり，内容量が上限が越えた場合には，まばたきを行なうか，意識集中状態から意識非集中状態への遷移を行なう．
6. 意識非集中状態では常に視線の非凝視を行なう．
7. 視覚記憶バッファの内容量が一定値を下回ると，意識非集中状態から意識集中状態への遷移を行なう（欠乏した視覚記憶を補おうとする）．
8. 擬人化エージェントの発話中と，ユーザの不安が高いと予測される場合には，視線の非凝視は行なわない（まばたきのみを行なう）．

図 4 はシステムもユーザも発話を行なっていない状態での状態変化の例を示している．図 2, 図 3, 図 4 を比較すると，視覚的短期記憶バッファの内容量 (VisualAmount) が一定範囲の値になるようにシステムの意識集中状態 (AgentAttentionState) が変化し，その結果としてまばたき (Blink) および眼球の注視方向の X 軸および Y 軸の値 (Ex, Ey) の値が制御されている．

3.5 擬人化エージェントの頭部運動制御

対人的コミュニケーションにおける視線活動は，眼球運動と頭部全体の運動の合成によって表出されると考えられる．しかし，本研究では，擬人化エージェントは一定のパターン（正弦波の加算）で周期的に頭部を上下左右に動かすものとし，特に上下運動の振幅を大きめに与えた．見かけの自然性を高めるための工夫として，擬人化エージェントの発話中は頭部運動を遅くする．図 2, 図 3, 図 4 においては頭部の上下角度は Hx で示している．

4. Galatea における実装

前述した身体動作の制御を Galatea for Linux において実装した．過去の報告 [Nishimoto 03] におけるモジュール構成からの変更点を表 1 に示す．各モジュールは C, C++, Java, Perl, Ruby などの言語で実装されている．下線は新たに追加したモジュールである．

10 種類程度の発話を認識し決められた応答を行なう対話パターンを VoiceXML によって記述した．この対話において，特にユーザが発話を開始した直後にシステムが非凝視状態から凝視状態に移行し，システムが応答発話を行なうまでの間，アイコンタクトを続けることができた．その他の音声入出力などの挙動は VoiceXML に基づいて実現された．

5. まとめ

擬人化音声対話エージェント Galatea において，ユーザとシステムの発話タイミングなどを監視し，まばたき，頭部運動，視線運動などを自律的に行う新たな機能モジュールの実現を目指して，ユーザの不安を予測するモデルや擬人化エージェントの凝視制御モデルおよび頭部運動制御モデルなどの利用に関する検討を行った．

今後はパラメータや制御モデルの精緻化を行ないつつ，頭部や視線の運動に関して二次遅れ系などの数理的モデルを導入することでさらに自然な身体動作表現を目指す [Nakazawa 04]．さらに，個性の実現や相手話者の振舞いへの適応などを実現しつつ，主観評価などを通じて提案手法の有効性を検討する必要がある．

表 1: Galatea のモジュール構成

モジュール名	機能
AM (AM-DCL)	各モジュールの制御
AM-MCL *	出力同期およびマクロ処理
DM-MCL *	DM 関連のイベント処理
SR-MCL *	SRM 関連のイベント処理
FS-MCL *	視線や頭部などの制御
DM	VoiceXML 処理系
FSM	顔画像合成
SSM	テキスト音声合成 (GalateaTalk)
SRM	音声認識 (Julian)
SIM	音声認識結果の後処理（意味解釈）
GUI	ユーザ向け画面
MON	対話監視者向け画面
SND	音声ファイル出力
PAR	並列出力制御
DIM	身体動作制御
GEV	イベント表示

*はブロードキャスト指定．下線は新たに追加したモジュール．

謝辞

音声対話技術コンソーシアム (ISTC) の実行委員および会員の皆様，Galatea Toolkit の利用者の皆様，助言や要望などをお寄せくださる皆様に感謝いたします．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．

参考文献

[Nishimoto 04] 西本 卓也, 荒木 雅弘, 伊藤 克亘, 宇津呂 武仁, 甲斐 充彦, 河口 信夫, 河原 達也, 桂田 浩一, 小林 隆夫, 嵯峨山 茂樹, 下平 博, 伝 康晴, 徳田 恵一, 中村 哲, 新田 恒雄, 坂野 秀樹, 広瀬 啓吉, 峯松 信明, 三村 正人, 森島 繁生, 山下 洋一, 山田 篤, 四倉 達夫, 李 晃伸: “Galatea: 音声対話擬人化エージェント開発キット,” インタラクション 2004 論文集, pp.27-28, Mar 2004.

[Nakazawa 04] 中沢 正幸, 西本 卓也, 嵯峨山 茂樹: “擬人化音声対話エージェントにおける視線制御モデルの提案,” 人工知能学会 SIG-SLUD-A303, pp.21-26, Mar 2004.

[Uematsu 04] 上松 恵介, 川本 真一, 中井 満, 下平 博, “擬人化音声対話エージェントにおける発話時の表情・頭部挙動の効果,” 日本音響学会講演論文集, 1-8-23, pp.51-52, Mar 2004.

[Nishimoto 03] 西本 卓也, 嵯峨山 茂樹: “擬人化エージェント Galatea のための VoiceXML 処理系,” 第 17 回人工知能学会全国大会, 2C2-04, Jun 2003.

[Kendon 67] Kendon, A.: “Some Functions of Gaze-direction in Social Interaction,” Aca Psychologica, Vol. 26, pp. 22–63, 1967.

[対人行動学研究会 86] 対人行動学研究会 (編): 対人行動の心理学, 誠信書房, 1986.

[Galatea] http://hil.t.u-tokyo.ac.jp/~galatea/

[ISTC] http://www.astem.or.jp/ISTC/