解　説

WWWと連携する擬人化エージェントとのHAI

Human-Agent Interaction with a Life-like Character linked with WWW

土肥 浩
Hiroshi Dohi
東京大学大学院新領域創成科学研究科基盤情報学専攻
The Graduate School of Frontier Sciences, University of Tokyo
dohi@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/

石塚 満
Mitsuru Ishizuka
東京大学大学院情報理工学系研究科電子情報学専攻
The Graduate School of Information Science and Technology, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: life-like agent character, human-agent interaction, face-to-face style communication, multimodal interface, affective computing

1. はじめに

HAI(Human-Agent Interaction) を実現する擬人化エージェントインタフェースは，人間の日常的な Face-to-face 型 (対面型) のコミュニケーションスタイルをメタファとしたインタフェースである．どこからともなく自分に向かって誰かが話しかけてくるよりは，相手の姿が見える方が違和感がなく落ち着く．自分が話し手となる場合はなおさらである．姿の見えない相手に向かって話すのは，我々が英語で国際電話をかけるときのように不安なものである．最近では音声認識が徐々に普及してきたが，四角い機械に向かって話すというのは面白くない．機械が本当に自分の話すことを聞いているのかどうかはっきりしない．しかも，自分の話したことがきちんと伝わったかどうか分からないまま，機械が勝手に動き始めるというのはあまりよい気持ちがしない．

『どうしてインタフェースにエージェントがいなければならないのか？』

これは HAI をベースとするインタフェースについて，よく受ける質問である．確かに日常よく使うワープロなどの道具としてのツールでは必ずしも擬人化エージェントなど必要なく，かえって煩わしいということにもなってしまう．しかし他方で，

『どうしてニュース番組にキャスタがいなければならないのか？』

ユーザが求めているのはニュースという情報であり，キャスタはその情報の流れを整理調整し，ユーザが情報を入手することを助ける．ニュースキャスタがいないと絶対に困るというわけではないのに，現実には多くの人が (個人的な好き嫌いは別にして) キャスタの存在を自然に受け入れている．

B.Reeves と C.Nass は著書 “The Media Equation (1998)”[Reeves 98] の中で，我々人間はコンピュータを含む人工物のメディアに対しても (進化の結果として遺伝子に織り込まれていて) 自然に社会性を感じ，人に接するのと同様な対応を示す傾向をもつことを明らかにしている．

擬人化エージェントは，多様化，複雑化する情報空間の中で，自然のコミュニケーションに近い形態で必要な情報を入手したりタスクを実行したりするのを手助けしてくれる協力者，パートナとなる．デジタルデバイド (情報格差) の解消が大きな社会問題になっており，情報技術に取り残されがちな高齢者や障害者にも使いやすいシステムの開発が期待されている．

HAI において顔や姿をもつことの第一の効用は，ユーザの認知的負荷の軽減であり，第二の効用は音声言語を補助する表情，身振りなどの情報 (ノンバーバル情報) も伝達し，理解容易度の向上を可能にすることにある．そして第三の効用として，背景あるいは擬人化エージェントの様子により対話の状況，コンテクストを明らかにし，音声対話の話題の範囲を自然に限定して，認識・理解の性能向上を可能にすることも加わる．

以上はコンピュータから人間へのマルチモーダル情報伝達の話だが，逆方向のユーザの状況や表情の認識情報は，コンピュータがより気配り，思いやりのあるインタラクションを生成する上で重要となる (Affective Computing)．

英語で単に “Agent (software)” というと，特に姿・形を持たず，自律的に何かの機能を担うプログラムやツールを指す場合が多い．顔や姿をもつものは，“Life-like (or Believable) Agent”，“Embodied Agent”，“ECA (Embodied Conversational Agent)”，“Anthropomorphic Agent” などという．最近では “Life-like Agent” がよく使われているが，“Life” という誰でも知っているが日本語に訳しづらい言葉に，さらに “-like” がくっついて

おり，なかなか適当な日本語がない．生命的エージェントと呼ばれることもあるが，擬人化エージェントという言葉の方が広く使われている．

なお筆者の一名により記された少し前の解説 [石塚 00] は本解説を補う部分があるので，合わせて参考にしてもらえればと思う．他の最近の解説には，[Cassell 00, Johnson 00, Cassell 01a, Gratch 02] などがある．

2. HAI における要素技術とその統合

2·1 エージェントキャラクタのデザイン

HAI ではエージェントの使われ方によって，以下のような種々のキャラクタが使われている．このうち顔あるいは肩から上のものは Talking Head と呼ばれる．

- 顔あるいは肩から上 (腕なし)
  Haptek(Haptek), Marco(Memphis 大), VSA-II(東大, テクスチャマッピング), SmArt(東大)[Barakonyi 01] など
- 上半身 (腕あり)
  Steve(USC/ISI), Ashow(産総研)[Hasegawa 95], MAICO (RWC / シャープ) [三吉 01] など
- 全身 (2D, 3D)
  MS Agent(Microsoft, 2D), Jack(Pennsylvania 大), Extempo キャラクタ (Extempo), Cyberella(独 DFKI), REA-agent(MIT), TVML エージェント (NHK/日立国際電気)[TVML], MPML-VR エージェント (東大) など

これらのうち，MS Agent は顔表情などを表現することは不得手であるが，Windows2000 以降の Windows OS に搭載されており，インタフェースが整っていて最も使い易いキャラクタシステムである．(ただし独自キャラクタを外注する場合の作成費用は 200〜400 万円程度となる．) また MPEG-4 の顔モデルを用いて標準的な擬人化エージェントを作成しようとする試みも見られる [Pasquariello 01].

顔や上半身のキャラクタではその表示位置が自然と限られるが，全身像のキャラクタは項目を指し示したりするために画面内の任意の位置に移動させても不自然さはない．2 次元スペースの場合，キャラクタを画面の上下方向にもうまく移動させることが必要になるが，MS Agent ではジャンプしたり，あるいはプロペラや翼で飛んで移動するような動作で実現を図る．

VRML などの 3 次元スペースでは，そのシーンを写しているカメラの位置や向きを変えることもできる．そのとき，キャラクタの見え方だけではなく，その後ろにある背景の見え方も同時に変化する．

擬人化エージェントにリアルな顔を使うことには賛否両論があり,『リアルな顔はユーザに過度の (人間と同程度の知能の) 期待を抱かせるのでよくない』という指摘もある．これは擬人化エージェントが使われる場面にも大きく依存する問題であり，一概にどちらかが優れていると決めつけることはできない．例えばニュースキャスタが漫画的なキャラクタに替わったら，不自然かもしれない．魅力的なキャラクタをデザインするには，アートのセンスが必要であり、簡単にキャラクタのバリエーションを増やせるという点では，テクスチャマップを使った自然感の高い顔の方に利点がある．

2·2 音 声 認 識

IBM 社の音声認識ソフトウェア ViaVoice が登場して以来，Windows 上で音声認識アプリケーションプログラムを比較的簡単に開発できるようになった．現在では多くの音声認識ソフトウェアが HAI システム構築に利用できる．

音声による操作は，使い方を覚えたり練習したりする必要がなく，誰でもが手軽に使えるという点で優れている．一方で，以下のような欠点も併せ持っている．

- 十分な認識性能が得られるとは限らない．
- 使用する場所を選ぶ．
- 特徴的なランドマークがある場合を除いて，任意の位置を正確に指示することができない．

音声認識では，大別して三つの認識モードが使われている．

- Phrase モード
- Context-Free Grammar モード
- Dictation モード

Phrase モードは，発話される可能性のある単語や短い文章を認識候補として事前に登録しておく．Context-Free Grammar モードでは，候補となる文章の中に単語リストを指定したり，文章の一部分が省略されるかもしれないことを指示できる．認識率は，認識候補となるフレーズや文法によって大きく変化する．認識候補の発音や長さがお互いに似ておらず，その候補数が少ないほど，認識率は向上する．たとえ認識率が 90 %を超える音声認識ソフトウェアでも，認識候補として発音や長さが近いフレーズを多数与えれば，認識率は極端に低下してしまう．

Dictation モードは，予め認識候補となるテキストを与えることなく，発話された音声をテキスト化するものである．内部では非常に高度な音声処理を行なっている．発話された言葉の前後関係によって最適な文章を確定するため，文末まできちんと発話されない場合には，発話の続きを待ってしばらくの間，制御が戻ってこないことがある．

擬人化エージェントインタフェースに音声認識を利用する場合，Dictation モードを使うのが理想的ではあるが，認識性能等の問題から現状では Phrase モードや Context-Free Grammar モードを使うのが一般的である．音声認識を常時動かしておくのではなく，音声入力を受け付ける場面があらかじめ決められており，それぞれの場面で 3〜5 個程度の短文あるいは単語を認識するものが多い．

音声入力を常時動かしておくなら，エージェント自身が発話する音声には反応しないような工夫が必要となる．

音声認識では、誤認識が発生することは避けられない．そこで入力された音声と予め登録したフレーズや文法との類似度をスコアとして返すものもある．発音が正しいのにスコアが低い場合は，次の二つのどちらかが考えられる．

(1) 偶然に音声の取得がうまくいかなかった
(2) ユーザが予め想定していないことを喋った

通常はユーザに対して再入力を促すが，(2) の場合には何度繰り返しても認識できないので，確実にループから脱出できる別の手段を用意しておく必要がある．

逆に誤認識しているにもかかわらずスコアが非常に高い場合も，対応が非常に難しい．

2·3 音声合成とリップシンク

擬人化エージェントとの HAI では，合成音声と唇の動きを正確に同期させるリップシンクが重要である．リップシンクが外れると，キャラクタの存在が逆に煩わしくなってくる．

Windows 環境では，SAPI(Speech API) が口形状をサポートしているので，英語のリップシンクを実現するのは簡単である．ただし，SAPI4(SAPI Ver.4) と SAPI5 では口形状 (Viseme) の取り扱い方が大きく変更されているので，注意が必要である．ところが日本語の場合には，うまくいかない．これは技術的な問題というよりも，SAPI の仕様上の問題である．SAPI は日本語の口形状を定義していない．その結果，日本語の口形状を得ようとしてもランダム値，あるいは NULL 値が返される．そのため日本語システムや UNIX 環境ではリップシンクをあきらめて，発話とは無関係に音圧や乱数によって唇をランダムに動かしたりする場合が多い．

正確なリップシンクをするには，一般に音声合成側で口形状情報を生成する必要がある．そのようなフィードバックのない市販の音声合成装置を使用する場合，日本語のリップシンクは困難であると思い込んでいる人が多いが，簡単なリップシンクであれば可能である．動いているものに対して，人間の眼はそれほど敏感ではない．したがって，口の形は音声と完全に一致している必要はない．ただし，しゃべり始めとしゃべり終わりだけはきちんと合っていないと不自然さが目立ってしまう．外国映画の吹き替えでもそれほど違和感を感じないのは，喋り始めと喋り終わりをきちんと合わせ込んであるからである．

我々が開発した VSA-II[Dohi 01] では，日本語の音声合成に市販のソフトウェアを使用している．デモを見た人から『どのようにして合成音声とのリップシンクを実現しているのか?』という質問をよく受けるが，実は我々は音声合成ソフトウェアとの同期をほとんどとっていない．それにもかかわらずリップシンクしているように見えるのは，発話テキストから発話タイミングを細かく予測しているからである．

図 1 VSA-II インタフェース．大画面テレビ電話のように見える自然感の高い顔をもつ．高速モータドライブ・ディスプレイがユーザの動きを自動的に追いかけて旋回し，アイコンタクト (視線一致) する．自然感をより高めるため，背景としてユーザの向きに合わせたライブ映像とリアルタイム合成している．画面右側は擬人化エージェントと連携して動作する Internet Explorer.

VSA-II では，まず発話テキストである漢字カナ混じり文を受け取り，そこからアクセント記号付きカナ文を自動生成する．次に，カナ文から口形状列を抽出する．音素レベルでは扱っていないので，各音素長は考慮していない．実はこのとき，口形状列として単純に母音を抽出するだけではうまくいかない．破裂音などを発音する際には，途中で口を一度閉じなければならないのである．例えば「母 (ハハ)」と「ママ」は母音列としてはどちらも「あ-あ」となるが,「母」は「あ」の口の形のままで発音できるのに対し,「ママ」の場合は途中で口を一度閉じないと発音できない．さらに拗音や長音，読点や句読点のタイミングを調整する．アクセント記号付きカナ文を音声合成装置に投入してから実際に合成音声が発話されるまでの遅延時間やキャラクタ画像生成の時間を考慮しながら，キャラクタの口形状を変化させていく．音声合成側が口形状をサポートする場合，発話中の音に対して口形状を先読みできないとキャラクタ画像生成時間分だけずれることになる．発話テキストから予測する方法では，この時間調整が簡単にできる．タイミングの誤差が累積しないような実装をするのが，フィードバックのないリップシンクをきれいに見せるポイントである．

嵯峨山らは IPA のプロジェクトで独自の日本語音声部分を中心とし，顔エージェントとリップシンクする公開用擬人化エージェントツールを開発している [川本 02].

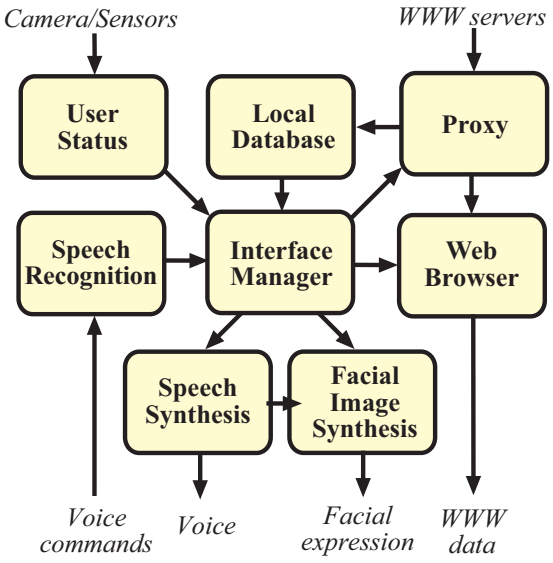


図 2 WWW と連携する擬人化エージェントインタフェースの構成例

3. WWW と連携する擬人化エージェントインタフェース

3·1 WWW との連携

現在では比較的性能のよい音声認識/合成ソフトウェアが容易に入手できるようになったため，簡単な音声対話システムを作ることはそれほど難しくはない．しかし，コンテンツが少なかったり，古い情報がいつまでたっても更新されないようなシステムでは魅力に乏しい．また出力チャンネルがテキストと音声だけでは表現力に限りがある．日常生活においても，人に何かを伝えたいときに言葉だけではうまく説明できなくて，大慌てで紙と鉛筆を探すのは誰もが経験していることであろう．

擬人化エージェントと WWW を連携させることは，それぞれの機能を相互に補完することになる．擬人化エージェント側からみれば，WWW は常に最新で膨大なデータを有する大規模マルチメディアデータベースである．ニュースや天気予報などの情報が常時更新されており，常に最新のデータにアクセスでき，またテキストだけでなく多くの画像や音声，動画などのマルチメディアデータを含んでいる．情報提示手段として，標準的 Web ブラウザを利用できることも大きな利点である．また WWW 側からみれば，擬人化エージェントはインタラクティブ・フロントエンドとして働く．

我々が開発した VSA は自然感の高い顔と簡単な音声対話インタフェースを備えた擬人化エージェントインタフェースであり，WWW との連携を実現している．VSA-I[土肥 96][土肥 99] は，複数の UNIX WS 上で走る独立した機能モジュールプロセスを共有メモリと TCP/IP で統合している．後継となる VSA-II[Dohi 01] は，Windows をプラットフォームとしている．

3·2 Web ブラウザのコントロール

我々が最初 1995 年に VSA-I と接続した Web ブラウザは，NCSA Mosaic であった．外部プロセスから Mosaic を制御するための API がまだ開発中であったため，公開されていた Mosaic のソースプログラムに直接手を加えることにより，外部プロセスからの制御を実現した．その後の Netscape では，X-window 上で XChangeProperty 関数を使用することにより，同様の制御が実現できるようになった．

現在の VSA-II では，Internet Explorer と連携している．Internet Explorer は ActiveX 化されており，Windows 上であればブラウザ制御は非常に簡単である．

3·3 音声による Web ブラウザコントロール

音声による Web ブラウザコントロールには，主に次の三種がある．

(1) 『戻る』『進む』『ホーム』のような URL を指定しない Web ブラウザの操作を音声で指示する．
(2) 音声の指示により，あらかじめ想定された Web ページにジャンプする．
(3) 音声の指示により，(音声による操作を前提としていない) 任意の Web ページから張られているリンクを辿る．

このうち (1) は，音声認識ソフトウェアと外部プロセスから制御できる Web ブラウザがあれば簡単に実現できる．(2) も同様に，簡単に実現できる．例えば『東京大学』という言葉を音声認識すれば，東京大学の Web サーバと接続するように Web ブラウザを制御すればよい．ただし，これはあらかじめ想定された世界から飛び出すことはできない．

(3) は，これまでに蓄積されてきた膨大な HTML データを利用するという観点からみると，最も望ましい形態である．HTML ファイルを解析すれば，タグ情報からアンカ文字列とそのリンク先の URL が分かる．したがってアンカ文字列が発話された時にその URL をオープンするように Web ブラウザを制御すればよいのであるが，音声による操作を前提としていない Web ページでは次の事由によりそれほど簡単ではない．

- マウスカーソルの位置に依存するものがある．
- 画像など，文字列を含まなくてもアンカにできる．
- URL など，アンカ文字列が必ずしも発話できるとは限らない．
- 同名のアンカがある．
- 膨大な数のアンカが存在する場合がある．

VSA-I[土肥 99] では新しい Web ページがオープンされる毎に，そのページ内に含まれるアンカ文字列とその URL を動的に抽出し，それに一連番号を付けてユーザに提示した．これにより，ユーザはアンカ文字列のかわりにアンカ番号を発話してもリンクを辿ることができるようになり，画像のアンカにも対応した．さらにアンカ

文字列の一部分を発話しても選択できるようにした．しかしながら最近の Web ページはフレームや Java を多用しており，音声による操作を考慮していない一般の Web ページを音声だけで操作することは，以前に較べて困難な状況となっている．

3·4 Web からの情報取得による音声応答

Web ページに情報があるからといって，いつもブラウザを操作する必要があるとは限らない．もし目の前に自分の探している情報を知っている人がいたら，その人に直接尋ねた方が簡単な場合もある．VSA では，バックエンドで Web ページから定期的にさまざまな最新データを取得してローカルデータベースを更新しておき，例えば天気予報が知りたいとき擬人化エージェントに『明日の天気は?』と尋ねると，すぐに『予報では，晴れのち曇り．降水確率は 0 %．最高気温は 25 度の見込みです....』というように答えてくれる機能を実現した．これは，天気予報などの Web ページではデータは随時更新されるものの，そのページフォーマットはほとんど変更されることがないことを利用して，音声による操作を前提にしていない Web ページからデータの切り出しを行っている．

これに対して，初めから音声対話を前提とした Web ページを作成しようとする試みもある．VoiceXML [VoiceXML] は，XML をベースとする音声対応の Web ページ記述言語であり，Web 上で対話型の音声応答アプリケーション (音声ポータル) を簡単に実現できるようにすることを目指している．さらに VoiceXML の仕様を拡張して，擬人化音声対話エージェントの制御に利用しようという研究もある [川本 02]．

4. 対話マネージメント

J.F.Allen ら [Allen 01] は，対話型インタフェースのタスクの複雑さとして，表 1 を示した．下にいく程，柔軟で自由度の高い対話が必要，あるいは実現されることになる．この性能向上の研究は自然言語対話，音声対話分野で行われているが，真に実用的なレベルとするにはまだ課題が多いのが実情である．

現在の多くの擬人化エージェントの対話マネージメントも低レベルにとどまっており，有限状態遷移による制御＋α といったレベルである．これで分岐があるマルチストーリ型対話を実現し，状態毎の音声入力認識，応答の生成を行っている．しかし，想定されるすべての状態やそこでのユーザの入力パターンをすべてあらかじめ用意するのは困難である．努力はしても想定しないユーザ入力や事態は避けられず，この対応をうまく処理する必要がある．典型的には聞き返したりメニューを示して選択させたりするなどであるが，ユーザに柔軟性，自由度の乏しさを感じさせてしまう．

表 1 対話とタスクの複雑さ ([Allen 01])

Technique Used	Task Complexity	Dialogue Phenomena Handled
Finite-state script	Least complex	User answers questions
Frame-based		User asks questions, simple clarifications by system
Sets of contexts		Shifts between predetermined topics
Plan-based models		Dynamically generated topic structures, collaborative negotiation sub dialogues
Agent-based models	Most complex	Different modalities (for example, planned world and actual world)

これはまさにコンピュータが知能をもつか否かを判定するチューリング・テストの状況に当たる．最初，1966 年に J.Weizenbaum により作成されたおしゃべり対話ソフトウェア ELIZA は，人工無能でもこのチューリング・テストに合格できることを実証するものとして生まれ，結果としてある状況では人格をも感じさせることができることを示した．最近では Chatbot と称されるこのようなおしゃべり対話ソフトウェアは，1991 年から毎年 Loebner 賞コンテスト [Loebner Prize] が行われ，技術的進歩が見られる．基本は入力テキストに含まれる文字列パターンに対して，予め用意された複数返答テキストパターンの中から一つをランダムに選び，穴埋めして返すことであるが，ユーザ入力の一部を返答テキストに組み込んだり，対話進行中に登場した複数のキーワードを組み合わせて返答の仕方を変える，重要語の認識や学習機能を組み込むなどの拡張が図られている．

このような Chatbot 技術を部分的に導入することは，擬人化エージェント対話の対応範囲，自由度の拡大に資すると考えられ，検討が行われている．このような人工無能プログラムだけでは目的をもった会話が成立しないので工夫が必要であり，例えばタスク依存の必要な部分さえ記述すればエージェントが自律性をもつような自然な対話が実現されることが期待される．

擬人化エージェントとのマルチモーダル対話では，Communicative Feedback と称される同意のうなずきや不同意の表現 (バーバル, ノンバーバルの両者を含む) も，親しみがある自然な対話にするために重要視される [Cassell 99]．

5. キャラクタ・エージェントを用いるマルチモーダルコンテンツの記述言語

単にマルチモーダルインタフェースでなく，擬人化あるいはキャラクタエージェントは新形態マルチモーダルメディア/コンテンツとしても有望視されている．USC Inst. for Creative Tech. のような米軍トレーニング用の専用的システムの開発もあるが，最近では誰でもがマルチモーダルコンテンツを記述 (オーサリング) できるツール体系の確立が重要になってきている．

コンピュータゲームに見られるように，プロのクリエータが労力をかけて制作すれば大変魅力あるマルチモーダ

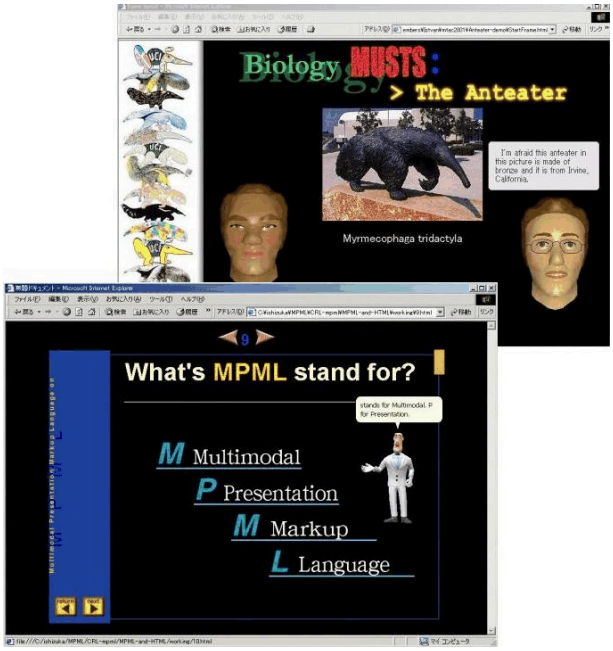


図 3 MPML によるマルチモーダルプレゼンテーションの画面例 (右上後部の画面のキャラクタエージェントは独自の SmArt Agent.)

ルコンテンツができるレベルのメディア技術はそろってきているが，これは必ずしも一般の人に使える形のものではない．HTML で誰でもが容易に Web コンテンツを記述できることによって Web コンテンツが飛躍的に増大したように，キャラクタエージェントによるマルチモーダルコンテンツも誰でもが使えるオーサリング環境が整備されれば，急速にコンテンツ制作，流通が広まるものと考えられる．Web ブラウザのコンテンツ表示機能の活用，Web 上でのコンテンツ流通を意図して，最近では XML に基づくコンテンツ記述言語が中心的に検討されており，以下のような開発例がある．

- VHML[Marriott 01]
- MPML[筒井 00, Tsutsui 00, Zong 00, Descamps 01]
- CML / AML[Arafa 02]
- APML[DeCarolis 02]
- RRL / NECA[Piwek 02]
- BEAT[Cassell 01b]

VHML (Virtual Human Markup Language) は Talking Head/Talking Human とのインタラクションを記述するためにカートン大 (豪州) で開発された言語であり，以下の六つのサブシステムで構成されている．

DMML(Dialog Manager ML)，SML(Speech ML)，FAML(Facial Animation ML)，EML(Emotion ML)，BAML(Body Animation ML)，GML(Gesture ML)．

低レベルから対話マネージメントといった高レベルの制御まで配慮されているが，そのために幾分複雑になっている．

MPML (Multimodal Presentation Markup Language) は，我々が開発を行っている XML に基づく記述言語であり，キャラクタエージェントの動作制御だけでなく，それによるプレゼンテーションの制御機能 (ページ単位のプレゼンテーション画面の提示，更には SMIL[SMIL] 準拠のメディア同期機能など) も備えている．誰もが容易に魅力あるマルチモーダルコンテンツを制作できることを目指しており，図 3 はこの MPML によるプレゼンテーション画面の例を示している．双方向音声対話も可能であるが，音声認識の性能は必ずしも実用的に十分なレベルではないことから音声入力部分の割合は数%程度であり，音声に関しては 90 数%が TTS(Text–To–Speech) による出力であるようなプレゼンテーション用途が最も実用的な適用領域となっている．

MPML では OCC モデル [Ortony 98] に基づく感情表現の記述も可能であり，エージェント用人工感情モジュールである SCREAM[Prendinger 02] と連携したエージェントの感情表現制御も可能になっている．また複数エージェントの動作制御が可能である．Flash3D のような新しい Web3D メディアとの連携も行えるようにしており，最近では携帯電話へのキャラクタを用いるコンテンツを配信，表示する MPML mobile version，3 次元 VRML 空間でのエージェントによるプレゼンテーションを記述する MPML–VR も作成されている．

XML に基づく記述言語は幾つか提案，作成されているが，標準化が図られなければ広く流通，普及することは困難となる．EU の IST(Information Science Tech.) プロジェクトではキャラクタエージェントによるマルチモーダルメディア技術に関して欧州のいくつかの大学，研究機関に資金を出しているが，この中では新技術開発・評価と並行して標準的記述言語の制定が大きな課題となっている．この記述言語の標準化に関する IST プロジェクト会合が欧州の大学，研究機関に加え，カートン大 (豪州)，東大の我々，南カリフォルニア大 (米国) のグループも参加して 2002 年 7 月にボローニャ(イタリア) でもたれ，標準化に向けて検討を始めることが合意された．記述言語の標準化は XML の形式に基づくことは合意されているが，論点を幾つか記すと以下のようになる．

- 誰のための記述言語とするか (一般の人々か，プロのコンテンツクリエータにも対応するかなど)
- 記述のレベル (低位，中位，高位レベル) と範囲
- 使用するコンポーネントモジュール (キャラクタシステム，音声認識/合成) の機能と標準インタフェース
- 感情やパーソナリティの表現 (種類と語彙) と制御法

2D，3D の差異もあることから 1 種のみの統一標準とするには困難な場合，少数複数のバージョンとなる可能性もある．

これとは別に，XML 相互運用性を目指す標準化団体 OASIS を足場にして，Virtual Human を XML で規定する標準設定に向けて HumanML[HumanML] の検討が行われている．しかし，こちらは議論が発散的で収束す

る方向にはなっていないようである．

6. 感性的エージェントへ向けて

記述言語の標準化と並んで，エージェントの生命性 (life-likeness)，信憑性 (believability) を向上させるために，感情，パーソナリティ (個性) の与え方が大きな課題になってきている．感情は短時間の心の動きであるのに対し，パーソナリティは半永久的な感じ方，行動の様式であり，この中間に中期間の心の状態を示すものとしてムード (気分) を置くこともある．

感情は OCC モデル [Ortony 98] が 22 感情を扱う最も包括的なモデルであり，簡単なものとしてよく使われるのは 6 種の基本感情 (怒り，恐れ，悲しみ，喜び，嫌悪，驚き) モデルである．パーソナリティに関しては，Big5(Extraversion, Agreeableness, Consciousness, Neuroticism, Openness) が軸としてよく参照される．

社会的存在としての性格を帯びて来つつあるエージェントは，感情やパーソナリティの表出等に加えて，社会性をもった行動をとることも重要になる．この面の研究はまだ多くないが，前節で述べた我々のグループによる人工感情モジュール SCREAM は，社会的関係と社会的距離により感情の表出を調節する社会的フィルタリング機構を備えている [Prendinger 01a, Prendinger 01b]．

感性的インタラクションに向けてユーザの感情を認識した上での対応も重要となり，このような研究分野は “Affective computing”[Picard 97] あるいは “Affective Interactions” と呼ばれている． 日本語で訳し分けるのは難しいが，“affect” は “emotion(情動)” の他に “mood (気分)” なども含んだ，より一般的な広範囲の感情を表すもの [Paiva 00] とされている．

ユーザの感情は，顔の表情 [Chandrasiri 01] や音声 (声の調子) から推定する研究があるが，最近では生体情報も利用されている．ユーザの状態を表す生体情報として，主に次のようなものが使われている．

- 皮膚抵抗 (SC, GSR: Skin Conductance または Galvanic Skin Response)
- 脈拍 (HR, BVP: Heart Rate または Blood Volume Pulse)
- 体温 (TEMP: Temperature)
- 呼吸 (RESP: Respiration)
- 脳波 (EEG: Electroencephalogram)
- 筋電図 (EMG: Electromyogram)
- 心電図 (ECG: Electrocardiogram)

このうち，皮膚抵抗，脈拍，体温の三つは比較的測定が容易であり，小型のセンサを装着するだけでよい．

皮膚抵抗は反応・緩和時間が早いことが特徴で，うそ発見器に使われているのと同じ技術である．医療機器として認定を受けた測定装置は非常に高価であるが，実験用としてはオペアンプ (あるいはトランジスタ)1 個でも実現できる．また脈拍は，赤外光を指 (あるいは耳朶) に照射し，その反射光量 (あるいは透過光量) をホトトランジスタで計測する．

生体情報は，ユーザが意図的にコントロールすることが難しい．[Wilson 00] では，被験者が意識しないようなストレスを与えた場合でも，生体情報は明らかな反応を示したことが報告されている．このことは刺激に対する正直な反応が期待できる一方，実験においても被験者を実際にある感情の状態に追い込まなければならないことを意味する．

これらの生体情報の変化と人間の感情は 1 対 1 に対応しているわけではなく，いろいろな仮説と検証が繰り返されている．次第に擬人化エージェントとのインタラクションにも採り入れられてくるものと思われる．

7. むすび

本稿では，VSA (Visual Software Agent) の開発経験を通して WWW と連携した擬人化エージェントとの HAI(Human-Agent Interaction) を実現するための技術と課題について解説し，次いで XML により WWW と連携するキャラクタエージェントを用いる新しいマルチモーダルメディアに関する最近の話題について紹介した．MPML についてはホームページ [MPML] でツール類が公開されているので，ご興味のある方はお試し下さい．

謝辞

本稿に紹介した我々の研究は，未来開拓学術研究「マルチモーダル擬人化インタフェースとその感性基盤機能」(H11〜15) により実施しているものである．日頃議論いただく石塚研メンバ，特に Dr. Helmut Prendinger に感謝致します．

◇ 参 考 文 献 ◇

[Allen 01] J.F.Allen, D.K.Byron, M.Dzikovska, G.Ferguson, L.Galescu, and A.Stent: Toward Conversational Human-Computer Interaction, AI magazine, Vol.22, No.4, pp.27-37 (2001)

[Andre 01] E.Andre and T.Rist: Controlling the Behavior of Animated Presentation Agents in the Interface: Scripting versus Instructing, AI magazine, Vol.22, No.4, pp.53-66 (2001)

[Arafa 02] Y.Arafa, K.Kamyab, S.Kshirsagar, A.Guye-Vuilleme, D.Thalmann: Two Approaches to Scripting Character Animation, AAMAS Workshop Notes (W14) on Embodied COnversational Agents: Let's Specify and Compare Them (2002)

[Barakonyi 01] I.Barakonyi and M.Ishizuka: A 3D Agent with Syntactic Face and Semiautonomous Behavior for Multimodal Presentation, Proc. Multimedia Tech. and Applications Conf. (MTAC2001), IEEE Computer Soc., pp.21-25 (2001)

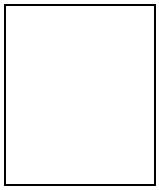
[Cassell 99] J.Cassell and K.R.Thorisson: The Power of a Nod and Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, Applied Artificial Intelligence, Vol.13, pp.519-538 (1999)

[Cassell 00] J.Cassell, J.Sullivan, S.Prevost, E.Churchill(eds.): Embodied Conversational Agents, The MIT Press (2000)
[Cassell 01a] J.Cassell: Embodied Conversational Agents: Representation and Intelligencs in User Interface, AI Magazine, Vol.22, No.3, pp.67-83 (2001)
[Cassell 01b] J.Cassell, H.Vilhjalmsson, T.Bickmore: BEAT: The Behavior Expression Animation Toolkit, Proc. SIGGRAPH 2001, pp.477-486 (2001)
[Chandrasiri 01] N.P.Chandrasiri, T.Naemura, and H.Harashima: Real Time Facial Expression Recognition System with Applications to Facial Animation in MPEG-4, IEICE Transaction on Information and Systems, Vol.E84-D, No.8, pp.1007-1017 (2001)
[DeCarolis 02] B.DeCarolis, V.Carofiglio, M.Bilvi, C.Pelanhand: APML, A Mark-up Language for Believable Behavior Generation, AAMAS Workshop Notes (W14) on Embodied Conversational Agents: Let's Specify and Compare Them (2002)
[Descamps 01] S.Descamps, H.Prendinger, M.Ishizuka: A Multimodal Presentation Markup Language for Enhanced Affective Presentation, Proc. Int'l Conf. on Intelligent Multimedia and Distant Learning (ICIMADE-01), pp.9-18 (2001)
[土肥 96] 土肥浩, 石塚満: WWW/Mosaic と結合した自然感の高い擬人化エージェントインタフェース, 電子情報通信学会論文誌 D-II, Vol.J79-D-II, No.4, pp.585-591 (1996)
[土肥 99] 土肥浩, 石塚満: Face-to-face 型擬人化エージェント・インタフェースの構築, 情報処理学会論文誌, Vol.40, No.2, pp.547-555 (1999)
[Dohi 01] H.Dohi and M.Ishizuka: Life-like Agent Interface on a User-tracking Active Display, HCI International 2001, pp.534-538 (2001)
[Gratch 02] J.Gratch, J.Rikel, E.Andre, J.Cassell, E.Petajan, N.Badler: Creating Interactive Virtual Humans: Some Assembly Required, IEEE Intelligent Systems, Vol.17, No.4, pp.54-63 (2002)
[Haptek] http://www.haptek.com/
[Hasegawa 95] O.Hasegawa et. al.: Active Agent oriented Multimodal Interface System, Proc. IJCAI-95, Vol.1, pp.82-97 (1995)
[HumanML] http://www.oasis-open.org/committees/humanmarkup/
[石塚 00] 石塚満: マルチモーダル擬人化エージェントシステム, システム/制御/情報, Vol.44, No.3, pp.128-135 (2000)
[Johnson 00] W.L.Johnson, J.W.Rickel, J.C.Lester: Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environment, Int'l J. AI in Education, Vol.11, pp.47-78 (2000)
[川本 02] 川本真一 他: 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)
[Loebner Prize] http://www.loebner.net/Prizef/loebner-prize.html
[Marriott 01] A.Marriott: VHML -Virtual Human Markup Language, (Online) http://www.interface.computing.edu.au/documents/VHML/, (2001)
[三吉 01] 三好秀夫: 人とコンピュータとの自然な対話, 映像情報メディア学会誌, Vol.55, No.11, pp.1403-1406 (2001)
[MPML] http://www.miv.t.u-tokyo.ac.jp/MPML/mpml.html
[Ortony 98] A.Ortony, G.L.Clore, A.Collins: The Cognitive Structure of Emotions, Oxford Univ. Press (1998)
[Paiva 00] A.Paiva: Affective Interactions: Toward a New Generation of Computer Interfaces, Affective Interactions - Towards a New Generation of Computer Interfaces, A.M.Paiva (Ed.), Springer-Verlag, pp.1-8, 2000
[Pasquariello 01] S.Pasquariello, C.Pelachaud: Greta: A Simple Facial Animation Engine, 6th Online World Conf. on Soft Computing in Industrial Applications, Session on Soft Computing for Intelligent 3D Agents (2001)
[Picard 97] R.W.Picard: Affective Computing, MIT Press, Cambridge, MA. (1997)
[Picard 01] R.W.Picard and J.Scheirer: The Galvactivator: A Glove that senses and Communicates Skin Conductivity, HCI International 2001, Vol.1, pp.1538-1542 (2001)
[Piwek 02] P.Piwek, et.al.: RRL: A Rich Representation Language for the Description of Agent Behavior in NECA, AAMAS Workshop Notes (W14) on Embodied COnversational Agents: Let's Specify and Compare Them (2002)
[Prendinger 01a] H.Prendinger and M.Ishizuka, Let's Talk! Socially Intelligent Agents for Language Conversation Training, IEEE Transaction on Systems, Man, and Cybernetics - Part A: Systems and Humans, Vol.31, No.5, pp.465-471 (2001)
[Prendinger 01b] H.Prendinger, M.Ishizuka: Social Role Awareness of Socially Situated Agents, Proc. Agents-2001, pp.270-277 (2001)
[Prendinger 02] H.Prendinger, M.Ishizuka: SCREAM: Scripting Emotion-based Agent Minds, Proc. 1st Int'l Conf. on Autonomous Agents and Multi-Agent Systems, (AAMAS-02), pp.350-351 (2002)
[Reeves 98] B.Reeves and C.Nass: The Media Equation, Cambridge Univ. Press (1998)
[SMIL] http://www.w3.org/AudioVideo/
[筒井 00] 筒井貴之, 石塚満: キャラクタエージェント制御機能を有するマルチモーダル・プレゼンテーション記述言語, 情報処理学会論文誌, Vol.41, No.4, pp.1124-1133 (2000)
[Tsutsui 00] T.Tsutsui, S.Saeyor, M.Ishizuka: MPML: A Multimodal Presentation Markup Language with Character Agent Control Function, Proc. (CD-ROM) WebNet 2000 World Conf. on the WWW and Internet (2000)
[TVML] http://www.strl.nhk.or.jp/TVML/
[VoiceXML] http://www.voicexml.org/
[Wilson 00] G.M.Wilson and M.A.Sasse: Listen to Your Heart Rate: Counting the Cost of Media Quality, Affective Interactions -Towards a New Generation of Computer Interfaces, A.M.Paiva (Ed.), Springer-Verlag, pp.9-20, 2000
[Zong 00] Y.Zong, H.Dohi, M.Izhizuka: Multimodal Presentation Markup Language Supporting Emotion Expression, Proc. (CD-ROM) Workshop on Multimedia Computing on the WWW (MCWWW-2000) (2000)

〔担当委員：××○○〕

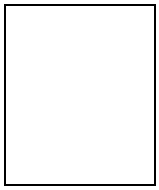
2002 年 9 月 11 日　受理

著　者　紹　介



土肥　浩

1985 年慶應義塾大学理工学部電気卒業. 1987 年同大学院修士課程修了. 同年東京大学生産技術研究所, 同大学院工学系研究科電子情報工学専攻助手を経て, 現在, 新領域創成科学研究科基盤情報学専攻助手. 研究分野はマルチモーダル擬人化インタフェース. ACM, 情報処理学会の会員



石塚　満(正会員)

1971 年東京大学工学部電子卒業. 1976 年同大学院博士課程修了. 工学博士. 同年 NTT 横須賀研究所, 1978 年東京大学生産技術研究所助教授, 同大学院工学系研究科電子情報工学専攻教授を経て, 現在, 情報理工学系研究科電子情報学専攻教授. 研究分野は人工知能, マルチモーダル擬人化インタフェース/コンテンツ, WWW インテリジェンス. IEEE, AAAI, 電子情報通信学会, 情報処理学会, 映像メディア学会, 画像電子学会, 日本顔学会の会員.