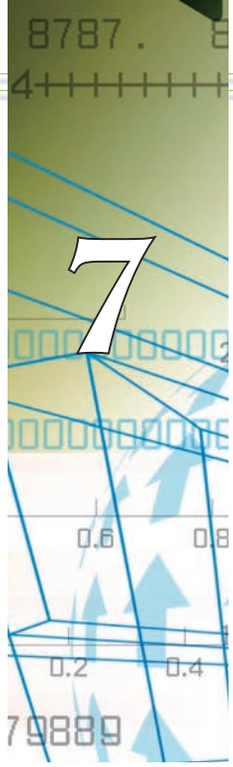


特集 音声情報処理技術の最先端

# 擬人化音声対話エージェント

**嵯峨山 茂樹**
東京大学 大学院情報理工学系研究科
sagayama@hil.t.u-tokyo.ac.jp

**西本 卓也**
東京大学 大学院情報理工学系研究科
nishi@hil.t.u-tokyo.ac.jp

**中沢 正幸**
東京大学 大学院情報理工学系研究科
nakazawa@hil.t.u-tokyo.ac.jp

ユーザと音声で対話をし，表情豊かな顔の動画像を持つ擬人化音声対話エージェントの技術について解説する．機械が人間的な外面とインタフェースを持つことは，今後のインタラクティブな知的なシステムにとって重要である．また，ロボットの知的な面を担う技術としても期待される．各種の開発事例を紹介した後，システムの構築に必要な技術要素と要件について述べる．最小限必要な要素は音声認識，音声合成，顔画像合成，対話制御などである．このような対話システムには相づち，割り込み，表情，ジェスチャなどの特有の特性があり，これらを活かしてこそ効率的な情報伝達が可能なので，人間同士の対話の解析を通して研究を進める必要がある．

## ■擬人化音声対話エージェントの意義

人間と機械のインタフェースは危機的状況にある．ますます複雑化する機械の機能には，従来型のインタフェースではすでに対応しきれない．たとえば，パソコンソフトや携帯電話の機能を全部使いこなしているユーザは少数だろう．

一方，人間は，人間同士の対面コミュニケーションでは高いコミュニケーション能力を発揮している．分厚いマニュアルを読む必要もない．それならば，機械が人間と対面コミュニケーションができるようになればよい．このような視点から，表情豊かな顔の動画像を用いてユーザと音声で対話を行う擬人化音声対話エージェント（Anthropomorphic Spoken Dialog Agent; 以下，擬人化エージェントと略す）への研究には，究極のヒューマンインタフェースとして大きな期待が持たれる．いわば，知的な人間型ロボットのソフトウェア版である．

現段階での擬人化エージェントの主な構成要素は，音声認識，音声合成，顔画像合成，対話制御などであり，さらに視覚入力やセンサ入力などを備える場合もある．そのような技術の応用は広い．オフィス用コンピュータソフトウェア，ゲーム，電子秘書，バーチャルリアリティなどを始め，およそ人間同士の対面コミュニケーションを機械に置き換える場面すべてに応用可能性がある．ネット経由で派遣される擬人化エージェントが抜群の交渉能力を持ち，膨大な契約をとってくるような，カリスマバーチャル店員の時代がきて，擬人化エージェントの開発の成否が各社の命運を分けるようになるかもしれない．また，音声対話機能だけ取り出しても，カーナビなどの車載用途，電話での音声アクセス，視覚障害者支援などへの応用が考えられる．

人間同士の対面コミュニケーションはきわめて高度でかつ効率がよい．人間は単に言語内容のみならず，パラ言語・非言語情報の伝達についても高度な能力を持ち，声の抑揚や顔の表情や身振り手振りで意思を伝え，相手の表情から反応を読む．一方が話している間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

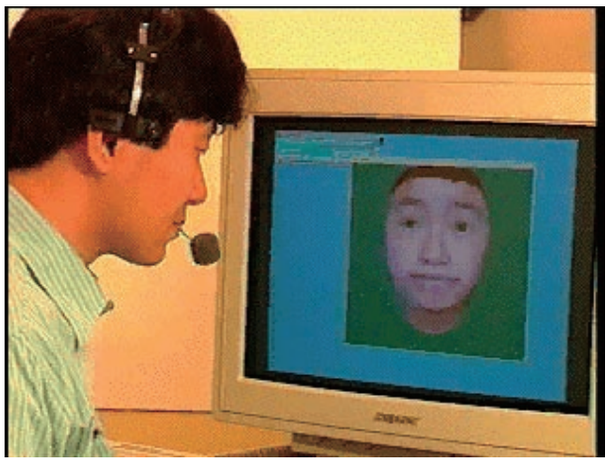

図 -1　擬人化エージェント Talkman（SONY）
http://www.csl.sony.co.jp/

図 -2　マルチモーダル対話システム Mr. Bengo
http://www.ntt.dis.titech.ac.jp/ja/research/multimodal.html

聞き返す．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しやすい．擬人化エージェントがそのような機能を持てば，人間同士の対話並みにコミュニケーションの効率と質を向上できる可能性を持つ．

擬人化エージェント技術は人間型ロボットの構成要素としても有用である．人間型ロボットには少なくとも身体運動能力，感覚・コミュニケーション能力，知的な思考判断能力など3種類の能力が必要であろう．現在，身体運動能力のような筋肉と小脳に対応する部分に関心が集まっているが，本来は感覚・コミュニケーション能力のような神経と大脳に対応する部分にももっと注意が向くべきだろう．擬人化エージェントは知的なロボットの実現にも必要な技術である．

擬人化エージェントの研究への期待は，音声認識や音声合成の研究者の側にもある．音声認識技術は高度に進歩したが，その割には期待したほど使われていないのはなぜか，という議論[1]の中で，音声対話は人間的なものであって，人間はあまりに機械的な機械と対話する気がしないのではないか，という指摘があった．究極的には，機械が知性と人格と個性と心を持たなければならないのではないか．たとえば電子秘書が時には頑固で時には思いやりがあるような個性ある一個の人間と感じられるとき，人間は心から対話をする気が起こるのではないか．すぐには無理としても，とにかく研究を開始して着々と進めなければならないのではないか．

人間同士の対話は人間の知的活動の中で大きな部分をなし，高度な創造活動の源泉となっている．究極的には，あるときは人間と打々発止と議論して人間の思考力や創造力を極限まで引き出し，あるときは優しく接して人に安らぎを与えるような機械が理想である．

## ■研究事例

以下では，音声対話システム事例を中心に，擬人化エージェントの研究を概観する．

東芝による TOSBURG II では，実環境を想定した頑健な音声認識を実現すると同時に，キャラクタによるアニメーションと合成音声による応答を行っている．また，SONY CSL の Talkman[2] では，連続音声認識技術による音声対話と，3次元的な顔のグラフィクスによる多様な表情のリアルタイム生成を実現した（図-1）．対話の典型的な状況として「話題の導入」や「意見の強調」などの分類を行い，これらに「すくめ顔」や「不安の表情」などを対応させ，顔の部品の動きに変換することで対話に合わせた表情生成を行っている．

Web ブラウザを音声で制御し，Web ページから取得した音声を読み上げることができるシステムとして，土肥・石塚による VSA[3] がある．また，Microsoft Agent は Internet Explorer での擬人化エージェントの利用を容易にした．擬人化エージェントと WWW を連携させることにより，音声対話システムを通じて膨大かつ最新の情報を得ることができる．

Mr.Bengo は，法律に関する専門知識を持ち，裁判における論争の支援を行うマルチモーダル対話システムである（図-2）．画像認識により利用者を識別し，裁判の有利・不利などの状況を擬人化エージェントの表情を通じて伝達することができる．また，MIT の REA（Real Estate Agent）は音声と身体動作を伴い不動産の売り込みを行う（図-3）．

NTT は擬人化エージェントを用いた音声対話システム DUG-1 および飛遊夢（図-4）において，ユーザの話す

図-3　Real Estate Agent (MIT Media Lab.)
http://gn.www.media.mit.edu/groups/gn/projects/humanoid/

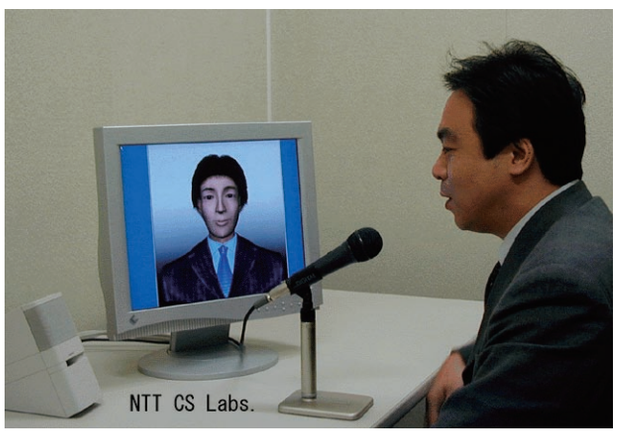


図-4　音声対話システム飛遊夢（NTT）
http://www.brl.ntt.co.jp/cs/humanoid-tech/index-j.html

図-5　インタフェースエージェント
MAICO（シャープ）

図-6　The CSLU Toolkit
http://www.cse.ogi.edu/research/


図-7　The August Spoken Dialogue System (KTH)
http://www.speech.kth.se/august/

内容を逐次的に理解して対応するなど，柔軟なやりとりや効率的な対話制御を実現している．また，自由発話入力によって擬人化エージェントと対話を行うことができるサイバーアテンダント[4]が開発されている．

シャープのインタフェースエージェントMAICOは，力学系を模した相づち生成のタイミング制御モデルを導入して[5]，ユーザの語りかけに対して「はい，ええ」などの相づちを交えて表情豊かに応答する（図-5）．

音声対話を実現するツールキットやフレームワークにおいては，後述するGalatea Toolkitのほかに，CSLU Toolkit（図-6），KTH August（図-7），MIT GALAXY-IIなどに擬人化エージェントが導入されている．また，擬人化エージェントを利用するコンテンツを誰もが容易に作成し視聴できることが重要である，という立場から，会話型知識流通システムPOC（Public Opinion Channel）や，コンテンツ記述言語MPML（Multimodal Presentation Markup Language）などの研究が行われており，教育，娯楽，コミュニティ支援など多分野で利用されている．

擬人化エージェントの制御は「人工の心」の実現にも繋がる．オムロンによるインタラクティブエージェントStellar[6]では，欲求と感情を価値判断に用いる心のメカニズムと，選択的注意や反射と熟考のプロセスを処理する意識のメカニズムを用いている．また，前述したMPMLと連携する人工感情モジュールSCREAMでは，感情モデルOCCに準拠してエージェントの感情を計算している．

なお最近，擬人化エージェントの研究を集めた書籍[7]が刊行されており，エージェントの動作の記述言語の議論も行われている．

## ■擬人化エージェントの構成

著者らが開発に参加し無償公開を行っているGalatea Toolkit[8]を例として，擬人化エージェントの構成を図-8に示す．このシステムは，音声認識，音声合成，顔画像合成（あるいは身体画像合成），統合制御および対話管理などの要素技術からなる．本ソフトウェアの入手方法は次のURLを参照していただきたい．

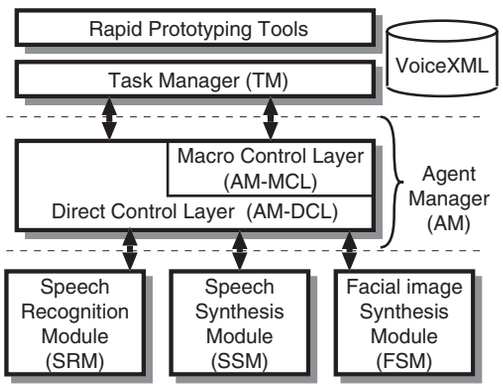


図 -8　擬人化エージェントの構成（Galatea Toolkit）

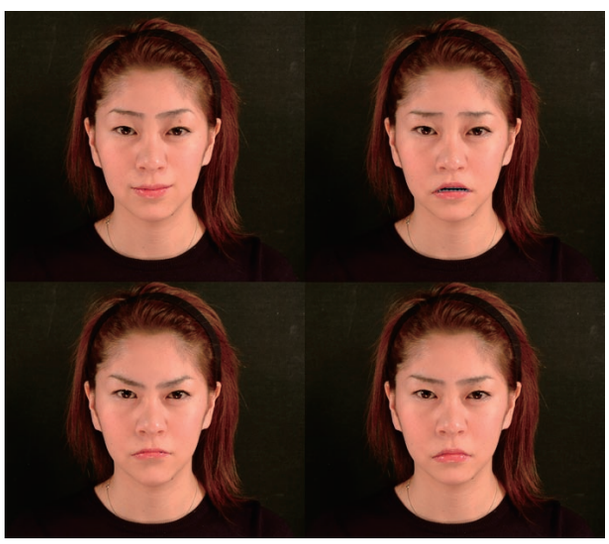

図 -9　顔画像合成モジュール（Galatea FSM）の表情合成例

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

以下，各要素について述べる．

## 顔動画像合成

顔動画像合成において，人間に似せた顔画像を表示し，音声合成に合わせて唇を動かす機能は必須である．さらに，表情（喜びや驚きなど）や身体動作（視線や頷きなど）を生成すれば，対話の状況に応じて有効に利用できる．さまざまな顔を合成でき使い分けられることも望ましい．キャラクタとしては，手作りのコンピュータグラフィクス（CG）のアニメーションがよいか，フォトリアルな人物画像がよいか意見が分かれる．ただし，前者は1体あたり数百万円ともいわれる作成コストが問題である．

Galateaの顔画像合成モジュールにおいては，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マウス操作で整合させ，1枚の人物写真から表情変化が可能なキャラクタを作成できる．表情としては怒り，喜び，悲しみ，驚き，嫌悪，恐れの6種を用意している（図-9）．また，音声合成と唇画像の同期のほか，自律的な動作（頷き，まばたき等）の機能も持つ．

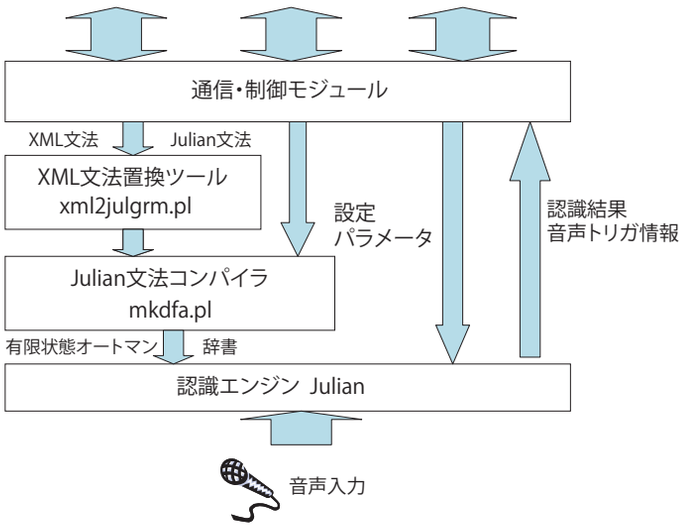


図 -10　音声認識モジュール（Galatea SRM）の構成

## 対話音声の認識

擬人化エージェントの音声認識においては，システムの動作中に認識処理の中断や文法の切り替えを行うなどの動的な制御が必要である．また，ユーザの発話内容に応じて即時にエージェントの表情を変えたり頷いたりする，というような処理のために，ユーザ発話の終了を待たずに逐次的に認識結果を得る機能が求められる．Galateaにおいては，連続音声認識エンジンJulianを用いて前述した機能を実現している．音声認識モジュールの構成を図-10に示す．

## 対話音声の合成

擬人化エージェントの音声合成においては，形態素解析テキスト音声合成の基本的な機能に加えて，顔画像との同期のために，各音素の継続時間長を出力する機能が必要である．また，ユーザの割り込み発話を許すために，合成音出力の中断が可能でなければならない．さらに，出力テキストにタグを埋め込み韻律や発音の詳細な制御を可能にすることが望ましい．また，多くの話者の種類を作れる機能も必要である．

Galateaでは，これらに対応した音声合成モジュールとしてGalateaTalk（図-11）を新たに開発した．GalateaTalkはHMMに基づく音声合成方式を採用しており，話者適応によりアプリケーション開発者が独自の話者を作成できる．詳細については本特集所収の解説をご覧いただきたい．

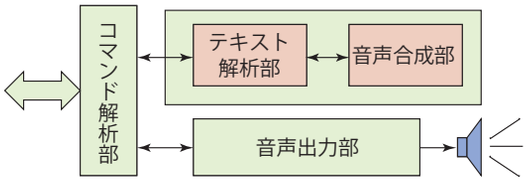


図-11　音声合成モジュール（GalateaTalk）の構成

```
<form id="main"> <field name="place">
  <prompt>
    <emotion type="HAPPY">場所をどうぞ．</emotion>
  </prompt>
  <grammar><rule><one-of>
    <item><token sym="とうきょう">東京</token></item>
    <item><token sym="きょうと">京都</token></item>
  </one-of></rule></grammar>
  <filled>（入力に応じた処理）</filled>
</field> </form>
```

図-12　VoiceXML による対話の記述例

## システム統合と対話の制御

擬人化エージェントのシステム統合においては，前述した各モジュール間の通信を行い，応答の出力時にはテキスト音声合成と唇画像を同期する処理が必要がある．さらに，タスクに応じた処理を行い，音声認識などのユーザ入力に応じて応答文や表情などを決定する対話管理も必要である．タスクの実装や記述は効率的に行えることが望ましい．

Galateaにおいては，統合制御モジュール（AM）が，各サブモジュールに固有の命令セットを使用して直接制御するレイヤと，唇同期などの同期処理など，対話管理に有用なマクロ命令を利用して制御できるレイヤの2つを提供している（図-8）[9]．また，アプリケーション開発者が対話タスクを記述するための手段としては，VoiceXMLに基づくLinux対応の対話処理系があり[10]，顔表情制御などのタグを付加した記述が可能である（図-12）．さらに，マークアップ言語を必要とせずにタスクを定義する手段として，Windows対応のラピッドプロトタイピングツール（図-13）が提供されており，その実装はモダリティの追加が可能なマルチモーダル対話向け言語XISL[11]に基づいている．

## その他の要素

以上に述べた必要最小限の構成要素に加えて，擬人化エージェントのキャラクタとして，顔のみならず上半身や全身の呈示，複数のキャラクタの同時表示，画像やテキストなどの出力との併用を，また，画像入力を構成要素に加えて，ユーザの顔の識別や認証，ユーザと視線を合わせるアイコンタクト，ユーザの頷きの認識などを考えることもできる．これらをユーザの発した意味や感情として統合的に解釈し，擬人化エージェントが表現すべき意味的内容や感情を決定し，音声合成や顔合成などの出力制御を行う，というようなマルチモーダル対話処理の枠組みに関する提案[12]では，図-14のように，モダリティ個別の処理，モダリティ統合の処理，モダリティ非依存の処理に分離し，それぞれを感覚，認知，意図のレベルに対応づける．

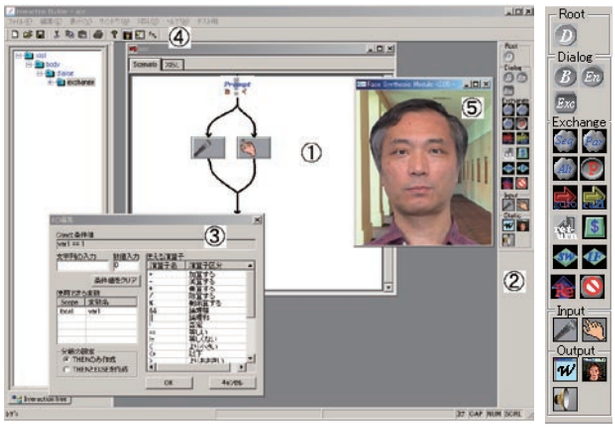


図-13　Galatea-IB の実行画面と対話部品バー

人間は統一性のある人格を持ち，他と接する．擬人化エージェントにおいて最終的に最重要な部分は知的な情報処理機構とともに，統一のとれた感情や人格や個性の発現であろう．

# ■今後の課題と展望

擬人化エージェント技術は，人間同士の対面対話に近づける目標を持つゆえに，特有の課題が浮かび上がってくる．

## 相づち・割り込み・マルチモダリティ

前述したように，音声対話は大変効率がよい情報伝達手段である．その大きな理由は，相づちや頷き，即時的な聞き返し，表情や態度などの高度なコミュニケーション能力にある．従来の音声対話技術のように相手の発声終了を待って疑問を呈する場合は，どの語が聞き取れない（認識できない）かを伝えるための対話が必要なのに対し，実際の人間同士の対話では，わずか「え？」という一語でどの語が不明であったかが伝えられる．

さらに，（顔と声の）表情，身振り，指示動作，心理表

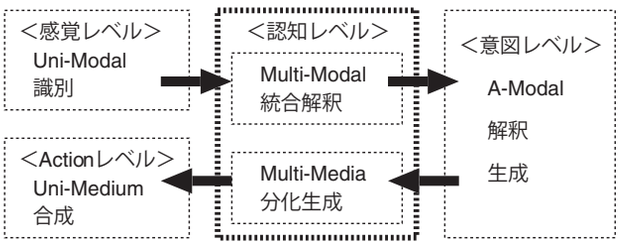


図 -14　意図理解・表出の階層モデル [12)]

出（そわそわ，落ち着き，無関心など），視線制御（驚いて相手を見る，視線をそらすなど），返答のタイミングと韻律，相手の発話中の注釈発言など，音声対話は多様な面を持ち，それゆえに大変効率がよいチャネルである．このような特性を活かしてこそ，音声対話の有利さが得られるのであり，この面の研究が今後重要である．

## マルチモーダル対話データの必要性

以上のような機能を実現するためには，人間同士の自然な対話における身体動作を分析する必要がある．

前述したMAICOの構築においては，ハーフミラーを搭載したモニタを2台用いて，別々の部屋にいる2人の正面からの画像をハーフミラー後方のカメラで収録し，光学式モーションキャプチャシステムを導入し，被験者の体につけた18個のマーカの位置を時間情報とともに記録した．

また，ハーフミラーを用いた顔画像付きの音声対話は千葉大学でも収録されており，音声情報に加えて相づち，頷き，視線活動（相手を見ているか，目をそらしているか）などのアノテーションが行われている．話し手が聞き手を見ながら発話単位終了時に頷くとき，聞き手がその頷きに誘発されて相づちや頷きを行う，などの現象が指摘されている．

音声対話における雰囲気に着目し，擬人化エージェントに対話を活性化させる仕草をさせる研究[13)]では，日常生活の中でよく遭遇し，ある程度持続する雰囲気として，「上機嫌」「不機嫌」「納得」「納得がいかない」「興味がある」「興味がない」の6つを分析の対象としている．

人間が発話中に行う頭部や目の動きなどを手作業で分析し，擬人化エージェントで再現することによって，これらの要因を個別に検討したり相互効果を評価することができる．Galatea Toolkitを用いた研究[14)]においては，映像の自然性に特に頭部挙動が影響を与えることが示唆されている．

## マルチモダリティ

擬人化エージェントの研究における今後の課題としては，感情豊かな発話に対応する音声合成，自然な表情変化や視線活動などを伴う動画像生成，状況やユーザの特性に応じた適応的な行動，入力に対する即時的かつ適切な応答，マルチモーダル入出力の効果的な利用，計算論的な心のモデルなどが考えられる．一方で，自律移動できるスクリーンに表示された擬人化エージェントはソフトウェア的なロボットと見なすことができ，高速な動作や精緻な表現などの利点が期待できる．

## システム開発ツールキット

このような研究のためには，さまざまな要素技術を自由に改良・改変でき，自由に組み合わせることができる柔軟なツールが必要であるが，単一の研究機関で開発・展開するには負担が大きい．著者らが参加している音声対話技術コンソーシアム（ISTC）ではGalatea Toolkitの開発を通じて，擬人化エージェントの研究開発を推進している．詳細は下記のURLを参照していただきたい．

`http://www.astem.or.jp/ISTC/`

**参考文献**


1）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？，情処研報，94-SLP-1-4, pp.23-30 (May 1994).
2）Nagao, K. and Takeuchi, A.: Speech Dialogue with Facial Displays: Multimodal Human-computer Conversation, Proc. ACL-94, pp.102-109 (1994).
3）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，信学論，Vol.J79-D-II, No.4, pp.585-591 (1996).
4）甘粕哲郎，山本俊一郎，小川厚徳，篠崎　翼，平沢純一，山口義和：サイバーアテンダント－自由発話入力に対応したマルチモーダル対話システム，音講論，3-Q-32, pp.207-208 (Mar. 2004).
5）向井理朗：非言語情報を用いたマルチモーダル対話インタフェースの試作，インタラクション2001論文集，pp.139-140 (2001).
6）牛田博英，平山裕司，中嶋　宏：心のモデルに基づくインタラクティブエージェント，第4回知能情報メディアシンポジウム講演論文集，pp.165-172 (1998).
7）Prendinger, H. and Ishizuka, M. (Eds.): Life-Like Characters - Tools, Affective Functions and Applications, Series: Cognitive Technologies, Springer (2004).
8）西本卓也 他：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).
9）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，情処論，Vol.43, No.7, pp.2249-2263 (July 2002).
10）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 (2003).
11）桂田浩一，中村有作，山田　真，山田博文，小林　聡，新田恒雄：MMI記述言語XISLの提案，情処論，Vol.44, No.11, pp.2681-2689 (2003).
12）新田恒雄：マルチモーダル対話の深化と記述言語の今後，情処研報，2004-SLP-50-3, pp.15-22 (Feb. 2004).
13）伊藤秀樹，重野真也，西本卓也，荒木雅弘，新美康永：対話における雰囲気の分析，情処研報，2001-SLP-40-18, pp.103-108 (Feb. 2002).
14）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，音講論，1-8-23, pp.51-52 (Mar. 2004).


（平成16年7月13日受付）