音声言語情報処理 33－1
（2000．10．27）

擬人化音声対話エージェント開発とその意義

嵯峨山 茂樹 †‡　　中村 哲*
† 北陸先端科学技術大学院大学 情報科学研究科 / ‡ 東京大学大学院 工学系研究科
〒 923-1292 石川県能美郡辰口町旭台 1-1 / 0761-51-1225 / sagayama@jaist.ac.jp
〒 113-8656 東京都文京区本郷 7-3-1 / 03-5841-6900 / sagayama@hil.t.u-tokyo.ac.jp
* ATR 音声言語通信研究所
〒 619-0288 京都府相楽郡精華町光台 2-2 / 0774-95-1370 / nakamura@slt.atr.co.jp

概要　擬人化音声対話エージェントの基本ソフトウェアの開発に向けて、その意義を議論する。IPA からサポートを受けて 3 年間の予定で開発に着手した。構成要素は、音声認識、音声合成、顔画像生成、エージェント統合処理の 4 部分である。IPA プロジェクトで開発された顔画像生成、音声認識の無償公開ソフトウェアを利用し、音声合成を新たに開発し、これらを統合する処理を加える。ヒューマンインタフェース危機への解決方向、産業応用の可能性、今後の音声技術研究の問題発掘、モルチモダリティの研究、研究キットとプラットフォームの提供などの多くの意義を持つ。

Developement of Anthropomorphic Dialogue Agent: a Plan and Its Significance

Shigeki Sagayama†‡,　*Satoshi Nakamura**
† Japan Advanced Institute of Science and Technology / ‡ University of Tokyo
1-1, Asahi-dai, Tatsu-no-kuchi, Nomi-gun, Ishikawa, 923-1292 Japan
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
* ATR Spoken Language Translation Research Labs
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan

Abstract　A plan of anthropomorphic spoken dialogue agent development is presented and its significance is discussed. With a 3-year financial supoport from IPA, a research team launched a project. The system will consists of four modules: speech recognition, speech synthesis, facial animation, and agent integration. Fully utilizing open softwares for speech recognition and facial image generation already developed under IPA supports, speech synthesis and agent integration will be newly developed. Significance of this plan is discussed in many aspects including solution to human interface crisis, industorial applications, finding new speech research issues, multimodality research, and open source provision of a software kit and platform for future research.

1　はじめに

音声認識および音声合成の研究開発は、朗読調の音声に関してはかなり実用レベルに達して来ており、実際、商用製品も多く販売されるようになって来た。電話回線を利用する情報案内などの自動化は企業レベルで多くの実用例が見られるようになった。

この時期に、音声認識・音声合成の研究者のレベルでは、どのような研究内容を次の研究のターゲットとするべきだろうか。この問いに対するさまざまな回答として、次の時代を準備するさまざまな研究やデータベース作成の努力が行われている。ここで原点に戻って考えてみたい。音声認識や音声合成は、「人間と機械が人間同士のように対話をする夢の技術」を実現するために始められたのではなかったか。音声認識や音声合成は、着実にその夢の実現に向かって準備しているだろうか。実際は、企業レベルでは製品開発や実サービスへの適用などに精力が傾けられ、夢の実現への歩みは遅くなっているように見える。

このような中で情報処理学会 音声言語情報処理研究会の「マルチモーダルツールワーキンググループ」(1998-2000) では、今後の音声研究者の研究目標をどのように持ち、進めればよいかを議論してき

た。そして、次世代の研究ターゲットとして擬人化エージェントを構想し、その研究プラットフォームを研究者の共同作業により構築して公開する計画を持った。この構想は、情報処理技術振興協会 (IPA) のサポートを受け、十数名の研究者により実行段階に入っている。

本報告では、この構想の意義と計画概要とこの分野の研究課題について述べる。

2 擬人化対話エージェント構想

2.1 擬人化対話エージェント研究の必要性

ここで言う「擬人化対話エージェント」とは、顔 (あるいは身体) の動画像と、音声認識、音声合成の機能を持ち、対話制御のもとで人間のユーザと音声言語で対話する機能を備えたインタフェースである。このような擬人化対話エージェントの研究には、多様な意義がある。そのいくつかを以下に挙げてみよう。

(1) ヒューマンインタフェース危機の解決に向けて

いままでの情報処理機器 (コンピュータなど) は、コマンドライン・ユーザ・インタフェース (CUI) からグラフィカル・ユーザ・インタフェース (GUI) へと進んで、使いやすさは格段に向上し、ユーザ層が広がった。しかし、アプリケーションの機能の増加とともに操作は複雑になり、ユーザが望む結果を得るには、厚いマニュアルを読んで知った機能を組み合わせてどのように操作すればよいか考えなければならなくなった。高度な情報機器を使いこなせるかどうかが、社会的な格差を生じる恐れも危惧される。高齢化社会を目前に、このようなヒューマンインタフェース危機に対処するには、GUI より先のインタフェース形態の研究を本格的に始める必要がある。

その一つの候補は、より人間的なインタフェースである。「機械を操作する」という形態でなく、人間同志の対面会話に近い形態で、人間と擬人化された機械が音声言語で情報をやりとりし、意思を伝えあうことにより、ヒューマンインタフェース危機の解決に向けて前進できる可能性がある。

(2) 産業応用可能性の面から

擬人化された音声言語インタフェースは、完成度が高くなくとも、それなりの応用分野が想定できる。新しいタイプのゲーム、極めて限られた場面での質問応答、情報提供、電子秘書などが挙げられよう。発展段階に応じて、産業的な応用が期待できる。

技術が進歩すれば、サイバースペース店員 (図 1) など、人間的な販売エージェントにより効率的なビジネスの可能性もある。また携帯機器の進歩の一つのゴールとして、人間的なパートナーとして感じることができるような機械が考えられるだろう。

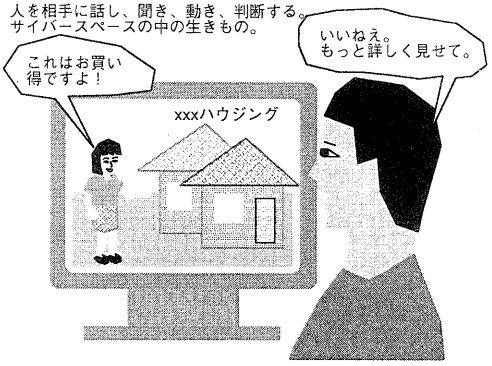

図 1: 業界売り上げ No.1 のサイバースペース名物店員 (未来の夢)

(3) 今後の音声技術研究の問題発掘

朗読調の音声認識技術では、最近の主流である統計的な手法による限り、広範な音声および言語データをより多く集めた者が最も高い性能を得ることができるような現象が見られる。また音声合成でも大規模なデータベースを利用する手法が使われている。このような手法でこのような分野だけを見ていると、知恵比べの時代が終わり、物量競争の時代に入ったようにも見える。ここで大学やアルゴリズムの研究者の役割は何なのか。

その一方で、音声認識および音声合成の技術は、いまだに未熟なレベルにあると感じている研究者も多い。人間のもつ音声言語能力にはいろいろな点で遥かに及ばず、理想的なヒューマンインタフェースの実現には程遠いと感じている。ヒューマンインタフェースの観点から見て、現在の技術では何ができるようになって、何ができないのか。

ここで、音声認識・合成の研究者らがその研究の効用として主張して来た人間と機械の対話に関して、どこまで人間らしく機械が振る舞えるようになって来ているのか、何ができていないのか、今後何が必要なのか見通す必要があるだろう。音声研究者の立場から、音声認識・音声合成の今後の研究のあり方を考える上で、朗読音声の認識・合成よりさらに人間に近いレベルで人間・機械の対話を目標とすることは、一つの理想である。

最新の技術レベルの音声認識については、IPA プロジェクト「日本語ディクテーション基本ソフトウェアの開発」(1997–2000) により、高度なソフトウェ

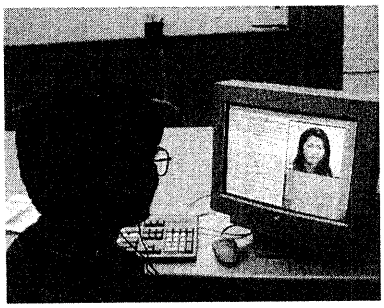

図 2: 擬人化エージェントシステム研究の雰囲気

アが無償公開されている。このような現時点の技術を組み合わせることにより擬人化対話エージェントを構築し公開して、問題点を明らかにすることは、今後の研究のために重要である。現在の技術の組合せでは、人間らしい機械の実現に向けてどのような研究が不足しているか、一目瞭然になるだろう。

(4) マルチモダリティの研究として

一方、顔画像の生成などの研究も進歩した。たとえば、IPA プロジェクト「感性擬人化エージェントのための顔情報処理システムの開発」(1995–1998) により、高度なソフトウェアが無償公開されている。このような状況により仮想的な人間の実現の可能性が増して来ている。音声認識と音声合成に顔の動画像生成を組み合わせて統合して、上記の夢の実現に継げられないだろうか。このような擬人化エージェントの研究はすでにいくつか行われているが、まだ検討の余地を多く残している。また、このような問題は、音声技術と画像技術のみならず、他のモダリティ、知能処理、知識処理、心理学や通信まで含めて、多くの分野から多くの研究者が参加して、総合的に進められなければならない。このためには、共通の研究基盤となるプラットフォームあるいはプロトタイプ (図 2) が必要である。現在のところ、最新の技術を組み合わせ、かつソースコードを公開するようなオープンプラットフォームは存在していない。

(5) 研究キット、プラットフォームとして

オープンなプラットフォームが実現されれば、その部分的な改良がどのように全体に寄与するのかが分かりやすくなる。また、画像認識をはじめとしてさまざまなモダリティやモジュールを追加して、発展させることもできる。研究の問題の発掘とともに、研究成果の実現の場として、大きな意義がある。また、音声技術や画像技術以外の分野 (たとえば心理学、言語学など) の実験的研究のツールとしての用途の可能性もある。

2.2 擬人化対話エージェントの夢

人間のように知的にユーザと会話ができ、人間の姿・形を持つロボットは、音声研究者のみならず人工知能関連分野の研究者の夢である。研究プラットフォームとして、「人間性豊かなサイバー人間キットが欲しい！」という素朴な気持を持つ研究者・技術者は、音声技術関係者のみならず広く存在するだろう。擬人化エージェントの究極の目標は、「個性と人格を持つ機械」であろう。人間的な存在感があり、自らの意志や嗜好があり、

- 話す・聞く (→ 音声認識・音声合成)
- 笑う・怒る・うつむく (→ 表情画像・動作画像)
- 感情・個性・人格 (→ 統合動作)
- サボる・働く・遊ぶ (→ タスク制御)

のような機能を持つことは、研究者としての夢である。その先には、ゲーム・遊び・仕事の融合、役立つ機械／役立たない機械の境界消滅、雑談・用件・愚痴などが考えられる。(考えるヒントとして、たまごっち、ポストペット、シーマン、AIBO、Star Wars の R2D2 の人格などが参考になる。)

3 擬人化音声対話エージェント基本ソフトウェアの開発計画

3.1 擬人化音声対話エージェントの構成要素

以上に述べたような動機で、現時点での要素技術を組み合わせて擬人化音声対話エージェントのオープンプラットフォームを提供するプロジェクトを開始した。その構成要素は、図 3 に示すように

- 擬人化音声対話エージェント統合基本ソフトウェア (新規開発)
- 対話音声合成基本ソフトウェア (新規開発)
- 対話音声認識基本ソフトウェア (IPA「日本語ディクテーション基本ソフトウェアの開発」(1997–2000) 成果を改造)
- 顔画像合成・制御ソフトウェア (IPA「感性擬人化エージェントのための顔情報処理システムの開発」(1995–1998) 成果を改造)

の 4 部分である。これらの部分を、関連分野を専門とする研究者が集結して構築する。

3.2 擬人化音声対話エージェント統合基本ソフトウェアの開発

動画像生成・音声認識・音声合成を統合し、それらの間に同期性と統一性を持たせるように、モジュー

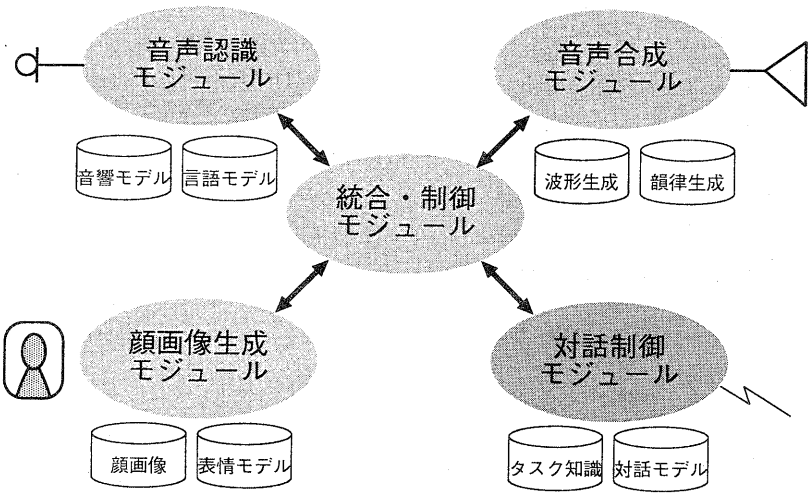

図 3: 擬人化音声対話エージェントの構成要素とモジュール間通信

ル間の通信及び統合制御を行う部分である。通信プロトコルと、制御内容の統一性が大きな研究開発課題となる。各モジュールは、将来置き換えられ得るように、柔軟な結合を目指す。このために、FIPA による分散エージェント技術の標準化や、対話システム記述言語 VoiceXML など、インターネットにおける世界標準の動きを注視しながら、モジュール間通信及び外部アプリケーションとの通信方式を決定する。一方、表情生成と音声合成は統一を取って、怒った顔が笑い声で話すようなことが起こらないようにする必要がある。このために、基本的統一制御のメカニズムが不可欠である。API を通して外部から制御を行うには、このような統一制御モデルへのマクロ指令と、それより細やかな個別制御を用意する必要がある。また、実際的な場面設定として、情報案内的なタスクと販売問答的タスクを実現例として、システムの検証を行うとともに、プロトタイプとして提供する。

なお、知能的な対話や推論などを含む、いわば「頭脳」や「魂」に相当する部分は、今回の計画には含めない。

3.3 対話音声認識基本ソフトウェア開発

IPA「日本語ディクテーション基本ソフトウェアの開発」(1997.4～2000.3) で開発されたソフトウェアパッケージを基盤にして、対話処理を考慮した機能拡張や、柔軟な制御を可能にする。具体的には、文法に基づく認識（タスクに応じた文法を用意することで認識できるようにする）、認識結果の棄却、不要語やポーズなどへの対応（タスク外の発話について、それを検出・棄却できるようにする）、認識処理の動的な制御（音声認識のオン、オフや、割り込み、時間切れなどに対応できるような動的な制御を可能にする）などの機能を実現する。

3.4 対話音声合成基本ソフトウェア開発

音声合成の研究は多数の研究機関で行われており、また、種々の商品が提供されているものの、フリーでエージェント開発に直接利用できる音声合成プログラムはない。音声合成プログラムは、テキスト解析部、韻律生成部、合成器の三つのモジュールで構成する。文で与えた発話内容を音声波形に変換するだけでなく、合成プログラム内部で決定した時間情報などの合成パラメータを外部へ提供することにより、顔画像生成プログラムとの同期を可能とする。各モジュールは独立に機能するように設計し、音声合成の研究においてモジュール単位で異なる方式の評価が行なえるようにする。テキスト解析では、「日本語ディクテーション基本ソフトウェアの開発」で開発された読みを与える形態素解析プログラム chasen を改良し、単語のアクセント型も得られるようにする。

3.5 顔画像合成・制御ソフトウェア開発

IPA「感性擬人化エージェントのための顔情報処理システムの開発」(1995.6～1998.3) で開発されたソフトウェアパッケージを基盤にして、よりクオリティの高いエージェントの表情合成とアニメーション制御、さらに従来にはなかった合成音声あるいは自然音声との唇の同期を実現する。具体的には、複

平常の顔　　怒った顔　　うんざり顔

図 4: 顔画像生成による表情生成の例

数方向から撮影した頭部画像に対して、標準ワイヤフレームを整合させ、3 次元の個人モデルを容易に作成するための GUI の構築、テキスト入力により生成された合成音声や自然音声と唇アニメーションとの同期の確立を目指す。さらに、成蹊大で研究されてきた顔表情制御方式 (図 4) を導入、拡張し、任意の表情を付加できるようにする。これには、3 次元の変形規則を導入する。また、ノンバーバルな情報として重要な頷きの制御、瞬きの制御を可能にする。最後にテキスト入力と表情付加のタイミングを制御し、任意の擬人化エージェントのアニメーションを作成するための GUI を開発する。

4 擬人化音声対話エージェントの将来

4.1 予想される問題

ひとたび擬人化エージェントを構想すると、欲しくなる機能には限りがない。2,3 年の期間で実際に実現できるのはそのうちのごく僅かな部分に過ぎない。しかし、このような問題の発掘は擬人化エージェント構想の大きな意義の一つである。実現の仕方も、おざなりなレベルから本格的研究までいろいろ考えられよう。ここでは、どのような問題が考えられるかを、(実現性を無視して) 若干挙げて見よう。これらのうちいくつかでも本計画の中で着手できれば幸いと考えている。

4.2 時間特性・同期・無矛盾に関する問題

マルチモーダルなシステムでは、時間的な同期やモダリティ間の矛盾解消などの新たな問題が浮上する。音声対話エージェントに関しても以下のような問題が予想される。

- モダリティ間の同期、無矛盾
  3 つのレベル: (1) 音声発声と音声器官動画像の時間同期、(2) 発話内容と音声器官形状の対応、(3) 感情表現を含む言語表現、合成音声の音色と韻律、顔画像の表情や動作がすべて矛盾しないこと (e.g. 怒った顔で、笑って話すような矛盾を避ける)、などが要請される。
- ユーザの割り込み
  以下の例 (U:ユーザ、S:システム) のように、

S: その件は 嵯峨山さんに...	あの、北陸の嵯峨山...
U: (割込んで)え、誰？	うん。

  システムとユーザは、対話の中で、相手の発話中に割り込むことがあり得る。
- システムの割り込み
  一方、未知語などを検出してシステムが割り込む例

U:明日、プー太郎 にね、嵯峨山くんにさ、伝えて。	
S: (頷く)	はあ？　　はい。

  や、逐次認識した結果に基づいて遮る場合

U: 社長には 12 日に会う約束...	
S:	12 日は日曜日ですけど。

  を可能にすることが望ましい。

4.3 エージェント統合に関する問題

基本的な必要事項: 各モジュール間を統合・通信・制御、柔軟なモジュール結合、通信プロトコルの設計・実現、制御内容の統一性・同期性の実現に加えて、

- シナリオに従う制御
- タスクドメインの記述 (VoiceXML の動画像つきへの拡張など)
- プロトタイプ開発ツール (GUI による rapid prototyping)
- 自律的な知能的対話のための自然言語処理、推論
- 人工的な人格や個性の演出、その統一原理

なども必要となる。但し、今回の計画では知能処理には手をつけない。あくまで外面的・表層的なツールキットの提供に留める。

4.4 音声認識に関する問題

基本的には、対話音声の認識、文法に基づく認識 (タスクに応じた文法を用意することで認識できる)、認識結果の棄却、不要語やポーズへの対応 (タスク外の発話の検出棄却など)、認識処理の動的な対応 (認識 on/off、割込み、時間切れなどの制御) が必要である。これらに加えて、音声対話においては、ユーザの非言語的情報が有用になる可能性があり、それを取り込めるような機構を持つことは、今後の技術展開のために望ましい。将来の問題としては、

- 認識内容逐次出力：文節ごと程度のタイミングで、複数候補、信頼度→ うなずき、問い返し、首かしげ、聞いているという表情などに利用。
- Barge-in 対策 (→ 音声合成中に話し始められる)
- 音響モデル、言語モデルの切替え機能を可能にするメカニズム
- 相槌（「はい」「うん」)、問い返し（「えっ？」）の高速認識
- いい淀み（母音継続、ポーズ）対策と検出
- Supra-segmental Features (韻律、ポーズ、いい淀み、速度変化) の検出 (→ 強調、非言語的(自信の有無)、心理的状態、感情認識)
- Office 環境での Hands-free 入力
- 話者交替検出 (→ 複数ユーザ対話、話題の切替え、より高度な対話)
- 音声／非音声識別、物音認識
- 発声速度検出 (→応答音声の速度を合わせる)

などの解決が必要になろう。

4.5 音声合成に関する問題

オープンソースの日本語 text-to-speech 音声合成システムはまだ存在していないので、これを実現することが望ましい。さらに、感情を表現できる音声合成 (韻律制御)、画像同期情報の生成 (→ 唇の動き制御)、

- 音節程度の単位で生成可能、生成完了の acknowledge を返す。(制御可能（中止、一時停止、破棄、acknowledge)) → ユーザ割り込みに対応、lip-sync
- 音素列、韻律指定、速度（音節長）指定、音色（感情表現）指定を反映した合成音声信号を生成すること
- 話者切替え機能
- 音声電気信号出力あるいは音響出力（device を制御する）
- 音節程度の単位で波形生成に必要な情報（韻律、速度、etc.）を生成出力する
- （感情タグつき）テキストから感情表現情報を生成する
- binary 出力も可能 → Barge-in 対策のための Echo Cancellor の計算のため

などが望ましい。

4.6 顔画像合成・制御に関する問題

顔・身体動画像の生成について将来の問題あるいは要望を語ればきりがない。音声合成と精密に同期した発話顔合成、任意の表情を顔合成、音声合成に付与する技術、うなずき・瞬きなどの non-verbal な情報による顔合成制御は基本として、

- 発音に対応する音声器官の動作 (→ 唇、顎、歯、舌、のどぼとけ)
- 発話内容と同期して指定された、ジェスチャ、指示動作、身振り、身体動作、強調動作
- 話者切替え
- 人物動画像・身体画像生成、身体のゆれ
- 各項目について自律動作（default 動作）と指令動作を実現
- 心理表現（そわそわ、落ち着き)、感情表現、表情、身体のゆれ
- まばたき、視線制御（直視する、目を伏せる）
- プリセット動作（反射的に見せる表情群）と、それらへの滑らかな動き補間→ 驚き、「えっ?」、「はてな？」、しらんぷり
- 特に指令を与えずとも適宜動作するような自律的な動作 (default 動作) の制御

など、さらに

- 指示語と対応した画像生成
- 眼の周囲の細部の合成 → 表情の詳細制御
- 眼鏡つけた顔のアニメーション
- 顔と髪の制御
- 歯、舌の表示
- 複数人物
- 食べる、走る、道具を操作する

なども含めて、要望は限りがない。これは同時に、本分野の研究の大きな可能性を示唆している。

5 結語

将来の音声認識研究の方向の一つの可能性を開くことを目的の一つとして擬人化音声対話エージェントのオープンプラットフォームの構築を目指して活動を開始した。本稿ではその狙いや意義、進め方、今後に予想される問題点などを述べた。本プロジェクトが成功裡に進み、広く使われるフリーソフトウェアとして結実することを願う。