ヒューマンインタフェース　97−11
音声言語情報処理　40−11
（２００２．２．１）

擬人化音声対話エージェントツールキットの基本設計

川本真一 *1 下平博 *1 新田恒雄 *3 西本卓也 *4 中村哲 *5 伊藤克亘 *6 森島繁生 *7
四倉達夫 *7 甲斐充彦 *8 李晃伸 *9 山下洋一 *10 小林隆夫 *11 徳田恵一 *12
広瀬啓吉 *2 峯松信明 *2 山田篤 *13 伝康晴 *14 宇津呂武仁 *3 嵯峨山茂樹 *1,2

*1 北陸先端大, *2 東大, *3 豊橋技科大, *4 京工繊大, *5 ATR, *6 産総研, *7 成蹊大, *8 静岡大, *9 奈良先端大, *10 立命館大, *11 東工大, *12 名工大, *13 ASTEM, *14 千葉大

あらまし　筆者らは，顔画像が容易に交換可能で，音声合成が話者適応可能で，対話制御の記述変更が容易で，更にこれらの機能モジュール自体を別のモジュールに差し替えることが容易であり，かつ処理ハードウェアの個数に柔軟に対処できるなどの特徴を持つ擬人化音声対話エージェントシステムを構想し，実装した．各モジュールのインタフェースを統一化して扱い，モジュール間の入出力は，*UNIX* システムで使われている標準入出力を用いる簡便な方法にてモジュール統合機構を実現した．いくつかの簡単な対話タスクについてエージェントを試作し，必要な機能に関する達成度を確認した．また，顔画像合成モジュールを制御する新たなモジュールの追加を容易に実現することができた．

キーワード　擬人化エージェント，音声対話システム，ソフトウェアツールキット

A Design of Anthropomorphic Spoken Dialog Agent Toolkit

Shin-ichi Kawamoto*1 Hiroshi Shimodaira*1 Tsuneo Nitta*3 Takuya Nishimoto*4 Satoshi Nakamura*5
Katsunobu Itou*6 Shigeo Morishima*7 Tatsuo Yotsukura*7 Atsuhiko Kai*8 Akinobu Lee*9
Yoichi Yamashita*10 Takao Kobayashi*11 Keiichi Tokuda*12 Keikichi Hirose*2 Nobuaki Minematsu*2
Atsushi Yamada*13 Yasuharu Den*14 Takehito Utsuro*3 Shigeki Sagayama*1,2

*1 JAIST, *2 Univ. Tokyo, *3 Toyohashi Univ. of Tech., *4 Kyoto Inst. of Tech., *5 ATR, *6 AIST, *7 Seikei Univ., *8 Shizuoka Univ., *9 NAIST, *10 Ritsumeikan Univ., *11 Tokyo Inst. of Tech., *12 Nagoya Inst. of Tech., *13 ASTEM, *14 Chiba Univ.

Abstract　*This paper discusses a design and architecture of a software toolkit to develop an anthropomorphic spoken dialog agent (ASDA) that is easy to customize. Such human-like voice dialogue agent is one of the promising man-machine interface for next generations. To develop such a software toolkit, this paper firstly discusses the basic requirements that ASDA system should have, and then designs the software modules of the systems to fulfill the requirements. A prototype agent system has been developed on the UNIX-base systems by using the software toolkit that is under development. Discussions of the current achievement of the toolkit that will become publicly available as a free software are given finally.*

Keyword　*anthropomorphic agent, spoken dialog system, software toolkit*

1. はじめに

今後のヒューマンインタフェース (HI) 技術において，機械があたかも一個の人間のように振舞い，人間の顔や姿を表現し，音声言語で話し聞くような擬人化音声対話エージェントは，大きな目標の一つであり，様々な研究開発が進められている [2, 3, 4] が，人間同士の対話に比べるとまだ初歩的である．この関連分野は極めて広く，心理学，自然言語処理，知識処理など多分野の多面的な協力が必要になるだろう．それらの研究促進のためにも，また成果の集積のためにも，多くの研究開発者が容易に使用・開発参加できるような擬人化音声対話エージェントのツールキットが，共通の研究プラットフォームとしてソースコードを含めて無償公開・提供されることが望ましい．

人間と機械のより自然な対話を目指した研究として，ロボットの自然な対話を目指した興味深い研究 [1] や，市販のソフトウェアを活用した擬人化エージェントの例もあるが [2]，後述するような高度な機能と高いカスタマイズ性を持たせたソースコード

無償公開のソフトウェアはまだ提供されていない．

筆者らは，擬人化音声対話エージェントを用いた研究開発のための共通のプラットフォームとなる，カスタマイズ可能なソフトウェアツールキットの開発を進めており，現在その基本動作を確認した段階にある．本稿では，多様なエージェント構築のためのツールキットとして必要とされる機能，それを実現するための問題点と解決法について議論する．更に著者らが実現したシステムについて述べ，目的の達成について評価し，今後の技術的課題を議論する．

2. 擬人化音声対話エージェントへの要求条件

コンピュータがあたかも人間のように振舞い，人間の顔や姿を表現し，ユーザの音声言語で話し聞くようなインタフェース実現のための擬人化音声対話エージェントに求められる要素について議論する．

2.1. 人間らしい対話実現のための要素技術

人間同士の音声対話では，キーボード対話などではあまり見られない特有の現象，例えば，相手の話の途中で相槌を打つ，最後まで聞かずに割り込むなどが見られる．また韻律によって，感情や意図を表現することが多い．これらが実現できる基本機能を人間・機械間においても実現すれば，音声対話の特性を生かした効率的なコミュニケーション実現のプラットフォームとして意味がある．

更に自然で効率的な音声対話を実現するために，ユーザに対する素早いレスポンスや各要素技術の高い次元でのバランスも重要である．このような対話における時間要素は，従来の音声認識・合成技術においては，大きな関心の対象ではなかった．

2.2. カスタマイズが容易な構成

研究開発の共通基盤の一つとしてツールキットを提供する以上，その利用者の広い要求に答えられることが望ましい．そのうちの重要な一つの要素は，エージェントの顔，音声，および対話タスクのカスタマイズ可能性である．これらは適用分野，利用目的などにより望み通りに容易に変えられることが望ましい．例えば，エージェントの機能毎に異なる顔や声などのインタフェースを設定する際，利用者の好みに応じて容易にカスタマイズできることが望ましい．また，擬人化音声対話エージェントを用いて音声認識・合成の研究成果の効果の実証・検討や，研究・開発した技術のデモンストレーションを行う際，適した対話タスク記述やデモンストレーションを作成する労力は極力省きたい．

2.3. 機能部品のモジュラリティ

ツールキットの産業的応用において，標準で準備されている音声認識・合成，顔画像合成のモジュールに替えて，一部に別のモジュールを使用したい場合や新機能のモジュールを追加したい場合があり得る．

また，音声認識・合成，顔画像合成などの要素技術研究における参照システムとしての利用を考えるとき，それぞれのシステムは単独で利用可能であることで効率的な要素技術の参照が行える．また，各要素技術の開発においても，単独利用できることは，開発効率を向上させることができるだろう．

更に，エージェントの運用に関して，各モジュールを分散して処理することでより高速，かつ効率的に動作する可能性があり，そのために，機能単位のモジュール化とそれらを統合する枠組が必要であろう．

2.4. ソース無償公開のソフトウェア

新しい HI 技術分野の発展の基盤として，擬人化音声対話エージェントのソフトウェアがソースコードも含めて，すべてが無償で公開されることは，極めて有意義であろう．この技術はまだまだ発展途上であるため，多くの研究開発者による改良を歓迎するものでなければならない．つまり，多くの分野の研究者が擬人化音声対話エージェントを使った研究に参入しやすい環境の整備が必要であり，理解と利用が容易な共通の研究開発プラットフォームの提供が重要である．更に研究開発において，種々の応用に無償で使用できる意義も大きい．

2.5. 従来のソフトウェアの要求達成度

それぞれの要素技術が研究段階のものであり，多くの難しい問題を含んでいるため，現状では上記の要求を満たすようなカスタマイズ性に優れたフリーソフトは無い．例えば，漢字仮名混じり文を明瞭に読み上げを実現し，ソースコードまで無償公開された音声合成ソフトウェアは存在しない．

3. 構成要素モジュールの設計

擬人化音声対話エージェントのプラットフォームとして，先に議論した要件を満たすためには少なくとも，擬人化音声対話エージェントの顔や声などをカスタマイズ可能にするための対話部品モジュール (音声認識，音声合成，顔画像合成) と，対話部品モジュールを統合し，対話を管理・カスタマイズする対話統合モジュールが必要と考える (図 1)．ここでは，各モジュールに必要な機能について議論する．

3.1. 音声認識モジュール

筆者らは，対話音声にも利用可能なモジュールとして，大語彙連続音声認識エンジン Julius[9] や SPOJUS[15] を開発してきた．

Julius では，言語モデルとして統計的言語モデルを用いていたが，対話記述や簡単な対話アプリケーションの作成・変更を容易にするためには，形式言語による言語モデル記述，および状況に応じた動的

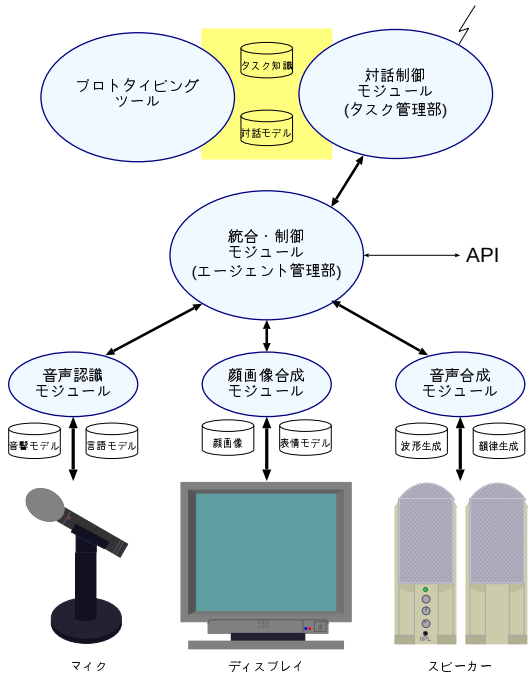

図 1: 擬人化音声対話エージェントプラットフォーム

な切替が重要である．更に，インタラクティブな対話を実現するためには，まずはユーザ発話に対する素早いレスポンスの実現が必要である．

この対処として，Julius を基盤に有限状態文法に基づく連続音声認識パーザ Julian[10](図 2) を開発し，文法切替や漸次的認識結果出力を実現する．

3.2. 音声合成モジュール

音声合成モジュールに求められる機能は，特定の人の声による任意の漢字仮名混じり文の明瞭な読み上げと，声質・韻律のカスタマイズ可能性である．

この要件を満たすために，**1)** 事例ベースシステム [12] より少ない学習データ量で実現可能であり，合成音声の品質はフォルマント合成系システムより高く，明瞭度も非常に高い．**2)** 韻律などのさまざまな制御が比較的容易である．**3)** モデルベースの声質変換，話者適応も可能であるというメリットを有する HMM 音声合成方式 [11](図 3) を採用する．

更に対話の自然性を確保するためにエージェントの音声発話時において，精密な合成音声と合成画像中の口の運動の同期 (以後 LipSync と呼ぶ) を実現するための情報共有，および他のモジュールとの連係を実現する．

また，発話する漢字仮名混じり文の対する部分的な強調や韻律制御などの機能を付加する．発話文に対する部分的な強調や韻律制御などの指定は，日本電子工業振興協会の作業グループによって制定された「日本語テキスト音声合成用記号の規格: JEIDA-62」[14] を基盤に拡張を検討する．

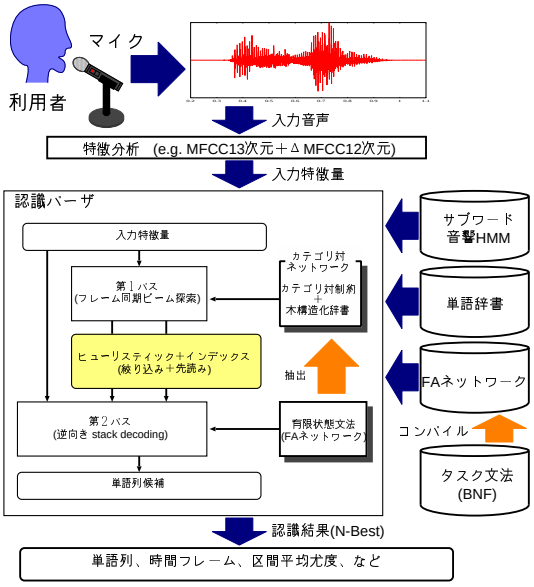

図 2: 音声認識モジュール

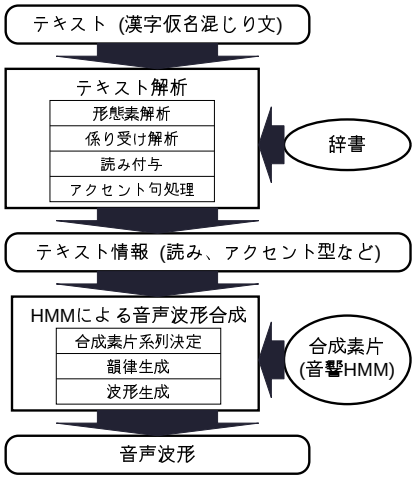

図 3: 音声合成モジュール

3.3. 顔画像合成モジュール

エージェントに任意の顔を持たせるためには，顔画像のカスタマイズが容易であることが重要である．コンピュータグラフィクス (CG) によるアニメーションの顔を用いる場合，相槌などの頭の動きや精密な LipSync を実現できる複数の顔を，各ユーザ毎に準備するのは困難である．

一方，筆者らが開発したソフトウェア [13] では，顔画像と標準 3 次元頭部モデルを整合させ，各個人のモデルを生成できるため，顔画像を準備するだけでエージェントの顔をカスタマイズできる．

更に，より人間らしい対話を実現するために，精密な LipSync のための他のモジュールとの連係，喜びや怒りを表現するための任意の表情付加機能，自然な瞬きの制御機能を付加する (図 4)．

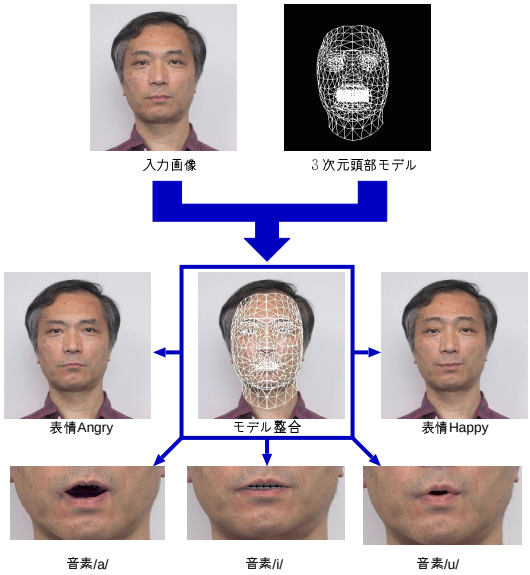

図 4: 顔画像合成モジュール

3.4. 統合モジュールおよびツール群

擬人化音声対話エージェントを構築するためには，それぞれの要素技術を統合する枠組として，以下に述べる対話部品モジュールを統合・制御し，対話を管理するための 3 つのモジュールを構築する．

3.4.1. エージェントマネージャ

要素技術のモジュラリティとカスタマイズ容易性を確保するため，各モジュールの通信を管理し，各モジュールの低レベルの制御を実装する必要がある．

このためエージェントマネージャ(AM) では **1)** 各モジュールのインタフェースを共通化することで，新たなモジュールの追加や機能拡張を容易にし，**2)** AM を介してモジュール間の入出力を扱うことで，モジュール間の入出力の扱いを簡単化するとともに，モジュールの独立性を高め，**3)** エージェントの音声発話における LipSync の機能など頻繁に使う機能は，マクロ処理コマンドとして提供する．

各モジュールが連動して円滑に動作するには，分散環境におけるシステム制御，情報管理などが必要となる．これらの開発例として，MIT で開発された Galaxy-II[5] を基礎とした DARPA のコミュニケータ・プログラム [6] におけるシステムや，SRI の開発した Open Agent Architecture(OAA)[7] などがある．このような標準化の動向は参考にすべき点が多くあるが，汎用的なシステム構成を目指すことによって，機能が複雑化・多様化することが多く，このことがシステムの利用を困難にする恐れがある．これらの標準化の動向に注目しつつ，最小限の変更でモジュール交換や機能追加などのカスタマイズを実現可能にするモジュール統合の設計でなければならない．

図 5: 擬人化音声対話エージェントとの対話風景

3.4.2. タスクマネージャ

タスクマネージャ(TM) に求められる機能は，音声対話の記述の標準化に沿って，対話記述が容易であり，音声対話のために書かれたタスク記述の再利用や部分利用を可能にすることである．更に，擬人化音声対話エージェント特有の機能を活用できるような対話記述の拡張性が求められる．

この要件を満たすために，開発の基盤として，XML を基盤とした対話タスクの記述・カスタマイズを容易にするための言語として標準化が進められている VoiceXML[8] を採用し，VoiceXML インタプリタ機能を TM に持たせる．

3.4.3. プロトタイピングツール

様々な分野の研究者が擬人化音声対話エージェントを使った研究に参入しやすい環境を実現するために，各モジュールの詳細の動作原理を知らなくても対話タスクの記述を含めた擬人化音声対話エージェントの開発・カスタマイズを容易に行うための統合環境が必要となる．

この統合環境として，対話システムの構築に必要なパラメータ設定やシナリオ記述・制御などを容易に行うことが出来る GUI 環境を提供する．

4. 擬人化音声対話エージェントシステムの実現

前章で述べた構成要素モジュールの設計に従い，モジュールに要求される機能を実現し，対話システムを試作した．システムとの対話風景を図 5 に示す．

このシステムは，対話部品モジュールおよび AM の基本機能の動作確認のためにいくつかの対話タスクについて実現した．簡単なタスクであれば，AM の提供するコマンドを使って直接タスクを記述することができる．実際に TM として，いくつかのタスクについて AM のコマンドを直接利用するプログラムで実現できることを確認した．

VoiceXML インタプリタによる TM，およびプロトタイピングツールについては，それぞれ単体での

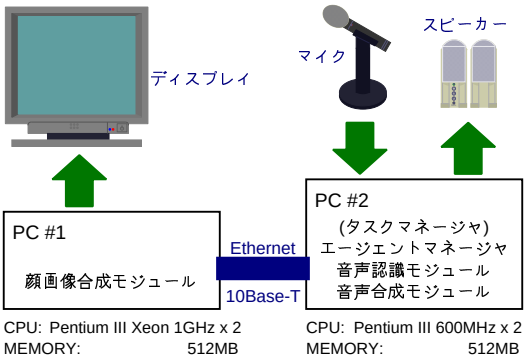

図 6: 擬人化音声対話エージェントの動作環境

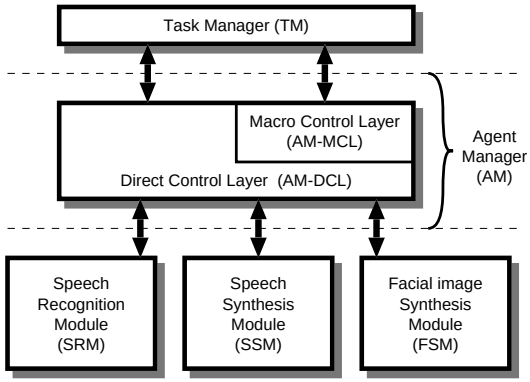

図 7: エージェントマネージャと各モジュールとの基本構成図

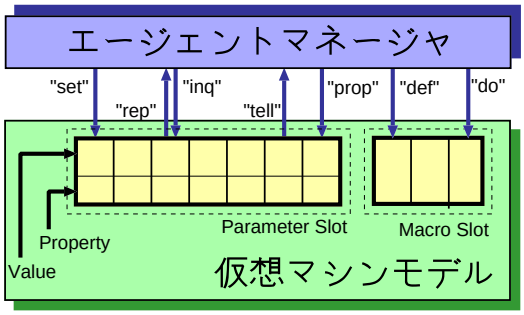

図 8: エージェントマネージャと仮想マシンモデルとの関係図

開発を行ない，部分的な機能を実現した．

図 6 に試作システムのハードウェア構成を示す．

4.1. 音声認識モジュール

有限状態文法に基づく認識，語彙・文法などの切替え機能，漸次的な認識結果の出力を実現した．更に，文法記述のための簡易的なツールキットを構築することで，対話タスク毎の文法生成を容易に行なうことができた．また速度面に関しては，パープレキシティ 10 程度の比較的簡単なタスクであれば，ほぼリアルタイムでの動作を実現した．

4.2. 音声合成モジュール

男女各 1 名分について，任意の漢字仮名混じり文章に対する明瞭な読み上げと部分的な韻律制御機能を実現した．これにより，疑問文発話の際の文末表現などを実現できた．また，合成音声の発話時間より短い時間での合成準備を実現した．更に，顔画像合成モジュールと連係することで，エージェントの音声発話時の LipSync が可能となった．

4.3. 顔画像合成モジュール

正面方向から撮影した 1 枚の顔画像と標準ワイヤフレームモデルを整合させることで，エージェントの 3 次元頭部モデルが生成し，特定の人物の顔による任意の表情の合成を実現した．更に，口腔環境内の歯のモデルを加え，より精密な口周辺の合成画像が得られた．また，顔画像合成モジュールの動作速度は，ハードウェアアクセラレーション機能を使い，平均描画更新レート 20[frame/sec] を達成した．

4.4. エージェントマネージャ

AM は 2 つの機能レイヤーで構成し (図 7)，各モジュールを図 8 に示すような仮想マシンモデルとして扱い，インタフェースを統一化した．実際，顔画像の自律動作を頭部回転コマンドを直接用いて制御する新たなモジュールの追加を容易に実現できた．また各モジュールを並列動作させ，効率的な処理を実現した．

音声対話において頻繁に利用される LipSync は，AM が提供するマクロコマンドとして実現した．これにより，対話部品モジュールは AM との通信のみを考慮して設計できるため，モジュールの独立性が確保でき，TM から見た利便性を向上させた．

モジュール間の入出力は，UNIX システムで使われている標準入出力を用いる簡便な方法にて実装することで，各モジュールの単独開発・単体での利用が容易となった．

5. 考察

試作システムとの対話を通じて見られた開発目標の達成度と，機能拡張の必要性について考察する．

5.1. カスタマイズ性

音声認識に関しては，文法の切替えが実現し，文法記述のための簡単なツールキットを整備することで，語彙・文法のカスタマイズが可能になった．

音声合成に関しては，任意の漢字仮名混じり文章の読み上げと韻律制御が実現し，音声発話のカスタマイズが可能になった．更に話者適応を実装されれば，声のカスタマイズも可能になると考える．

顔画像合成に関しては，任意の人の顔画像からエージェントの 3 次元頭部モデルを生成し，このモデルを利用して瞬きや笑顔などの表情の生成する顔・表情のカスタマイズ機能を実現した．

5.2. 機能部品のモジュラリティ

AM は，各モジュールのインタフェースを統一化し，モジュールの追加などのカスタマイズが容易な設計を実現した．更に，モジュール間の入出力は全て AM を介して行った結果，高いモジュールの独立性を実現できた．

今後は，Julian に代わる音声認識エンジンとして SPOJUS[15] の追加作業などをすることによって，さらなるモジュラリティの検証ができると考える．

5.3. 人間らしい対話の実現

音声合成と顔画像合成が連係し，エージェントの音声発話時の LipSync はほぼ問題なく動作した．更に歯のモデルを加えることによって，より精密な口周辺の合成を実現し，音声発話の自然性が向上した．

よりインタラクティブな対話を実現するために，筆者らは，音声認識時の第 1 パスの結果の漸次的な出力を実現した．また各モジュールは並行に動作しており，エージェントが音声発話中でも音声認識可能である．これらの情報をうまく利用することで，例えばエージェントの対話理解状態の開示のための相槌 [16] や割り込みなど，より効率的な対話を実現できると考える．

5.4. 今後の展望

エージェントが話を聞いている状態や発話が中断された状態，発話中に関しても強調などの動作が LipSync や表情だけではなく，頭部動作も連係した振舞として合成できれば，より明示的なシステムの状態開示を行なうことができ，人間らしい擬人化エージェントの実現に役立つと考える．

またエージェントの発話に関して，発話内容の部分的な強調が必要な場合，それに対応する表情や頭部動作のタイミングを制御するすることで，より人間らしい発話に近づけることができると考える．

今後は，各モジュールのさらなる機能拡張・改良を進めるとともに，標準的な分散オブジェクト環境アーキテクチャCORBA[17] 等の導入も検討しながら，開発を進める．

6. おわりに

本稿では，擬人化音声対話エージェントを将来の HI の重要な技術要素として位置づけ，その分野の共通プラットフォームとなり得る高いカスタマイズ可能性を備えたツールキットの実現を目指し，それに必要な要素とその実現技術について論じた．

本稿で示した擬人化音声対話エージェントのツールキットは，2000 年 3 月より十数ヶ所の研究機関の共同で開発が進められており [18]，ソースコードも含めて無償公開する予定である．本システムは，擬人的な外見と，音声対話する機能を持つものであり，これに知能的な情報処理部を接続することにより，広い用途が予想される．利用者を広く募り，そのフィードバックによりシステムを改善し，更に利用者を広げて行きたい．本システムに関心を持たれた各方面の研究者からの御意見を歓迎する．

謝辞

本研究の一部は，情報処理振興事業協会 (IPA)「独創的情報技術育成事業」の支援を受けた．

参考文献

[1] 松坂, 東條, 久保田, 田宮, 古川, 早田, 中野, 小林: “複数話者による対話システム,” Interaction’99, pp.33-34, 1999.
[2] 土肥, 石塚: “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, Feb. 1999.
[3] 向井, 関, 中沢, 綿貫, 三吉: “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,” Interaction2001, pp.139-140, 2001.
[4] Joakim Gustafson, Nikolaj Lindberg and Magnus Lundeberg: “The August Spoken Dialogue System,” Proc. of Eurospeech99, pp.1151-1154, 1999.
[5] Stephenie Seneff, Ed Hurley, Raymond Lau, Christine Pao, Philipp Schmid and Victor Zue: “GALAXY-II: A Referece Architecture for Conversational System Development,” In ICSLP-1998, pp.931–934, 1998.
[6] DARPA Communicator Program, 1998. http://fofoca.mitre.org/.
[7] OAA (The Open Agent Architecture). http://www.ai.sri.com/~oaa/.
[8] VoiceXML Version 1.00: VoiceXML Forum, 2000. http://www.voicexml.org/.
[9] 河原, 李, 小林, 武田, 峯松, 伊藤, 山本, 山田, 宇津呂, 鹿野: “日本語ディクテーション基本ソフトウェア (98 年度版) の性能評価,” 情報処理学会研究報告, 99-SLP-26-6, May 1999.
[10] 李, 河原, 堂下: “文法カテゴリ対制約を用いた A* 探索に基づく大語彙連続音声認識パーザ,” 情報処理学会論文誌, Vol.40, No.4, pp.1374-1382, Apr. 1999.
[11] 吉村, 徳田, 益子, 小林, 北村: “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 電子情報通信学会論文誌, Vol.J83-D-II, No.11, pp.2099-2107, Nov 2000.
[12] ニック・キャンベル, アラン・ブラック: “CHATR: 自然音声波形接続型任意音声合成システム,” 電子情報通信学会技術報告, SP96-7, Mar. 1996.
[13] 森島, 八木, 金子, 原島, 谷内田, 原: “顔の認識・合成のための標準ソフトウェアの開発,” 電子情報通信学会技術報告, PRMU97-282, Mar. 1998.
[14] (社) 日本電子工業振興協会: “日本語テキスト音声合成用記号の規格,” JEIDA-62-2000, 2000.
[15] 甲斐, 中川: “冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価,” 電子情報通信学会論文誌, Vol.J80-D-II, No.10, pp.2615-2625, 1997.
[16] 平沢, 川端: “音声対話システム Noddy – ユーザ発話途中でのうなずき・相槌生成 –,” 情報処理学会研究報告, SLP20-9, 1998.
[17] CORBA (The Common Object Request Broker Architecture). http://www.corba.org/.
[18] 嵯峨山, 中村: “擬人化音声対話エージェント開発とその意義,” 情報処理学会研究報告 2000-SLP-33-1, Oct. 2000.