ヒューマンインタフェース 69－7
（1996. 11. 15）

人に近いインタフェースを目指して
～擬人化インタフェースRachelの試作(1)～

鈴木薫、山口修、福井和広、田中英治、倉立尚明、松田夏子
東芝 関西研究所

非エキスパートユーザのためのコマンドプロセッサとしてマルチモーダル擬人化ユーザインタフェース(multi-modal Humanoid User Interface = mmHUI)を提案し、その検証の第1段階として試作したナレーションエージェント(Rachel)を紹介する。Rachelは人物を画像的に検出すると挨拶をして自己紹介を始める仮想人間(CG)である。顔を持つ計算機への反応を探るべく社内公開して100名を超える人に体験してもらった結果、多くの関心と興味を引くことができたが、その応答遅れのために自然性への評価は高くなく、技術的な課題の多さを確認した。

A multi-modal Humanoid User Interface (mmHUI)
- An Experimental Narration Agent (Rachel) -

Kaoru SUZUKI, Osamu YAMAGUCHI, Kazuhiro FUKUI,
Eiji TANAKA, Takaaki KURATATE, and Natsuko MATSUDA

Toshiba Kansai Research Laboratories
8-6-26,Motoyama-minami-machi, Higashinada-ku, Kobe 658, Japan

We are focusing on a multi-modal Humanoid User Interface (mmHUI) as a new command processor of social systems for none-expert users. We developed our first experimental system named "Rachel", a virtual narration agent. After detecting the user by processing images, Rachel greets the user and introduces herself, complete with facial expressions. We have demonstrated Rachel to more than 100 people, and found that many people were interested in a computer which has her own face and voice. However, many technical gaps still remain in building a perfect human-like agent.

1　はじめに

現在のパソコンではGUI(Graphical User Interface)が計算機との対話手段の主流である。そのようなパソコンの多くは主としてオフィスで特定業務（文書作成、表計算、メール、データベース）に使われており、GUIはこれらの目的のために安価でそこそこ使いやすい環境を提供している。しかしながら、パソコンが今まで以上に広く使われる場面を想定した場合、GUIが適当な対話手段であるという保証はなく、むしろGUIを超える対話手段が必要である可能性が高い。

人の対話能力はもともと人間同士の対話のために発達したものである。従って、人が最も自然に対話できる相手はやはり人であるといえる。機械を人に見立てることには賛否両論あろうが、パソコンが人のように対話でき、言うことをきいてくれる相手であれば、特別な訓練を加えることなく誰もがパソコンを扱うことができるようになるだろう。この認識に立って筆者らは「マルチモーダルな擬人化ユーザインタフェース＝mmHUI (multi-modal Humanoid User Interface)」を指向している。

人間の感覚器と表現器は複数のチャネルを持つ、いわゆるマルチモーダルになっていることは周知のとおりである。筆者らの考えるmmHUIは、マルチモーダルの考え方を拡張するとともに、計算機を仮想的な人間としてユーザに認識させることを意図したものである。それは、お店では店員、家庭ではお手伝いさん、オフィスでは秘書といったもののメタファとなる。

間瀬1)は車の運転を例に挙げ、「もっと右」とか「少し戻って」などというように言葉を使って機械に細々と命令を与えるのではなく、例えば「京都駅まで」と言えば「新幹線口ですね」というように気をきかせさえする、マクロなレベルで命令できるマルチモーダルのインタフェースエージェントが求められることを指摘して、対話の様式が人間に近いメタファ（擬人化メタファ）の有用性を説明している。

また、長谷川ら2)は、このような擬人化メタファがユーザを認証すること、すなわちユーザがシステムに個人として認識されることの有用性を実証するシステムを提案している。

以下、本報告では提案するmmHUIの背景と考え方、mmHUIを検証するために最初に試作したナレーションエージェント実験システムの紹介を行う。

2　multi-modal Humanoid User Interface

2．1　ポストGUIの必要性

GUIの登場以前はキーボードからのコマンド入力がパソコンの唯一の UI (User Interface) であり、コマンドにより直接的に機能を呼び出せる反面、呪文のようなコマンドを知らなければパソコンを操作できない、エキスパートユーザのための操作体系であった。その後、マウスというポインティングデバイスとアイコンという視覚的かつ直観的な画面要素から構成され、画面上で一覧して選択するという極めてシンプルな操作体系を用いたGUIの登場により、多くの作業を呪文を覚えることなく実行できるようになり、より広いユーザ層（例えば子供たち）がパソコンを使えるようになった。

しかし、このようなパソコンがネットワークで外部と繋がり、社会システムの端末として利用される近未来を考えると事情が若干異なってくる。まず、システムが提供するサービスや情報はシステム側で勝手にインストールあるいはアップデートされるので、ユーザはその全貌を把握できなくなる。また、その数も種類も格段に増加することが予想される。膨大なサービスや情報をGUIの枠組みで限られた表示領域にアイコンで示すのにも限界が現われる。そこでこれらを体系付けて階層化する必要がでてくる。事実、現在のウィンドウシステムは膨大な数のファイル（アプリケーションプログラムやデータ）の表示と操作に際してこのような階層化を行っており、また、WWWでもリンクを使って情報の階層化を行っている。この結果、ユーザは難解な呪文から解放された一方で、この階層的なメニューを探索せ

ねばならないという新たな宿命を背負うことになる。

この探索の繁雑さを回避するためには、目的のサービスや情報を的確に呼び出せる強力なサーチエンジンが必要である。実際、WWW上では幾つかのサーチエンジンが稼働しているが、検索の的確さには問題が残っており、これはこれで新たにサーチエンジンを使いこなすエキスパート性がユーザに求められている。

今後、発展すると見込まれる社会システムのように、ユーザがシステムに対する知識も習熟度もまちまちな不特定多数の非エキスパートであることが前提となるシステムでは、そこで求められるUIは以下の要件を満足するコマンドプロセッサであると筆者らは考える。

- ●非エキスパートユーザが訓練なしに十分使いこなせる単純な操作体系
- ●難解な呪文によらず、階層メニューによらず、的確に目的を果たせるコマンドシステム

2．2　Multi-Modal User Interface

GUI以前のコマンドライン方式にも長所があった。それは、コマンドさえ知っておれば即座に目的を果たすことができた点である。GUIが速度／操作性の点で今日のように快適な環境を提供できるまでに発達しても、コマンドライン入力を使うエキスパートユーザが多いのはそのためである。非エキスパートユーザにとって障害となっていたのは、コマンドが呪文のように難解であった点と、それをキーボードから正確に打ち込まなければならなかった点である。それならば、呪文を平易にして、多少不正確でもそれを受理してくれるようにすれば、コマンドラインの長所を活かしつつ、非エキスパートユーザにも障害とならないUIが構築できるはずである。これを実現する最も有効な手段は音声を主体にしたコマンド入力と、このコマンドとして人間の日常的な表現（自然言語、身振りなど）を受理することのできるマルチモーダルインタフェースである。人間は自分が対話する際に意識的／無意識的に使う情報のうち機械が扱えないものが多ければ多いほど、無理をして機械に近づいてやらなければならなくなる。逆に、機械が拾い上げられる情報の数や種類を可能な限り増やして、人と機械の対話を人間同士のそれに近づけてくれる強力なコマンドプロセッサが筆者らの求めるものである。

2．3　コマンドプロセッサmmHUI

GUIは直感性に優れ、視覚的にわかりやすく、視覚的要素ならびに空間的要素が機能にマッピングされたコマンドプロセッサである。その多くの特長の中で意外と重要なのが親しみやすさである。GUIを取り入れたウィンドウシステムはいわゆるデスクトップメタファを提供する。また、アイコンはフォルダや書類や文房具という机上に一般的に置かれるもののメタファである。世の中に何を象徴するのかわからないアイコンが氾濫しつつある現在でも、ではその中身を文字で説明したアイコンを使うかというと、相変わらずグラフィカルなアイコンが利用されるのはなぜだろうか。結論から言えば、GUIはそのグラフィカルな性格のために美しく楽しく価値があるのである。エージェントが顔を持つことも同様である。音声で対話できるのであれば、顔がないよりもあった方が良い。それも自分の気に入った顔が良い。顔があれば、自分を認識してくれて挨拶の1つもしてほしい。忙しいときには忙しい顔をしてくれても良い。ユーザは誰と話しているかを意識できるし、機械の状態を顔色で察知することもできる。気に入った顔が気に入った声としゃべり方で作業（生活）を優しくサポートしてくれれば、誰もが計算機を使おうという気にならないだろうか。筆者らのmmHUIはそんな「雰囲気の良い計算機」のコマンドプロセッサとして企画された。

以下、第3章ではmmHUIの有効性を検証するための実験システムについて説明する。mmHUIに関する最も大きな疑問は顔のあるインタフェースが利用者

に与える印象である。そこで、アプリケーションとして顔が必然となるナレーションエージェントを設定し、顔の形態上／動作上の品質、表現要素（口、目、表情など）、同期、速度がエージェントの自然性に与える影響の調査と、mmHUIの基礎技術の獲得を図ることとした。

システムは人物を検出すると、所定の挨拶を音声出力した後、本題のナレーションを始める。また、人物が離れるとこれを検出してナレーションを中断し、お別れの挨拶をする。そして、発話中には音声パワーの有無に合わせて口を自動開閉し、また人物を常に追跡して視線と顔向きを変えることも行う。なお、筆者らはこの実験システムをSF映画「ブレードランナー」に登場するアンドロイド（人間の振りをする機械）から名前をとってRachelと呼んでいる。

3　ナレーションエージェント (Rachel)

3．1　システムの構成

試作した実験システムは、図1に示すように人物同定モジュール、ナレーション制御モジュール、音声検出モジュール、エージェント映像生成モジュールから構成される。システムは2つのWSを用いており、画像処理のみを行うWSとエージェント映像の生成と音声処理を行うWSから成る。各モジュールは、それぞれのWSにプロセスとして実装されており、各モジュールはUNIXソケットのプロセス間通信によって情報を交換しあう。以下に個々のモジュールの働きについて説明する。

3．2　人物同定モジュール

人物同定モジュールは、（システム）の前に人物が存在するか否かを、TVカメラによって取得した画像を処理し判断する。カメラはエージェントの表示されるディスプレイの前に図2のように設置される。よって、エージェントと向き合う方向にシステムの利用者が存在することを判断するには、画像中から正面顔を検出すればよい。

カメラによって入力された画像から顔領域を抽出する方法について述べる。検出は部分空間法により行う。

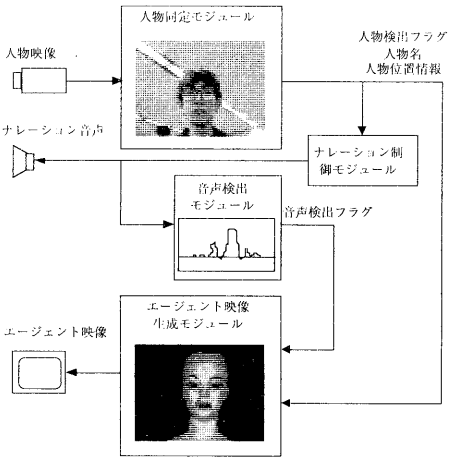

図1　システム構成

検出に用いる辞書は不特定多数の人物に対応するため、図3中央に示すように顔画像を眉から口の上唇までの矩形で切り出し、髪型などの影響を受けないようにした。人物検出用には８５０枚の画像を用いて顔辞書を生成し、個人認証用には各々約３０枚の同一人物画像を用いて個人辞書を生成した。

入力画像（図3左）から任意の大きさと位置で切り出された部分画像と顔辞書画像との類似度を求め、最も類似度の高い場所に顔領域が存在すると仮定する。その類似度が設定した閾値よりも高い場合、その領域を人物の顔領域とする（図3左）。

顔領域が検出されると、次にこの領域を個人辞書と照合して最も類似度の高い人物を特定する。このときの類似度がある閾値より高い場合にはその人物が検出されたとし、低い場合には未知人物が検出されたとする。

顔検出／照合処理は一秒間に５～６枚の速さで実行され、顔領域の検出が所定回数連続して成功した場合に人物が存在すると判断する。逆に、顔検出を所定回数連続して失敗すると人物が立ち去ったとす

る。

本システムで実現されるエージェントは、人物を検出するとその方向に顔を向けて挨拶と自己紹介を開始する。その際に、個人照合できた人物にはその名前を呼んで話し始める。また、人物が離れるとエージェントは話しをやめて待ち受け状態になる。これらの機能を実現するために、人物同定モジュールは人物の存在／非存在を示す人物検出フラグ、検出された人物の名前、画像中の顔領域の重心を示す人物位置情報（図３右）をナレーション制御モジュールとエージェント映像生成モジュールに出力する。

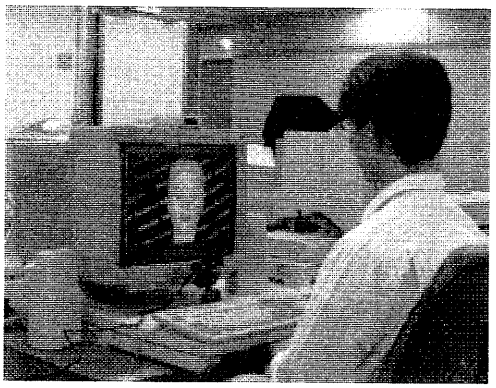

図２　実験風景

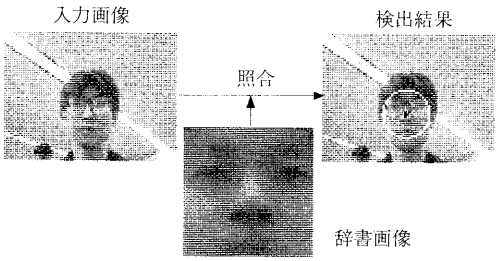

図３　人物検出・同定処理

３．３　ナレーション制御モジュール

人物検出フラグと人物名を受け取ったナレーション制御モジュールは、その人物に向けた挨拶音声を出力し、その後に続けてナレーション（実験では自己紹介文）音声を出力する。もし、検出された人物が未知人物である場合には一般的な挨拶を行う。また、人物検出フラグがオフになった場合には人物が離れたと解釈して、ナレーション音声を中断してお別れの挨拶を出力した後、待ち受け状態になる。

３．４　音声検出モジュール

音声検出モジュールの入力ラインはナレーション制御モジュールの音声出力ラインに接続されている。音声が出力されると、音声検出モジュールはこれを検出し、有音区間／無音区間を区別できる音声検出フラグをエージェント映像生成モジュールに受け渡す。この音声は合成音声でも録音音声でもかまわないが、今回のシステムでは実在人物の声を録音して生成したAIFFファイルを再生する。

ここで、音声検出モジュールの動作を説明する。一般に見られる音声の波形の概形を図４に示す。図中の番号は状態遷移図５で示される各状態に対応している。図４の縦軸は入力音声のパワー、横軸は時間である。入力音声のパワー（以下Pw）は次式で定義される。

$$Pw = 10 \times \log10\ (Am \times Am / block)$$

ただし、Amは入力音声の入力値、
blockは移動平均を取るサンプル数、
Pwの単位は音圧レベルを示すdBとする。

最初、無音状態(0)であったのが、入力音声のパワーがピーク始端検出閾値を超えると音声仮定状態(1)となり、ピーク終端検出閾値を下回った時点で終了仮定状態(4)に遷移する。これは、一般に音声波形が山と谷を持って連続しているという特性を用いている。この時の閾値は背景雑音パワー（無音状態(0)の期間におけるパワーの平均値）からの相対的なパワーとして決定するが、音声の場合、一般的にピークの後に雑音が多いという特性から、パワー上昇時のピーク始端検出閾値よりもパワー下降時のピーク終端検出閾値の方を高く設定しておく。この結果、雑音の影響を極力排除して音声部分のみを抽出しやすくなる。

音声では２番目以降のピークでパワーが大きくな

る傾向がある。逆に音声以外の雑音では最初のピークしか存在しないか、２番目以降のピークが続いても大きなパワーを示さない傾向が見られる。最初のピークを超えて終了仮定状態(4)から再び入力音声のパワーがピーク始端検出閾値を超えると、次音声仮定状態(2)となる。次いで、この状態が一定時間継続することにより次音声状態(3)となる。音声検出プロセスはこの時点で始めて音声が入ったと認識する。次音声仮定状態(2)あるいは次音声状態(3)から、ピーク終端検出閾値を下回ると終了仮定状態(4)となる。

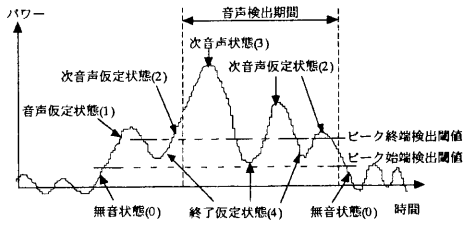

図４　音声波形拡大図と各状態の関係

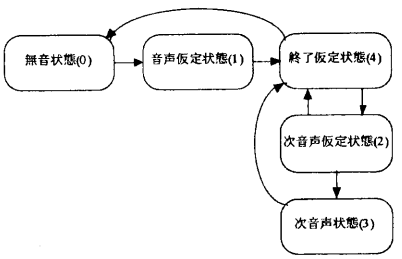

図５　波形分析時の状態遷移

また、さらに雑音への反応を鈍らせるために、これとは別に最初のピークの大きさにも閾値を設け、これを超えないピークを検出しても無音状態(0)から音声仮定状態(1)への遷移を行わないようにする。さらに、十分な強さを持つピークでも設定値以下の継続時間のものはスパイク状の雑音として無視する。

音声検出後に入力レベルがピーク始端検出閾値を下回った場合、終了仮定状態(4)から音声の終了、すなわち無音状態(0)へ遷移する。

３．５　エージェント映像生成モジュール

人物検出フラグがオンになると、エージェント映像生成モジュールはエージェントの顔向きを人物検出位置に向けて発話状態に入る。発話状態中のエージェント映像生成モジュールは音声検出フラグに従って有音区間で口をランダムに開閉させる。また、発話中のエージェントの頭部は上下に小刻みに揺れるように制御される。これにより、ナレーション音声の中に有音／無音区間がどのような分布で混在していても、エージェントの口は的確に開閉し、かつ発話中の頭部の動きを表現できるようになる。

エージェント映像は約7000ポリゴンのデータ（図４）からCGにより生成される（図５）。データは頭髪、額、顔面、唇、眉、睫、まぶた、顎－側頭部－後頭部、首、胴、腕、眼球、歯の各パーツから成り、各々専用のマテリアル情報を与えられる。また、形状の見かけの良さを最大に保ちつつデータ量を最小に抑えるべく、各パーツは必要に応じて各々異なる精度でポリゴン化される。このデータがグーローシェーディングとZバッファ・アルゴリズムを用いてGWS (ONYX RE2) で約2～３frame/secの映像更新速度で描画される。

眼球は人物位置に応じて計算される方向に向けて制御され、顔向きはその50%の大きさで制御される。また、注視時に眼球を微妙に揺らすことで生きているような表現を実現する。瞳孔から虹彩にかけては実画像から切り出した画像にハイライトを加筆したテクスチャをマッピングする。また、この視線に応じてまぶたの開度が自動的に調整される。

顔向き制御では頭部まわりの全パーツが目標値の100%で座標変換を受ける。一方、首のねじれに関しては首から胴にかけてがその高さに応じて100%から0%まで傾斜された比率で座標変換を受けるように計算される（図６）。また、首の前後左右の傾斜に関しては首の付け根にピボットを設定し、これ

を回転中心として100%で座標変換されるが、首の付け根付近のみ100%から0%で比率を傾斜させることで、首の付け根まわりの肩などが無用に陥没してしまう変形を防止する。

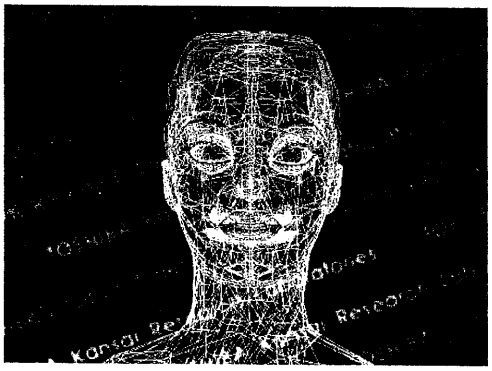

図4　ポリゴンモデル

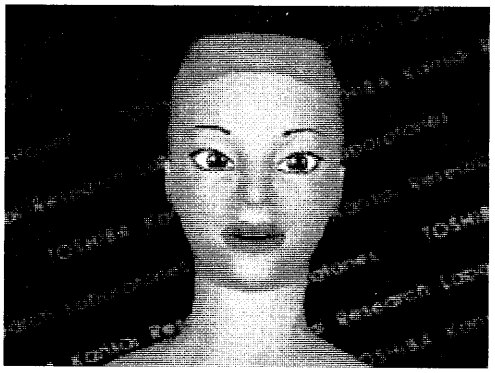

図5　生成画像

図6　顔向き制御例

有音期間における口形状の変形は、「あ」の口形状をベースにしてこれをランダムな開度に内挿することで実現される。変形量は口周辺の所定のポリゴン頂点に対して働く引力場で計算する。また、描画時の計算量を少しでも削減するために、閉口時には歯の描画を省略している。

3．6　考察

本システムを社内公開実験したところ、描画更新速度が低く音声と口の同期が不十分であったため、自然性についての評価は高くなかった。しかし、顔を持つしゃべる計算機への期待と関心は高く、研究の方向が誤りでないことが確認できた。

4　結論

以上、本報告ではポストGUIとしてmmHUIを提案し、その実験システムとして試作したナレーションエージェントについて報告した。

実験により顔のあるエージェントへの期待と関心は高いとの結論を得ることができたが、速度と品質の点で自然で生きているように見えるには課題の多いことも確認された。また、キーとなる音声認識／音声合成に関しても今回は実装しておらず、これらの要素技術を包含したマルチモーダル自由対話を実現することが大きな課題である。

今後は、このシステムをベースに様々な実験を試み、知見と技術の蓄積を行うとともに、応用システムへの展開を図る予定である。

参考文献

1)間瀬健二,「マルチモーダル・インタフェースのための画像処理」,第2回画像センシングシンポジウム講演論文集,pp.123-128,Jun.,13-14,1996

2)長谷川修,「擬人化エージェントによる対話型視覚情報学習システム」,第2回画像センシングシンポジウム講演論文集,pp145-150,Jun.,13-14,1996