“GazeToTalk”:
メタコミュニケーション能力を持つ非言語メッセージ利用インタフェース

○知野 哲朗　福井 和広　山口 修　鈴木 薫　田中 克己
{chino,fki,osamu,suzuki,tanaka}@krl.toshiba.co.jp
(株) 東芝 関西研究所
〒 658-0015 兵庫県神戸市東灘区本山南町 8-6-26

本稿では、視線検出技術、音声認識技術、および擬人化エージェント CG 技術によって、ユーザとシステムとの間の自然なインタラクションを可能とする “GazeToTalk” システムを提案する。本システムでは、まず、視覚情報処理によって擬人化エージェントへのユーザの注視を検出し、ユーザの音声入力意図を認識する。このユーザからの注視に対して、フィードバックとして、擬人化エージェントの顔表情およびジェスチャを提示することで、ユーザとシステムの間でのアイコンタクトを実現した。さらに、このアイコンタクトの成立状態に応じて音声入力を制御し、ユーザが入力を意図した音声だけに適切に反応し、周囲雑音や入力を意図しない音声による誤動作を起こさないインタフェースを実現した。また、システムの試用実験によって、非言語メッセージをその社会的機能に着目して活用することが、 HI のユーザビリティ向上への有益はアプローチであることを確認した。

“GazeToTalk”:
A Non-verbal Interface System with Speech Recognition, Gaze Detection, and Agent CG Output.

Tetsuro CHINO　Kazuhiro FUKUI　Osamu YAMAGUCHI
Kaoru SUZUKI　Katsumi TANAKA
Toshiba Kansai Research Laboratories
8-6-26 Motoyama-Minami-Cho Higashinada-Ku
Kobe, 658-0015, JAPAN

In this paper, we propose a new human interface system named “GazeToTalk”that is implemented by vision based gaze detection, acoustic speech recognition, and animated human-like agent CG with facial expressions and gestures. A natural interaction by vioce and non-verbal messages without use’s consious operation in real environment is realized.

1 はじめに

[Bolt80] など、80 年代初頭頃から開始されたマルチモーダルインタフェース研究は、近年の音声・画像情報の認識・生成技術の進展と、Post GUI としての期待の高まりにより活発化し、今までにも数多くの試作システムが開発されている [Maybury93]。しかし、その多くは、主にマルチモーダル入力された参照表現の照応解決 [Kobsa86, Koons93] や、その生成 [Wahlster91] などを課題とし、マルチモーダル入力の統合あるいは生成に関心を置くものであって、ユーザとの自然なインタラクションの実現には至っていない。

また、Post-GUI 実現へのもう一つの有力なアプローチとして、[Apple87] などに端を発する擬人化エージェント技術を挙げることが出来る。これは、(1) 表現力が高く、(2) より自然であり、(3) アプリケーションも広い、といった特徴を持つ [Wahlster97]。そのため、ユーザとの自然なインタラクションの実現に大きく寄与し、また、マルチモーダリティを HI に導入する意義を与えるものでもある。

これまでも擬人化インタフェース技術を用いた様々なシステム [竹林 94] [神尾 94] [鈴木 96] [知野 97, 河野 98] を開発し、研究を進めてきた。本研究は、この流れに沿ったものであり、非言語メッセージの活用によって、次節で説明するメタコミュニケーション能力を実現した新しいインタフェースを提案する。

2 メタコミュニケーション

メタコミュニケーションとは、「コミュニケーションを成立させるためのコミュニケーション」を意味する [黒川 94][1]。人間同士の対話では、あいづちや、うなずき、視線一致などによって、会話の開始や発話交替、あるいは話者聴者といった役割の認識がなされたり、あるいは情報の伝達状態の通知、確認、あるいは問い返しなど、通信路の確立、維持、解除などが実現されていると考えられる。こういったメタコミュニケーションは、対話能力の中で大きくかつ本質的な部分を占めるものであり、そこでは非言語メッセージ [Vargas87][2] が重要な役割を担っている。ところが、従来のインタフェースでは、例えばユーザとシステムとのインタラクションの開始、中断、あるいは終了などは、*a priori* とされ、メタコミュニケーションに関してまで考慮されたものはほとんど無かった[3]。

本研究では、このメタコミュニケーション能力の欠如こそが、現在のヒューマンインタフェースの重大な問題点であるとの立場を取る。そして、ユーザとのインタラクションに於いて、非言語メッセージを積極的に活用することにより、HI にメタコミュニケーション能力を与えることで、自然でより使いやすいインタフェースの実現を目指すものである。

3 “GazeToTalk” システム

従来の音声インタフェースの大きな問題点の一つに、「誰に向かって話されているのかが分からない」ことが挙げられる。そのため、システムへの入力を意図していない発声 (*e.g.* ユーザの隣にいる人に対する発声, 一人言) や、あるいは周囲雑音などによって誤動作が発生してしまっていた。これはメタコミュニケーション能力の欠如の一つであると言える。

他方、音声入力の受け付け可否を、従来の Push-To-Talk 方式[4] の様に、例えばボタンやマウス操作などによってユーザに制御させる様にすると、(1) 非接触、(2)*hand-busy* 時にも使用可などと

[1] メタコミュニケーションには、ゲームのルール説明や機器の操作方法の説明などを指す場合もある [Yvonne89]。

[2] コミュニケーションに於いて授受されるメッセージの内で文字として書き表せないものの総称。アクセント、イントネーションといった韻律から、表情、身振り、姿勢、さらに年齢や、服装までもが含まれる。

[3] 音声認識技術の問題点を、HI への応用を含んで広く検討した [嵯峨山 94] においても、この観点の指摘はなされていない。

[4] ボタン等を押すなどユーザの身体的な操作によって音声入力を可能とする方式。

いった、音声メディアの本来の利点をスポイルしてしまうことになる。

そこで、本研究では、画像処理による視線検出技術と、エージェント CG 技術[5]を用い、(1) 画面上の擬人化エージェントに対するユーザの注視を検知し、(2) エージェントの表情によってユーザへフィードバックを返すことで、(3) ユーザとシステムとのアイコンタクトを実現し、このアイコンタクトによって (4) 音声入力の受付可否を制御する “GazeToTalk” システムを開発した。図 1は、本システムの内部構成を示しており、また図 2は、本システムの使用の様子と画面例を表している。

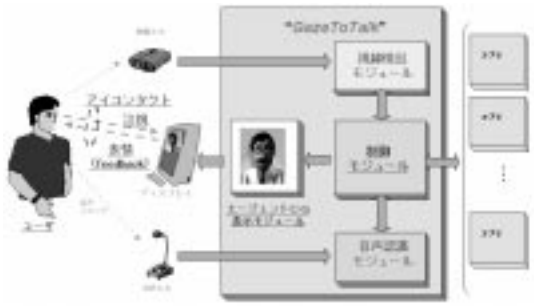
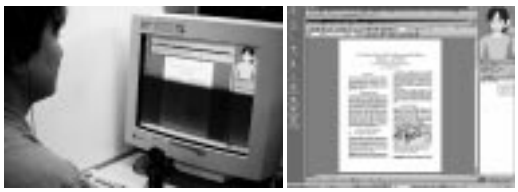


図 1: “GazeToTalk” システムの内部構成



(a) (b)

図 2: システム使用の様子 (a) と画面例 (b)

図 2(a) に於いて、ディスプレイ前下部に設置されているのがカメラであり、ここから得られる動画像情報が視線検出モジュールに渡される。またユーザの装着しているマイクからの音声信号は、音声認識モジュールへと渡され認識処理される。制御モジュールは、各モジュールから得られる情報に基づいて各モジュールを制御し、エージェント CG 表示モジュールは、エージェントの表情および身振りを動画像として生成し、ディスプレイを通じてユーザに提示する。

続いて、各モジュールの概要を説明する。

3.1 視線検出モジュール

本視線検出モジュールは、リアルタイムでユーザ注視位置の検出を行なう。ここでは、カメラから逐次得られる動画像に対する画像処理 [福井 97] を拡張した処理が行なわれ、(1) 画像からの顔領域の抽出および位置トラッキングによるユーザ検出と、(2) 目鼻等の部品領域候補の同定および目周辺のパターン情報の照合による注視位置の判定を行なっっている。本モジュールでの認識処理はソフトウェアのみによって実現されており、ディスプレイ面 9 分割の分解能での高速 (Indy R5000 180MHz 使用で 5 回 /sec) な視線検出を実現している。

さらに本モジュールは、画面特定領域への注視の他、ユーザの到来 / 離脱の検出や、ユーザの視線が画面上のエージェント以外の不特定領域内を推移している状態である ” 非注視状態 ” の検出なども行なう。 この非注視状態は、ユーザが本システムを使用中ではあるが、例えば画面上の他のアプリケーションを操作中であって音声入力を意図していない状態であることを検出するために利用している。

3.2 音声認識モジュール

今回のシステムでは、音声入力の受け付けの可否の制御に注目しているため、音声認識モジュールには、 PC 上で動作するオーソドックスな不特定話者対応の音声認識 [金沢 96] を用いた。認識対象語彙は、あらかじめ用意した語彙セットの中から、その時点でアクティブなアプリケーションに対応して随時自動的に設定され、音声コマンドとして認識できるようにしている。

[5] HI の窓口として擬人化されたキャラクターを CG で生成し利用する技術を、こう呼ぶこととする。

3.3 エージェント CG 表示モジュール

制御モジュールの指示に基づき、男女 2 体のエージェントの身振りを伴う多彩な表情をアニメーションで提示する (現在, 男女各 15 種類)。図 3,4にその画面イメージの抜粋を示した。

NORMAL HAPPY PARDON SORRY BOW

図 3: 女性エージェントの画面イメージ (抜粋)

NORMAL HAPPY PARDON SORRY BOW

図 4: 男性エージェントの画面イメージ (抜粋)

本モジュールは、 OpenGL を用いて実装されており、 PC 上で 3D Graphics をリアルタイム生成している[6]。また、エージェントの表示位置、サイズおよび表示状態も制御モジュールから制御可能としている。

ここでは、ユーザに対して、非言語メッセージとしての “ 表情 ” が提示されるが、そのタイミング (レスポンス) はシステム全体の効果および使用感に重大な影響を与える。しかし、本モジュールが実行される計算機の負荷状態の変動は避けることが出来ない。そこで本システムでは、エージェント CG の制御命令の時間的粒度を、各時点における計算機の負荷状態に応じて随時制御するこによって、常時安定した表情の提示を可能としている。また、本機構はエージェントモジュール単独での、専用高速 GWS 等パフォーマンスの異なる機械でのクロスプラットフォーム開発や表情提示タイミングのチューニング等を可能し、システムの開発効率向上にも寄与した。

[6] PC(PentiumPRO 200MHz,Windows-NT) で 8fps、アクセラレータボード追加で 24fps の速度のアニメーションを提示可能。

3.4 制御モジュール

本システムのようなマルチモーダルシステムの制御では、 (1) 複数の経路から非同期に入力がなされ、 (2) 複数の制御対象を同時に制御する必要がある。また、今回注目している非言語メッセージの特徴の一つとしては、意図を瞬時に伝達可能であることを挙げることが出来る。そのため、 (3) レスポンス (タイミング) がシステム全体のクリティカルな要因の一つとなる。以上の要件から、本システムの制御モジュールの基盤として、新たにタイムアウト付きの FSM (Finite State Machine) を開発した。

3.4.1 タイムアウト付き FSM

タイムアウト付き FSM は、ステートの集合、通常アークの集合、およびタイムアウトアークの集合から構成される。ステートおよび通常アークに関しては、従来の FSM と同様である。各タイムアウトアークにはタイムアウト時間と出力記号が情報として付与されている。また、各ステートは高々 1 つのタイムアウトアークの始点となり得る。

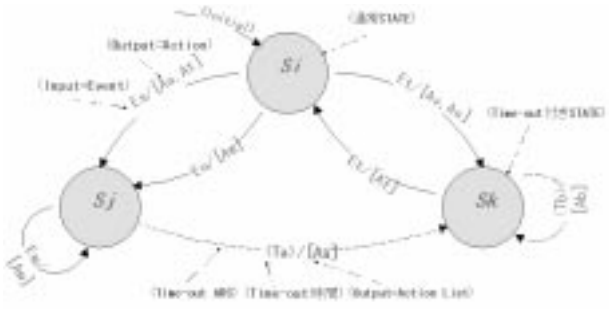


図 5: タイムアウト付き FSM の例

タイムアウトアークの機能の概要は以下の通りである。ある状態 S_t を始点とするタイムアウトアーク A_t が存在し、 FSM が、 S_t に推移した時点から他の状態に遷移することなく A_t のタイムアウト時間が経過した場合に、 A_t の出力記号を出力し、 A_t の終点のステートに推移する様構成した。

3.4.2 制御フロー

タイムアウト付き FSM の各アークの入力記号に、ユーザ注視 / 非注視検出、ユーザ到来 / 離脱検出など、本システムが検知可能な事象を割り振り、出力記号に、音声認識開始 / 停止、視線検出要求、あるいはうなづき表情提示など、各モジュールへの制御信号のリストを付与することによって、本システムの制御フローを表現し、これに基づいて制御モジュールが動作するよう構成した。

図 6は、制御フロー (抜粋) の概要を表している。システムは、通常右端のステート S_a にあり、ここを始終点とするループ状のタイムアウトアークにより t_1 毎に注視検出要求を行なう。ここで、ユーザのエージェントへの注視を検出すると、 S_b への遷移が発生し、音声認識が開始されるとともに、ユーザに対するフィードバックとして、“ 眉間の閃き ”[黒川 94] と呼ばれる表情が提示される。

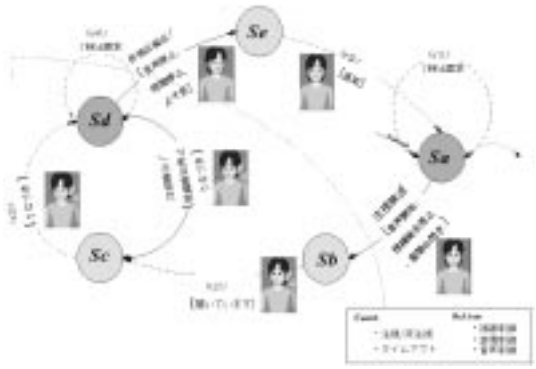


図 6: 制御フロー (抜粋)

その後、周辺雑音レベルに基づく音声検出パラメータの設定が行なわれた後、タイムアウトにより、 S_c への遷移が起こり、耳に手をかざしたエージェントの身振りによって、音声受け付け状態であることが伝えられる。その後、ユーザのエージェントへの注視が検出される度に、S_c-S_d 間のループによって、ユーザへうなづきの表情が提示され、通信路が維持されていることがユーザにフィードバックされる。その後、ユーザが画面の他領域へ視線を移すなどして非抽出が検出されると、 S_e へ推移し、音声認識が停止され、エージェントが視線を逸す表情が提示され、さらに t_5 後通常状態 (S_a) へ復帰する。

なお、現時点の試作システムの制御フローの全体の規模は、図 7に示した通りである。

図 7: 制御フロー全体概要

本制御フローに基づく処理によって、例えば視線検出モジュールへの視線検出要求は、図 6および図 7の濃い色で示されたステートに於いてのみ出力される。また音声認識処理も、音声受付けが必要な極限られた数のステートにある間だけ起動されるため、各認識処理の負荷は必要最小限に押えられている。

4 考察

4.1 検討

従来のシステムでは、視線検出技術は、例えばユーザの注意対象領域の推定に使われる程度であった。また、擬人化エージェントも、喜びや悲しみといった感情の表現、あるいは指し示しジェスチャなど、実際にはシンボリックな意味の明示的表現手段と一つとして使われることが多かった。会話における非言語メッセージの役割 (社会的機能) について言及している研究としては、 [Cassell98] を挙げることが出来る。しかし、そこでの論議は、計算機上でシミュレートされたソフトウェアエージェント同士の会話に限定されており、実際の非言語メッセージを認識する能力を持つものではなく、かつ実時間

動作も不可能であって、人間とのインタラクションを行なうことが出来るものではない。

これに対し、本研究では、実働するヒューマンインタフェースシステムに於いて、視線、表情、および身振りといった非言語メッセージを、ユーザとの通信路の確立、維持などといったメタコミュニケーションに利用している。つまり、非言語メッセージの社会的な機能に着目した利用を行なっていることが、本研究の一つの特徴である。

非言語メッセージの社会的機能は、通常人間同士の会話では、無意識的に利用され効果を発揮している。そのため、これを適切にインタフェースに組み込むことは有益であるだけでなく、そのシステムを利用するユーザにとって、認知的コスト増加をほとんど生せず、かつ学習不要であるという利点を持つ。

さらに、擬人化インタフェースに対しては、[Reeves&Nass96] の主張する以下の性質がより強くなり、この性質への対応能力の欠如は、そのままそのインタフェースの重大な欠点となる。

> ユーザはシステムに対して社会的ルールを無意識的、自動的に適応し、かつシステムが社会的であることを期待する。
>
> [Reeves&Nass96]

つまり、擬人化エージェントを用いた HI の社会的能力は、むしろ必須の要件である。そして、本システムは、その社会性の一部を実現したものであると言える。

実現上の観点からは、本システムは、視線検出、音声認識、および 3D-CG 生成と言った処理負荷の高いモジュールを複数利用して実現しているが、図 6および図 7に示したように必要最小限の状況だけで各モジュールを動作させているため、全体の処理負荷を小さく押えることに成功している。

なお、本技術の応用先としては、駅務 / 金融端末や、エンタテイメント / エデュテイメント領域などが想定できる。

4.2 知見

本システムの試用によって以下の知見を得た。

まず、本システムでは、ユーザが他の人物と (音声で) 対話している途中であっても、自然にシステムへの音声入力を行なうことが出来た。つまり、ユーザが、例えば他の人物へことばで説明を行なうなど、システムへの入力以外の意図で音声を使っている状態で、本システムへの音声入力が必要になった時に、説明相手である人物の顔からエージェント CG へ視線を移すことで、特別な切替え操作無しで、音声コマンド入力によるアプリケーション操作を行なうことが出来た。また、本システムでは、ユーザが他の人物と音声で相談しながらワードプロセッサソフト (MS Word) を利用した (キーボード入力による) 文書作成を行なっている状態で、ユーザがエージェントを注視し音声コマンドを発声することで、作業を中断することなく、表示形式の変更などを行なうことが出来た。

これらは、従来の音声インタフェースでは不可能であったことであるが、実際の使用状況ではよく起こり得る状況であると考えられるため、本システムの有用性を表すものであると考えられる。

また、オフィスなど、断続的な周囲雑音のある状況下では、ユーザが雑音の少ない時を見計らって、エージェントを注視することで、余分な操作無しで適切に音声入力を行なうことが出来た。また、ユーザがシステムに対して音声入力を行なっている間に、例えば他の人物がユーザに近付いてきたり、あるいは話しかけるなどしてきた場合にも、ユーザがその人物の方向を向くだけで、自動的に音声入力の受付が停止され、その人物との対話のための発話によるシステムの誤動作を回避することが出来た。さらに、その人物との会話が終了した時点で、ユーザがエージェントの方向を向き直すことで、特別な操作なしでシステム音声入力を再開することができたなど、本システムの有用性が確認された。

4.3 主張

著者らが重要であると考えているのは、「相手に喋りかける時は相手に視線を向け、受け手がその意図を受け入れる場合には表情によってフィードバックを返すこと」が、人間同士の対話の場合と同じであって、社会的コンセンサスのとれたルールに沿った、「非言語メッセージによるメタコミュニケーション能力を実現している」といえる点である。

つまり、本研究の主張は、以下の通りまとめることができる。

> (無意識の行動による) 非言語メッセージを、メタコミュニケーションという観点から、状況に応じて的確に解釈 / 利用することで、自然で、円滑で、かつユーザの負担増加を起こさない新しいインタフェースを実現出来る。

今後、この方針に基づく新しい HI の実現を目指す。

4.4 課題

まず、本システムの試用によって明らかとなった現状システム固有の問題点としては、「行なう作業の種類によっては、画面上の作業領域とエージェント CG の表示領域の空間的配置に起因してインタラクションに不自然さが発生し得ること」などを、挙げることが出来る。試用から判明するこういった問題点については、今後さらなる検討を行なう予定である。

本システムの一般的課題としては、(1) 各認識モジュールのロバスト化、(2) ダウンサイジングなどが挙げられる。また、視線に反応する擬人化エージェントは、システムへの親しみやすさをもたらし、計算機システム利用へのバリアを軽減する効果が期待される。そして、本システムを動作を見た一般者の反応からもそれが推察されたが、こういった HI の効果の (3) 定量的評価も大きな課題として挙げられる。

5 おわりに

本稿では、試作した “GazeToTalk” システムについて報告した。また、その試用によって、自然で使いやすい HI 実現という目標に対し、非言語メッセージを活用し、メタコミュニケーション能力を適切に実装していくことが、有望なアプローチであることを確認した。

今後、メタコミュニケーション機能の解明を図るとともに、活用する非言語メッセージの拡大によって、本研究を進める。

参考文献

[Bolt80] R.A.Bolt, ”Put-That-There”: Voice and Gesture at the Graphics Interface, *Computer Graphics*,14(3), pp.262-270, 1980.

[Kobsa86] Kobsa,A., Combining Deictic Gestures and Natural Language for Referent Identificaiton, ACL, *Proc.of COLING86*, pp.356-361, 1986.

[Apple87] Apple, The Knowledge Navigator, (*Concept Video*), 1987.

[Vargas87] バーカス, 非言語コミュニケーション, 新潮選書, 1987.

[Yvonne89] W.Yvonne, 植草訳, Cognitive Aspects of Computer Supported Tasks, (人とコンピュータのサイコロジー), John Wiley&Sons Ltd.,(BNN), 1989,(1991).

[Wahlster91] W.Wahlster, Designing Illustlated Texts: How Language Production is Influenced by Graphics Generation, *Proc. of EACL'91*, 1991.

[Koons93] Koons, D.B., Sparrell, C.J., Rhorisson, K.R., Integrating si-

multaneous input from speech, gaze, and hand gestures, In Maybury, M.T. (ed), *Intelligent Multimedia Interfaces*, pp.267-276, 1993.

[Maybury93] Maybury, M.T., Intelligent Multimedia Interfaces, The AAAI-Press/MIT-Press, 1993.

[神尾 94] 神尾, *et.al.*, マルチモーダル対話システム MultiksDial, 信学論, Vol.J77-D-II, No.8, pp.1429-1437, 1994.

[黒川 94] 黒川, ノンバーバルインタフェース, オーム社, ヒューマンコミュニケーション工学シリーズ, 1994.

[嵯峨山 94] 嵯峨山, なぜ音声認識は使われないか・どうすれば使われるか?, IPSJ-SIG-SLP,94-SLP-1-4, 1994.

[竹林 94] 竹林, 音声自由対話システム TOSUBRG-II, – ユーザ中心のマルチモーダルインタフェースの実現に向けて –, 信学論, Vol.J77-D-II, No.8, pp.1411-1428, 1994.

[Reeves&Nass96] B.Reeves, C.Nass, The Media Equation - How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge Univ. Press, 1996.

[金沢 96] 金沢, *et.al.*, 雑音免疫学習を用いたサブワード HMM に基づく 雑音環境下での音声認識, 日本音響学会平成 8 年度春季研究発表会講演論文集,2-5-13, pp.83-84, 1996.

[鈴木 96] 鈴木, 山口, 福井, 田中, 倉立, 松田, 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)-, IPSJ-SIG-HI, 69-1, 1996.

[長尾 96] 長尾, インタラクティブな環境を作る, 認知科学モノグラフ (2), 共立出版, 1996.

[Wahlster97] W.Wahlster, E.Andre, Intelligent Multimedia Interface Agents, IJCAI-97, *Tutorial*, (SP1), 1997.

[知野 97] 知野, 河野, 屋野, 池田, 鈴木, 金沢, 音声入出力、 タッチジェスチャ入力、およびエージェント CG 出力を持つマルチモーダル対話試作システム, IPSJ-SIG-SLP, 1997.

[福井 97] 福井, 山口, 形状抽出とパターン照合の組合せによる顔特徴点抽出, 信学論 D-II, Vol.J80-D-II, No.8, pp.2170-2177, 1997.

[河野 98] 河野, 屋野, 池田, 知野, 鈴木, 金沢, ATMS ベースのマルチモーダル入力統合方式を用いたインタフェースエージェントシステム, 人工知能学会誌, 1998.3(採録予).

[Cassell98] Jastine Cassell, *et.al.*, Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multile Conversational Agents, in *Reading in Agents*, 1998.