解 説 ネットワーク社会を支援する新しい知能メディア技術

6. ネットワーク利用者を支援するマルチモーダルヒューマンインタフェース

Multi-Modal Human Interface for Network Users by Tsutomu MIYASATO and Kenji MASE (ATR Media Integration & Communications Research Laboratories).

宮里 勉[1] 間瀬健二[1]

1 ATR 知能映像通信研究所

1. はじめに

あたかも人間同士で対話しながら相手に意図や命令を伝えるかのように，コンピュータや機械を使うことができたらどんなにか便利だろう．

近年，ハードウェア技術の進歩によって高解像度のディスプレイや高速なCPU，あるいはグラフィックス専用のVLSIなどが安価に利用できるようになってきた．また，コンピュータのダウンサイジングなどにより機器のパーソナル化が進む一方で，ネットワークを介して多くの資源を利用し，人々と協調して作業を進めたいという要望が増加している．そのような要望とともにクローズアップされるのが機械の使い勝手の悪さ，すなわちヒューマンインタフェースの良否の問題である．

このような背景および画像処理，音声処理，自然言語処理，CG技術，仮想現実感などの諸技術の発展にともない，テキストを中心とするキーボードやマウスやボタンのような入力デバイスを手で操作し，テキストやグラフィックスで出力を得るという従来型のヒューマンインタフェースも変化しつつある．そして，音声情報，映像情報，触覚情報など複数の情報センシングによる入力と，グラフィックスや動画像と音声を融合した出力との自然なインタラクションを可能とするマルチモーダルインタフェースの研究が盛んに行われている．

しかし，本当に便利だろうか？ いったいどんな状況で機械やそのインタフェースに何を期待するのかよく考える必要がある．たとえば，車で，ある目的地に行く時に，「もうちょっと右」とか「少し戻して」とか教習所の教官よろしく，言葉を使って車に命令したいのだろうか？ 私たちが期待しているのは，タクシーのドライバのように，「京都駅まで」といえば「新幹線口ですね？」と気を利かせ，安全に早く連れて行ってくれるようなシステムではないか．そして，「新幹線口」と判断したのは乗客の荷物や話し方の情報を用いたのかもしれない．つまり，“誰”が“どこ”でそのメッセージを発したかということが，“どんな内容”か以上に重要なことがあるし，メッセージの解釈に大きく影響している．

このように，同じ信号でも状況やほかの信号との統合で別の解釈にもなりうる情報群を処理できるインタフェースが必要である．本稿で述べるマルチモーダルインタフェース(以後MMHIと呼ぶ)は，情報群の主たる発信源である人間から，状況に適切なメッセージを抽出して機械との自然な対話を実現してくれるインタフェースであり，種々の情報の集約によるユーザへの気軽さの提供や解釈の曖昧性の除去などに非常に有効である．

本稿では，MMHIの具体例を示して現状や課題を解説する．なお，紙面の都合で音声・画像処理のアルゴリズムに関する詳細は省く．

2. マルチモーダルヒューマンインタフェース

2.1 マルチモーダルヒューマンインタフェースとは？

マルチモーダルヒューマンインタフェースとは，それを利用するユーザの認知情報処理の複数の様式(モダリティ)が統合あるいは組み合わされて用いられるインタフェースである．

類語にマルチメディアインタフェースがある

が，マルチメディアインタフェースは単にメディア(音，映像，触覚など)が複数になっていることを表すのに対し，それぞれのメディアがいろいろな形態で使われ情報伝達を行っている時に，マルチモーダルインタフェースと呼ぶと考えられる．たとえば，同じ音でも言葉としての音声，韻律，擬態語，摩擦音や落下音，のように分類するとモダリティを考えることができる．あるいは人差し指を伸ばす動作の映像は，1という数字，物体の指示，口にあてて静かにという命令，など数種類のメッセージを手の同じ映像というメディアから伝達するときにマルチモーダルだということができる．

したがって，MMHIでは，グラフィックスやテキストなどの複数のメディアが同時に利用でき，人間の動作，ジェスチャ，顔の向きや視線・表情なども入出力情報として使われる．

2.2 なぜマルチモーダルヒューマンインタフェースなのか？

現在の情報機器とのインタフェースは，効率，操作性，認知的負担度などの点で満足できないものが多い．

我々人間の特徴は言語を使用して互いにコミュニケーションをすることにあるが，しかし，単に言語だけが使われるのではなく，図-1のようにいろいろなメディアを介して意識的あるいは無意識に五感に訴えるような情報の送出を行っている．

したがって，コミュニケーションにおいて十分な意思疎通が成り立つには，互いの五感情報が十分に伝わることが重要であり，プレゼンテーションや電気通信においてもマルチメディア化が進むのは自然な流れである．また，当然，コンピュータとのインタラクションも複数のモダリティを用いるマルチモーダルインタフェースが主流になることは疑いのないことである．

従来の対話システムでは，機械に向かって話しかけるというユーザの心理的抵抗感や応答の遅れなどによって，ユーザがシステムの状態を把握できなくなるという問題が生じる．そこで，ユーザが意図する操作が直ちに実行でき，ユーザに認知的負担をかけないインタフェースが必要であり，視覚メディアなどを用いて，システムの状態を分

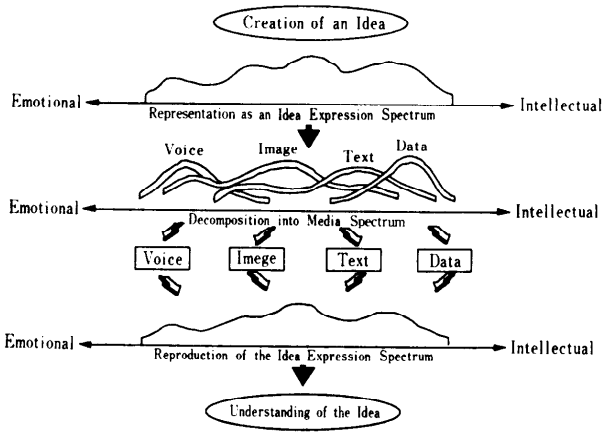

図-1 メディアスペクトル[1]

表-1 マルチモーダルヒューマンインタフェースの特徴

システムへの入力	冗長性	複数メディアを使うことで，ユーザは同じことをさまざまに表現することが可能である．
	気軽さ	ユーザは普段慣れているメディアを使うことができる．
	頑健性	複数のメディアからの情報を総合するので，各メディアからの情報は完全でなくてもよい．
	状況依存性	ユーザの状況に応じてメディアを使い分けることが可能である．
システムからの出力	論理性	文字や表を中心とする論理的な表現が可能である．
	直感性	図やグラフを用いることで数値的な比較を直感的に表現できる．
	実感性	写真，動画によって実物に近いものを実感できる．
	情感性	視覚的，聴覚的なメディアを利用することで情感的な表現が可能となる．
	対話性	情報を複数メディアで表現できるので状況に応じた情報の利用や体験により正確な伝達が期待できる．

かりやすくしたり，ユーザの発話を促進したりすることができる．

したがって，MMHIの要件としては，人の各感覚に対応するモードを備え，機械がユーザの直感や感性に直接訴えかけることができる表示チャネルをもち，人の通常の行為のモードを備えて各種の意識的，無意識的な身体動作が直接的に機械に伝わること，があげられる．表-1にMMHIの特徴をまとめる．

2.3 マルチモーダルヒューマンインタフェースとVR

MHHIと非常に密接な関連がある技術として五感情報を人工的に作り出すVR(バーチャルリアリティ)がある．

最近では，パソコンに限らずワークステーショ

ンでもGUI(グラフィカルユーザインタフェース)が一般的になっており，モニタ画面上に表示されるアイコンを指示することで，ファイルの開閉など，各種のコマンドを起動できる．

アイコンにより，コマンドなどの機能が可視化されて直観的にその機能が分かるようになる．この考えをさらに発展させると，空間自体をアイコン化したり，あるいは機械がユーザの身体のサイズに合わせて伸縮したり，機械自らが腕を伸ばして必要なスイッチやボタンを指し示したり，あるいは喋ったりすることも考えられる．そのようなことはすでに実現されつつある．すなわち，機器の操作においていちいちマニュアルを読まなくとも，実物に重畳して操作ガイドを示すことができる．たとえば，複写機操作のマニュアルで「A4紙をトレイに入れて….」に代わり，実際のA4紙トレイの中にA4紙が入る映像が実物に重畳されて示される，などである．そのようなことは，計算機で仮想の空間を生成するVR技術により可能である．

ネットワークを介した在宅勤務も，単に普段の自宅の空間内で端末に向かうのではなく，VRによって自宅に居ながらにしてある時は本社の会議室内であったり，またある時は資料室の中に入り込むことが可能となる．そして，資料室の例でいうと，「部屋に入ってすぐ右側の棚の上から3段目の中央付近にある赤背表紙の本」のように体の感覚を使って書物や資料を探せるので，端末からのキーワードの入力による検索よりも容易であり，記憶にも残りやすい．

3. 入力・認識中心のユーザインタフェースから出力・表現を考慮した対話指向のユーザインタフェースへ

ユーザの心理的抵抗感を軽減する1つの手段として，システムの擬人化がある．たとえば，擬人化された案内役のキャラクタがユーザからの質問に対話的に答えるなどであり，その際擬人化キャラクタは表情などを豊かに使って対話を円滑に行うことが期待される．

これまでのインタフェースのような，どちらかというと一方向的なユーザあるいはシステム主導型のものとは異なり，人間とコンピュータの間でインタラクションの主導権が必要に応じて移り変わるスタイルが重要になってきている．このような能動性のあるインタフェースはインタフェース・エージェントと呼ぶことができる．このエージェントは人間同士の対話を模倣することを目指しているため，人間とエージェントとの対話はマルチモーダルになる．

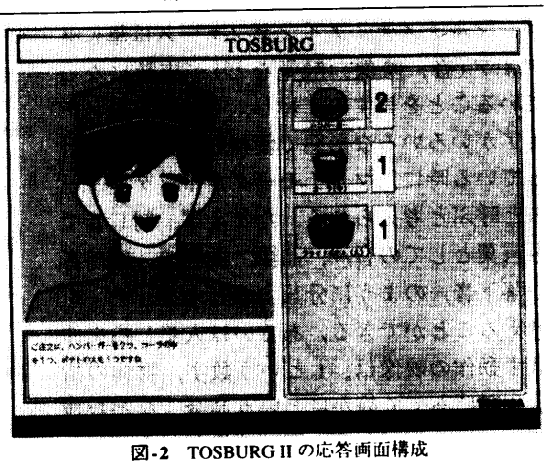

図-2 TOSBURG IIの応答画面構成

3.1 マルチモーダルヒューマンインタフェースを活用した対話システムの具体例

マルチモーダル対話システムとして，音声認識応答に擬人化されたキャラクタの表情を使ってユーザにシステムの状態を伝達するTOSBURG IIや表情アニメーションシステム，またポインティングデバイスと音声認識を使ったFingerPointer，さらに実際の人間をコンピュータグラフィックス(CG)で再合成して遠隔通信会議の臨場感を高めた臨場感通信会議などがある．

(TOSBURG II[2])

ユーザと計算機との自然な対話の実現には，音声認識にとどまらず，対話状況や内容を理解・考慮して人間と計算機が互いに影響し合う音声対話処理と，音声以外の視覚メディアなどを併用するマルチモーダルインタフェースが重要となる．

TOSBURG IIは，擬人化エージェントのCGで表現された反応と音声対話技術とがうまく組み合わせられた不特定ユーザ向きの実時間音声対話システムの例である．不特定多数のユーザがアクセスする身近な例として，ハンバーガ店での注文対話を対象にしている．

人間同士の会話のようにシステムを擬人化し，音声に対して自然に音声で答えやすくするように，さらに店員の顔の表情が変化し唇が動く2次

元のカラーアニメーションで提示される．したがって，ユーザは擬人化店員の表情から，対話音声理解の結果の確信度と対話の進行状況を理解できる．また，音声認識結果の確認応答提示の際には表情に加えて，注文品，個数，サイズも視覚的に提示し，応答文の内容もテキストで表示するとともに，規則合成音声で発声する，というマルチモーダル応答になっている．したがって，合成音声の品質が十分でない場合でも，ユーザはシステムの応答を容易に確認・理解でき，音声メディアの特徴の一過性の欠点が補われている．

図-2 に画面表示の構成を示す．ここで，合成音声と画面下の応答文はあいさつや確認などのメッセージ出力に用い，画面右にはシステムが理解している注文品目と個数およびサイズをユーザに提示する．また，店員の表情を対話状況に応じて変えるとともに，その口の動きの開始タイミング・時間を合成音声に合わせている．対話の開始時は笑顔で始まり，対話の停滞時や誤りの指摘には悲しい表情をし，最後はおじぎをして終わる．

(表情アニメーションシステム[3])

このシステムは，コンピュータ関連の製品，たとえば，ワークステーションやパソコンの価格，サイズ，重さ，機能，そしてCPUの仕様など，について音声対話形式でユーザの質問に答えることができる．その際，システムは音声応答とともに顔の表示による種々の会話的表情を示す．

顔の表情は3次元CGにより作成され，筋肉の収縮と口の動きや顎の向き，目の動き，顔の向きなどを制御して表情が生成されている．さまざまな表情は表面上への写真を用いたテクスチャマッピングによってリアルな顔を表現している．なお，対話における表情は，次の3つに分類されている．

(1) 構造的表情：特定の語・句・節を強調する表情，発話の構文的側面にかかわる表情，話の全体構造にかかわる表情(例：話者の質問がシステムの予想外であった場合の，"驚き"［いわゆる目が点］の表情)，

(2) 話者表情：発話中の事柄を例示する表情，発話中の事柄に情報を追加する表情(例：「知りません」という時の眉を急に上げて口元を下げて戻す表情)，

(3) 聴者表情：話者の発話に対する反応として作る表情(例：入力音声が認識できない時や認識結果が文をなさない時などの，"不安"［眉を下げた困ったような顔］の表情)．

図-3 FingerPointer

図-4 臨場感通信会議による3地点の参加者の協調作業

音声対話の過程では，対話で重要ないくつかの典型的な状況を認識し，これらを表情と関連づけている．

(FingerPointer[4])

計算機が音声と表情を使えば，人間側も音声と表情を組み合わせた入力を使えるようにするとバランスがよい．FingerPointerは，手の動きである指示動作と指文字のジェスチャ認識と音声認識を組み合わせてマルチモーダルな入力を可能にしている(**図-3**)．手の動作を画像処理により抽出してポインティングデバイスとして用いるとともにその多義性を音声で補完することにより，計算機に対するコマンドを完成させている．FingerPointerを用いたプレゼンテーション応用では，**「動作：人指し指でポインティング」**「音声：ここから」，**「動作：指を動かす」**「音声：ここまで」という一連のアクションによりプロジェクタ上に線を引いたり，**「動作：指を3本示す」**「音声：これだけ進めて」というアクションでス

©1996 土佐尚子
図-5 音声に含まれる感情に反応するMIC

ライドを3ページ前進させるといった，マルチモーダルなインタフェースが実現できている．

(臨場感通信会議[5])

ネットワークを介しての協調作業には，従来のTV会議システムにはない人間の感情までも伝える臨場感が求められる．TOSBURG IIや表情アニメーションシステムは擬人化されたキャラクタの表情を使っているが，臨場感通信会議システムでは，人物の表情だけでなく身体全体をCGで合成している．CGで生成した仮想的な3次元空間内に会議参加者の人物像を合成立体表示することにより同一の会議室にいるかのような感覚を再現する．CG人物像は3次元で生成した像であり，視線の動きも再現されるためリアリティ性が高い．また，CG人物像を用いるほかの利点として，表情の制御がある．日本人の表情は欧米人に比べて乏しいといわれるが，CG顔により表情をオーバにしてめり張りをつけることができる．したがって，CGによる表情合成により，ノンバーバルな情報を相手に誇張して伝達することや会議相手の文化や国民性に合わせた表情の翻訳も可能となる．

臨場感通信会議では，面談会議の実現のほかに，遠隔地間での協調作業も可能である．たとえば，遠隔地間で話し合いながら物体を製作する際に，仮想の物体を手で掴んで移動させることができるほかに，ジェスチャと音声認識を併用しての，仮想物体の配置，結合，移動，拡大・縮小，形状の変形，が行える(図-4)．

3.2 擬人化エージェントを用いた感性コミュニケーション技術

3.1節の擬人化エージェントでは極力実物に近づけようとしていたが，ここであげるMICと呼ばれるキャラクタは，よくみると外見は実物と異なるが，全体の動作や反応が生き生きとしているというエージェントの感性面が重視されている[6]．

マイクを通してMICに話しかけると，MICは声の高低，強弱，抑揚などを喜び，怒り，驚き，悲しみなどの感情に対応させる(図-5)．そして人間の声に含まれる感情に応じてCGによる表情と動作で反応する．したがって，同じ発音でも話し手により反応は微妙に変化する．

感情への対応化には，多数の人により感情を込めて話されるいろいろな単語をニューラルネットであらかじめ学習しておき，学習された声の調子と入力される声を比較して最も似ているパターンを対応する感情としている．

MICへの話しかけはちょうど赤ん坊をあやすのと似た感覚を人間に与える．MICに優しく「こらこら」と話しかけると笑顔を見せるが，怒気を含んだ「こらこら」では声を上げて泣き出す．話しているうちに，冷たい機械だと思っていた計算機が人間味ある応答をすることに驚かされ，親しみを感じるようになる．

4. 今後の課題

4.1 人間の動きや状況の認知

これまで研究されてきた音声認識やジェスチャ認識によるメッセージ抽出は，もっぱら声や動作をキーボードやマウスのかわりに用いることを目指してきたのではないだろうか？ たとえば人間の動作はマウスなどでは入力しきれない多様なメッセージを発している．動作検出や認識という要素技術をさらに発展させてその人の状況を認識したうえで，どのように動作に解釈を与えるかの研究が重要となってくるだろう．

たとえば，ヒューマンリーダ・プロジェクト[4]では人間が発するあらゆるノンバーバルな情報をメッセージとみなして，表情を含むジェスチャ認識や個人識別，それらの要素技術のインタフェースへの応用について検討がなされた．前述のFingerPointerもその一例である．また，スマートルーム・プロジェクト[7]では同様な要素技術の研究の上に，さらに人間の状況理解をいかに利用してスマートで賢いインタフェースにするかの研究へと発展させている．

4.2 自律エージェントの振舞い

コンピュータ上に擬人化されたキャラクタが現れてユーザとやりとりをするようになると，その擬人化に合わせてエージェントは知性をもち自律

的に振る舞うことが期待される．これまでは，もっぱらエージェントの自律性について，インタフェースが目的とするタスクについてはよく議論されているが，擬人化したインタフェース・エージェントにとって本質的な自律性とは何かはまだよく議論されていない．タスクに依存しながら，状況に合わせて自律的に制御してインタラクションの仕方を変え，さらに，ユーザの振舞いを学習してユーザに適応的に変化していくインタフェースを考える必要があろう．そのときに知能情報処理による，状況理解技術，ユーザ適応技術，インタラクション制御技術などが必要になってくると思われる．

5. むすび

人間にはメッセージ発出の仕方に複数のモダリティがあり，人間同士のコミュニケーションはマルチモーダルである．人間対人間の場合はインタフェースというより，マルチモーダルコミュニケーションと呼ぶ方が適切であるが，人間同士のコミュニケーションでマルチモダリティがいかに使われるか，ノンバーバル言語や動作学の研究に学ぶことができる．

近い将来，ネットワークを意識しないモーバイルコンピューティングやコンピュータが遍在する環境のユービキタス（Ubiquitous）コンピューティングが日常的になると思われる．その際には，電子メイルの普及により，隣の席の人との会話においても肉声よりも端末のスクリーン上でのテキストコミュニケーションを好む人が出てきているように，従来の対面会話様式に基づかないまったく新しい人間同士のマルチモーダルコミュニケーション様式が出現するかもしれない．

いずれにしろ，マルチモーダルヒューマンインタフェースはますます重要となることは間違いないことであろう．そして，CG による合成表情が自然になればなるほど，それを実現している高度な技術に気づいてもらえないように，将来の MMHI の技術もあまりに自然であるがゆえに誰にも気にされなくなるであろう．

参考文献

1) Watanabe, H. : Integrated Office Systems: 1995 and Beyond, IEEE Communications Magazine, Vol. 25, No. 12, pp.74-80 (1987).
2) 竹林洋一：音声自由対話システム TOSBURG 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.1417-1428 (1994).
3) 竹内彰一，長尾　確：新たなコミュニケーションモダリティとしての表情，情報処理学会研究報告 IM-9, pp. 25-34(1993).
4) 末永康仁，間瀬健二，福本雅朗，渡部保日児：Humanreader:人物像と音声による知的インタフェース，信学論，Vol.J75-D-II, No.2, pp.190-202 (1992).
5) 岸野文郎：ヒューマンコミュニケーション－臨場感通信，テレビジョン学会誌，Vol. 46, No. 6, pp. 698-702 (1992).
6) 中津良平：アートと工学の融合によるエージェント生成，Infor-Tech '96, pp. 25-32 (1996).
7) Pentland, A. : Smart Rooms, Scientific American, pp. 68-76 (Apr. 1996).

（平成 8 年 11 月 1 日受付）

宮里　勉（正会員）

1953 年生．1976 年電気通信大学電子学科卒業．1978 年東京工業大学大学院修士（電子システム）課程修了．同年国際電信電話（株）（KDD）入社．研究所を経て 1993 年より ATR に出向．現在（株）ATR 知能映像通信研究所第五研究室長．工学博士．仮想環境を応用したコミュニケーション環境の生成とヒューマンインタフェースが主な研究テーマ．電子情報通信学会，テレビジョン学会各会員．

間瀬　健二（正会員）

1956 年生．1979 年名古屋大学工学部電気学科卒業．1981 年同大学院修士（情報）課程修了．同年日本電信電話公社（現在 NTT）入社．1988 ～ 89 年米国 MIT メディア研究所客員研究員．1995 年より（株）ATR 知能映像通信研究所第二研究室長．工学博士．コンピュータグラフィックス，画像処理とそのヒューマンインタフェース，コミュニケーション支援への応用が主な研究テーマ．訳書「ロボットビジョン」（朝倉書店，共訳）．IEEE，電子情報通信学会，日本情報考古学会各会員．