パネルディスカッション(1)

マルチモーダルインタラクション
～今、どのような視点を必要としているのか～

畑岡信夫
Nobuo Hataoka
(株)日立製作所中央研究所知能システム研究部
Central Research Laboratory, Hitachi Ltd.

1．はじめに

本パネルディスカッションでは、現在研究が活発化しているマルチモーダルインタラクションに関して議論を行なう。具体的には、何故マルチモーダルが台頭して来たのか、そして現状の技術レベルを踏まえて、今後への問題提起を行ない、真の技術とするために必要なポイントを明確にする。

2．マルチモーダルインタラクションとは

2.1 マルチモーダルインタラクションの解釈

人と人との意志伝達の観点から見た場合、我々は音声や身振り、手振り、そして顔の表情などあらゆるモダリティを使って、コミュニケーションを行なっている。即ち、マルチモーダルなインタラクションを使用したコミュニケーションである。これを人と機械との関係に置き換えて、次の２つの観点から、マルチモーダルインタラクションを捉える。第１は、マルチモーダルインタラクションを可能とするインタフェースである。マルチモーダルインタフェースと呼ばれている。第２は、マルチモーダルインタラクションが可能なシステム、あるいは仕組みである。ＡＴＲの臨場感通信システム[1]やＭＩＴメディアラボのバーチャル犬サイラス等が例として挙げられる。以下、本稿では第１の捉え方であるマルチモーダルインタフェースに関して、言及する。

2.2 マルチモーダルインタフェースの定義

マルチモーダルインタフェースは、複数の入出力手段を、同時または逐時に用いることが可能な、人間にとって自然なインタフェースとして注目されている。まず、モードとモダリティの違いを、モードは音声や画像のように記憶された情報の形態(メディア)を表し、モダリティとは情報それ自身と情報の認識、知覚までも含んだ概念として捉える。従って、マルチモーダルインタフェースとは、複数の入出力手段を有し、これらの入出力手段の解釈をも行なうインタフェースと定義される。

マルチモーダルインタフェースは、モダリティの入力順序（継時、同時）と情報統合の仕方（独立、統合）の２軸から表１のように分類される[2]。その中で、音声を利用した場合には、音声と他の入力手段を同時に使うことができるというメリットがある。我々は、音声とポインティングジェスチャを同時に利用し、かつこれらの情報を統合した共働的(Synergistic) なシステムを既に提案している[3][4]。

表1 マルチモーダルインタフェースの分類

		USE OF MODALITIES	
		Sequential	Parallel
Fusion	Combined	ALTERNATE (交互)	SYNERGISTIC (共働)
	Independent	EXCLUSIVE (専用)	CONCURRENT (同時)

3．マルチモーダルインタフェースの観点からの問題提起

(1)入出力モダリティの選択と認識技術

入力モダリティとしては、音声や、ジェスチャ（身振り、手振り、ポインティング等）、表情、視線等がある。出力モダリティとしては、ＣＧ、アニメーション等がさらに加わる。モダリティの選択が第１の課題であり、音声認識技術や画像認識技術を駆使して、これらの入出力モダリティを如何に認識して、利用するかが課題となっている。

(2)情報統合の方式

入力された複数のモダリティをどう統合するかが課題となっている。我々は、情報統合テーブルを利用した方式を提案している[3]。また、時間同期を詳細に解析し、利用する試みもある[5][6]。

(3)インタフェースの拡張、例えばエージェント型

エージェントは、ユーザの指示に従って自律的に動作を行なう代理人モデルとして捉えられる。エージェントを用いた場合、エージェント型インタフェースとエージェント型システムという２つの概念がある[7]。前者は、ユーザインタフェースから見た応用であって、画面上にエージェントが表示され、ユーザとシステムとのインタフェースを仲介する。後者は、ソフトウェア構造、ネットワーク管理から見た応用であって、複数のエージェントが協調しあって作業を効率良く進めて、生産性と性能向上、通信コストの低減を行なう。対話機能を有したエージェント型インタフェースが開発されており[8][9]、３次元ＣＧや動画など、エージェントの表示方式も課題となっている。

(4)対話制御方式

エージェント型インタフェースでは、対話制御が課題となっている。初心者から熟練者までが、それぞれの観点から自由にシステムを扱えるようにするために、エージェントを介した対話機能を高度化することが重要となっている。具体的には、システムに不慣れな段階では、システムからの誘導に基づいて入力を行えるシステム主導の方式、その後熟練するに従って、ユーザが自由に操作できるユーザ主導の方式へと適応的に移行する対話機能が必要となっている[8]。

4．最後に

マルチモーダルで、エージェント型インタフェースは、ユーザにとって、使いやすく、そして親しみのあるインタフェースとなっている。我々が日常使っている音声やジェスチャなどを使って自由に計算機や機械を操作できたら、という夢が研究開発の根底にある。エージェントモデルを用いたインタフェースは、電子秘書などへと展開するコンセプトであり、今後活発に研究開発される技術となっている。

参考文献

[1]望月、岸野：第８回HIシンポジウム予稿集、pp.13-16 (1992)

[2]Nigay L., Coutaz J.：INTERCHI'93、pp.172-178 (1993)

[3]安藤、北原、畑岡：電子情報通信学会論文誌、Vol.J77-DII (1994.8)

[4]安藤、菊池、畑岡：第10回HIシンポジウム予稿集、pp.589-594 (1994.10)

[5]Nigay L., Coutaz J.：CHI'95、pp.98-105 (1995.5)

[6]菊池、安藤、畑岡：第10回HIシンポジウム予稿集、pp.547-554 (1994.10)

[7]Maes P.: FRIEND21 International Symposium (1994)

[8]安藤、菊池、畑岡：音響学会春季講演会予稿集、3-P-15、pp.191-192 (1995.3)

[9]伊藤、他：音声言語情報処理研究会資料、No.5、pp.31-38 (1995.2)

[10]新田、他：音声言語情報処理研究会資料、No.1、pp.31-38 (1994.10)

対話における Non-categorical チャネル幅の拡大とリアルタイムの入出力融合

岡 隆一、木山 次郎、伊藤 慶明

新情報処理開発機構 つくば研究センタ

oka@trc.rwcp.or.jp

1. はじめに

機械翻訳の研究で有力な方式に Memory-based Translation (MBT) というものがある。これは、二言語間で同じ意味をもつ例文対を沢山記憶しておき、入力文に近いものを検索して、他方の対の部分を一部変更して出力（翻訳結果）するという方式である。いま、これとマルチモーダル・インタラクションとの間でアナロジーをとるとする。入力（音声やジェスチャ動画像など）を MBT の一つの言語表現に対応させ、出力（合成音声や CG など）を他の言語表現に対応させるとすると、「入出力の対」がつくれる。もちろん、ここにおける「入出力の対」は MBT のように対をなすものが同じ意味をもつものではなく、「入力」に対する応答で「出力」が定まる。

さて、MBT では翻訳が双方向であることから、例文対の表現はそれぞれ対等の意味をもつ。一方、マルチモーダル・インタラクションにおけるインタラクションの双方向を考慮すると、「入出力の対」のそれぞれは対等の扱いが自然である。しかし、マルチモーダル・インタラクションのモデルでは従来「出力」は「入力」に比べて低く扱われていたのではなかろうか？ 入力の理解には多くの努力がなされてきたが、その努力に見合うものがこれまで出力の中に十分生かされているかについては疑問がある。このことを別の側面からいうと、出力に必要な入力の部分は何であるかを決定するメカニズムの研究があまりなされていない、ということである。もちろん、入力をよく理解しないと出力は決められないという反論は当然ある。しかし、MBT は自然言語を理解しないと翻訳はうまくいかないという立場に対峙しているが、実際はその方法が他のものよりうまくいく場合が多い。

マルチモーダル・インタラクション・システムのモデル化に関する研究で、出力にドライブされるシステムの設計法がもっと多く考案され、そのシステムの実用面での評価、またそのときのユーザが作り出す入力自体の評価、制御などについてもっと探求されていいと思う。

2. 出力パターンにドライブされるシステムのモデル化

前章で述べた「入出力の対」に加えて、入力から出力へと変換するシステムを新たに想定し、三者の関係をモデル化することを考える。

最も簡単にモデルを作成するには、時系列を入出力とする Back propagation の学習則をもつ Jordan network を使うことであろう。Jordan network を用いて上述の内容をモデル化すると図 1 のようになる。図 1 では Back propagation で学習を行なうが、ここで「入出力の対」の出力側が教師信号となる。その意味で「出力」にドライブされるマルチモーダルインタラクション・モデルの構成となる。また、このシステムにおいて以下の点に留意する。

- リアルタイムにカップリングした「入出力の対」による学習の実行
- 「入出力の対」の構成における（認識）粒度自由の単位の使用
- 「入出力の対」のパターン表現（特に、スポッティングによる入力パターンの抽出）の使用

上記 1 番目の学習時に使用するデータ作成は、まず、人間の対話者の片方の音声やジェスチャ動画像や画像などによる入力データ収集と、それ対する他方の音声、ジェスチャ、画像などの出力データの収集で行なう [6]。次に、収集した入出力データとは独立の音声やジェスチャの標準パターンを作成し、入出力データへ先に提案した Reference Interval-free 連続 DP（RIFCDP）[2][4] などを適用し、実際に学習に使用するデータを作成する。また、2 番目の認識単位の粒度自由の意味とスポッティングとの関係については文献 [4] で述べている。図 1 のネットワークの結合係数は real-time の入出力データで直接的に構成されることにより、出力データ（教師データ）のもつ役割が極めて大きくなる。これにより、入出力の対をなすものが対等に扱われる。さらに、入出

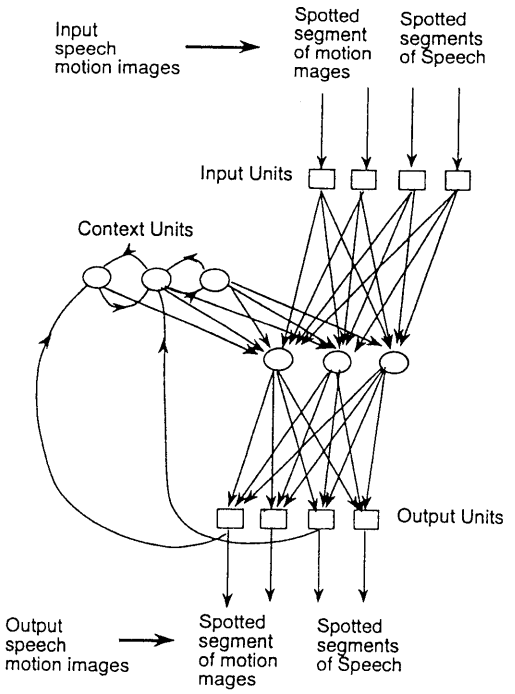

Figure 1: A learning procedure drived by real-time input/output patterns using a Jordan network for constructing a multi-modal interaction system.

力データがパターンであるため、システムの記述に、記号としての表現を用いることは極めて少なくなり、マルチメディアの融合が図り易くなる。このことは、入出力間のチャネルを non - categorical とし、かつ、入出力の対応が実時間で密接に結び付くことよりチャネルの幅の拡大を必然的に伴う。チャネル幅の拡大は、現在「入出力の対」データがパターン表現であるので、Back propagation のような教師つき学習によって行なうのが得策である。

3. リアルタイム I／O の臨場感とユーザの Error Tolerance

ここではリアルタイムの強いカップリングをもつ入出力データに基づくインタフェースの設計を主張する。この設計の考え方はまた入出力間のチャネル幅を広くする。例えば図 1 での出力は音声や動画像となる。チャネル幅の広がりとは、計算機の内部状態をユーザへよく知らせることであり、それは計算機がなぜその応答をするかの（明示的ではないかもしれないが）説明機能となっている。すなわち、計算機側がユーザに応答するとき、それ以外に多様な可能性をどう保持しているいるかを、応答以前の形で示している。応答候補の単なる羅列ではなく、低い優先順位の応答候補と高い順位の応答候補が絡みあっている状況も示せるのである。これらの計算機の状態の開示と（必ずしも明示的でないものであっても）説明機能は、ユーザに次の反応をするための考える材料を与え、ユーザと計算機とで共有している時間がユーザにとって無駄でないという感覚を結果的に与えることになる。

このような状況の出現は、計算機とユーザとの間でのある種の一体感の形成であり、計算機へのユーザの親近感の生成でもある。これらの感覚は 2. で述べた出力にドライブされたシステム設計と合わせると、計算機の認識エラーに基づく irregular な動作についてユーザが寛容であるという状況をつくれる。

文献 [5] で示されている対話システムは音声とジェスチャの同時使用ができ、これらの入力に応答する Graphics や合成音声の出力がリアルタイムで動作している。ここでいうリアルタイムとは、単に応答が早いという意味ではなく、音声や動画像の分析フレーム (8 msec or 30 msec) ごとにシステムの最終結果を出力すること、を意味する。文献 [5] のシステムは図 1 のような Jordan network を使用していなく、（モデル学習機能を内蔵しない）別のアーキテクチャである「連続オートマトン」[1][4] とよぶものでリアルタイム I／O の機能が実現されている。一般にリアルタイムをそこまで徹底することは常識として定着しているわけではない。ユーザをあまり待たせない程度に早く応答すればいい、というのが多くの設計思想であるだろう。しかし、上記のリアルタイムの I／O アーキテクチャに基づくとき、上述した意味のある種の臨場感をユーザに与えるシステムが容易に設計できるのである。

参考文献

[1] 岡、伊藤、木山、張:“概念スポッティングのための画像オートマトン”、音講論, 3-4-12 (1995-03).

[2] 伊藤、木山、小島、関、岡:“標準パターンの任意区間によるスポッティングのための Reference Interval-free 連続 DP（RIFCDP）”、信学会、音声研究会、(1995-06).

[3] 木山, 伊藤, 岡:“Incremental Reference Interval-free 連続 DP を用いた任意話題音声の要約”, 信学会, 音声研究会, (1995-06).

[4] 岡、木山、伊藤:“認識単位の粒度自由・並列アーキテクチャとその実現のための Reference Interval-free 連続 DP”、情報処理学会、音声言語情報処理研究会、(1995-07).

[5] 伊藤、木山、関、小島、張、岡:“同時複数話者の音声会話およびジェスチャの統合理解による Novel Interface System”、情報処理学会、音声言語情報処理研究会、(1995-07).

[6] 木山、伊藤、岡:“知識自動獲得を目的としたマルチモーダル対話データ収集”, 音講論, 3-P-16 (1995-03).

音声認識によるマルチモーダルインタラクションへの視点

嵯峨山 茂樹
NTT ヒューマンインタフェース研究所
E-mail: saga@nttspch.hil.ntt.jp

1 はじめに

「マルチモーダル」という語は、人によって多少異なった意味で用いられているが、いずれにしても、音声認識はその中の重要な位置を占めるだろう。この考え方は、古典的な音声認識の応用の考え方、すなわちキーボードの置き換えであるとか、電話を通したデータ入力などの、いわばシングルモーダルの入力に対して、新しい可能性を与えるものである。そのキーワードは複数入力間の協調(synergy)であろう。本稿では、音声認識をより有効に利用する意味でのマルチモーダルインタラクションと、音声認識への将来の期待としてのマルチモーダルインタラクションについて考えたい。

2 音声認識の利点

音声認識の利点を以下のように整理しておこう。

他に手段がない場合: 手が塞がっている場合などに音声認識が有効な情報入力手段であることは指摘されてきた。電話回線からの入力は今後も重要。さらに今後は、特に車載機器への入出力 (運転中は目と手が塞がっている) が大きな市場となろう。その後、福祉・高齢者関係が重要。や、手が塞がっている場合 (作業中、運転中)、身体障害の場合、などはいままでにも考えられてきた。この辺が当面ターゲットになる分野だろう。

快適さ: 音声認識の効用として、使いにくいキーボードを使わないで済む、という点が長く考えられてきた。

他入力手段との協調要素として: 音声入力と他の入力モードとの協調動作(synergy)は重要な面である。キーボード入力以上に効率を上げるためには、「キーボードの代わりに音声認識」ではなく、「キーボードとともに音声認識」を使うことになろう。「使い易い」という動機は二次的である。むしろ、「手も声も使え」になるかもしれない。効率の面から積極的に導入する理由になる。

効率: 発話はタイピングより速度が速いと言われてきた。しかし、今後はマルチモーダルの意味での効率に重点。

サイズ、重量: コンピュータはますます小型・軽量化し、(1) キーボードつきポータブルコンピュータ (ノートパソコン) → (2) 携帯端末 (手書き文字認識入力) → (3) 音声入出力コンピュータ ... と進むだろう。究極には (3) と携帯電話 (あるいはPHS) が融合するだろう [1]。究極的に情報処理機械を小型にするには音声認識合成によるしかない。

コスト: 究極のコンピュータは、CPU, マイクロフォン、イヤフォン、電池 だけからなる、極めて安価なものになるだろう。キーボードつきは高級品。[1].

面白さ: 音声認識は面白い。機械が自分の声でコントロールできる。自分で声を出すからすっかりその気になって没入するようなゲーム機や、バーチュアルリアリティなどが考えられる。バーチュアルワールドのエージェントたちと会話をするには、音声認識・合成は不可欠。

3 マルチモーダル入力要素としての音声認識

3.1 多手段の混在・協調の形態

「マルチモーダル」という語が多義的に理解されている。いくつかの入力モードが併存する ("OR" logic 的) ものと、いくつかの入力モードが協調する ("AND" logic 的) もの、さらに、いくつかの入力モードを順次用いるもの、がある。

音声認識 / マウスによる制御: 例: 自由発話大語彙住所音声の認識 (番号案内タスク)[3]

音声認識 / キーボード / マウスポインタの混在: たとえば、住所入力タスク [5]。画面上の情報入力スロットは、キーボードでもプルダウンメニュー選択でも音声認識でも同等に使える。

音声認識 / マウスカーソルの分担協調: マウスポインタを頻繁に移動しがちな作図ツールやCAD 作業では、マウスポインタは描画に専念し、線の太さや種類、図形の種類、などの属性指定やコマンドなどは、音声で行なう [4][6]。協調により入力速度が向上する。

音声認識 / 読唇の協調: 音声以外の情報を協調的に用いる [7]。

音声認識 / キーボードの協調: 例えば、カナ漢字変換において、カナをキーボード入力し、漢字の別の読みなどを音声で入力。協調により入力精度が向上させられるだろう。

音声認識 / 手書き文字認識の協調: 文字認識との協調: 手書き入力し、その読みなどを音声で入力。どちらも認識性能は100% でないが、協調により入力精度が向上するだろう。

4 マルチモーダルシステムの課題

筆者が考える、マルチモーダルシステムにおける今後の課題を述べよう。

認識誤りの扱い: 音声認識では避けられない認識誤りをどのように扱うか。キーボードなら速い操作で backspace (あるいは delete) キーを押すことができる。"音声 backspace" はどうすれば実現できるか。マルチモダリティはこの点で有用かも知れない (実際 "backspace" をタイプすればよい)。

複数候補の扱い: 音声認識では認識候補を複数生成することができる (手法合がある)。これをマルチモーダル入力でどのように結び付けるか。カナ漢字変換のキー操作を真似て「音声・文字変換」においても、スペースキーを叩くと次候補に変わり、リターンキーで確定する [5]、というような、すでに社会的に認知された方式を取ると慣れやすい。

語彙制約なし、任意の音声・文字変換の技術: 語彙制約のない任意語を認識する要望は強い。音声認識を行なうためには、それ以前に認識語彙のリストを定義する必要があり、そのためにキーボードから多くの語彙を入力せねばならないのでは、何のために音声認識をするのかわからない、という意見もあろう。このような事前語彙定義を必要としない技術は現実的である。音声だけで行なうのはかなり難しいが、マルチモーダル部品としてならば、ペンやカーソルで修正などを通して、使えるようになる可能性がある。

マウス、キーボード、カーソル、タブレットとの協調: マルチモーダル入力は、まだ社会的に認知された一般形式を持っていない。キーボード配列のように、歴史的な理由で受け入れられているようなものに比べて、まだ確立しているものがない。標準化に近い意識で、一般ユーザに浸透して行くようなスタイルの追求が必要。

応用開発インタフェースが重要: 大抵の場合、ソフトウェアからの呼び出しの形で音声認識を含むマルチモーダルサブシステムを使うことを考えると、これをソフトウェア部品として使うための、いわゆる API (Application Programming Interface) は非常に重要である。標準化を考える時が来るだろう。

音声認識に関わる問題点 [2] の解決: 音声認識について指摘されている問題点は、音声認識を含むマルチモーダルシステムにおいてもそのまま問題点であったり、マルチモダリティによって解決されたりする。重要に思われる問題点は、

- ヒューマンインタフェースが未熟だから使いにくい。
- うまい対話制御が重要。これがうまくいっていないから使われない。
- 音声認識の使い方の社会的コンセンサスがまだない。
- 人は機械に向かって話すのには抵抗がある。音声認識技術は本質的に嫌われる。
- 音声認識誤りがどのように起こるのか、どう発声すれば避けられるのか分からない。この不透明感がユーザにとって最も辛い。
- キーボードなどの他手段に比べて入力効率が決して良くない
- 音声認識を使う人間を訓練する必要がある。キーボードだって一日で使えるようにはならない。
- メンタルモデル (システムがユーザにどのように見えているか) とシステムイメージのギャップを埋める手段が確立していない。

エージェント: エージェント型マルチモーダルインタフェース [8] が提案されている。この考え方は、特にインターネットなどの通信と融合して重要であろう。例えばネットワークショッピングで、百貨店を呼ぶと、販売員エージェントが派遣され、音声合成と音声認識を介して顧客に商品の説明をして注文を取る。人間自体がマルチモーダルなので、コンピュータ側のマルチモダリティは人間を模擬する (エージェント) ことが究極の形である。

ヴァーチュアルリアリティにおける音声認識: ヴァーチュアルリアリティの世界 (例: NTT の「インタースペース」) などでは、ヴァーチュアルな世界の生き物が喋り、人間の声を理解するのが自然であろう。そうすると、音声認識と音声合成の必要性は必然的になる。このような世界の登場人物や生物が、マルチモーダルである。触っても、話し掛けても反応する。これらはその世界で動き、話す。

参考文献

[1] 嵯峨山茂樹: "音声認識," 日経バイト 7 月号 (100 号記念特集号), 日経マグロウヒル社, pp. 212–221 (1992).

[2] 嵯峨山 茂樹: "なぜ音声認識は使われないか・どうすれば使われるか?" 情報処理学会研究報告, 94-SLP-1, Vol. 94, No. 40, pp. 23-30, 1994.05.

[3] 吉岡 理、南 泰浩、山田 智一、鹿野 清宏: "電話番号案内を対象としたマルチモーダル対話システムの作成," 音学講論, 1-8-19, pp. 37–38, 1993.10.

[4] 西本、志田、山岡、小林、白井: "音声・マウス・キーボードを用いたマルチモーダル作図環境," 音学講論, 1-7-21, pp. 41–42, 1994.04.

[5] 荒井 和博, 吉岡 理, 嵯峨山茂樹, 山田 智一, 野田 喜昭, 井本 貴之, 管村昇: "音声認識機能を持つ住所入力システム," 電子情報通信学会 1995 年総合大会講演論文集, SD-9-7, 情報・システム 1, pp. 379–380, 1995.03.

[6] 井本 貴之, 山田 智一, 嵯峨山茂樹: "音声認識サーバを用いたマルチモーダル入力," 電子情報通信学会 1995 年総合大会講演論文集, SD-9-6, 情報・システム 1, pp. 377–378, 1995.03.

[7] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, A. Waibel: "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition," Proc. ICASSP 95, pp. 109–112, 1995.

[8] 伊藤克亘、長谷川修、栗田多喜夫、速水悟、田中和世、山本和彦、大津展之: "音声・視覚・画像をもつインタラクションシステム," 情報処理学会研究報告 95-SLP-5、pp. 31–38, 1995.

音声言語情報処理 7−16−4
（1995. 7. 21）

マルチモーダルインタラクション
—ユーザ中心の新しいインターフェイスの視点から

安村 通晃*

慶應義塾大学 環境情報学部

1 はじめに

コンピュータと人とのインタラクションでは、現在、最も盛んなのは、GUIに代表されるビジュアルインターフェイスである。GUIが最初にデザインされて、ほぼ1/4世紀が経過した。ポストGUIとして、マルチモーダルインタラクションに大きな期待が寄せられつつある。

ここでは、ユーザ中心の立場から現状のビジュアルインターフェイスの問題点を簡単に述べ、さらに、新しいインターフェイスを模索する意味で、マルチモーダルインタラクションにどのような点からアプローチすれば良いかについて、述べたいと思う。

2 GUIの利点と問題点

GUIに代表されるビジュアルインターフェイスでは、

- 今まで隠されていた情報、データ等が見えるようになる(可視性)
- 処理するコマンドなどを記憶再生する代わりに、メニューに現れた中から再認することにより記憶の負荷が減る(記憶の外在化)
- 操作対象物に相当するものが画面に表示され、それを直接操作できる(直接操作)

といったメリットがある。しかし、現状のビジュアルインターフェイスでは、

- キーボードとマウスによる入力とウィンドウからの表示だけであり、視覚に極度に偏重したインターフェイスである。
- コンピュータ側は、わずかに視覚的な表示をするのみで、極めて受動的なインターフェイスである。
- 仮想世界の表示のみが行なわれ、現実世界との関連性が薄い。

といった問題点がある。

3 マルチモーダルシステム

これに対し、マルチモーダルインタラクションにおいては、以下の点を活かしたインターフェイスを用意する。

- 身振りや音声など、視覚以外のモダリティも用いる。
- コンピュータ側に目や耳、口といった能動的な機能をもたせる。
- 従来の文字情報だけでは、表わし切れない感情や気持ちといった二次情報の伝達も可能にする。
- 複数のモダリティの複合化、連動化を行なう。
- モダリティ間の変換を支援する。

感覚器でいえば、従来は、手の動き(キーボードとマウス)以外は、視覚しか用いなかったが、マルチモーダルインターフェイスにおいては、音声対話と画像認識を意識的に用いる。人の体や手や腕などの身体運動もコミュニケーションの手段とする。

*e-mail: yasumura@sfc.keio.ac.jp

4 研究のアプローチ

これまでも、マルチモーダルに関連した研究の試みは少しずつなされてはいるが、まだ十分に広がりを見せているとはいい難い。より一層、マルチモーダルの研究を推進するためには、次のような点に注意して研究を行なう必要がある。

1. 技術志向から人間志向・タスク志向へ
2. 完全主義から共生・共存主義へ
3. 機械と人間の関係に関する認知的理解
4. 遊びと柔軟な発想

すなわち、従来のマルチモーダルの研究においては、たとえば、音声認識を用いた研究においては、音声認識の精度を向上させることの絶対的必要性が力説されるといった、技術志向が極めて強かった。音声認識の精度を上げる努力は必要ではあるが、もっと重要なことは、音声をどのように利用するか、どういう場面で用いるか、といった点である。その意味で人間志向、タスク志向へと研究の方向性を切替える必要がある。また、これと関連して、従来は、ともすれば、ある技術ならその技術だけを徹底的に追求する、といった完全主義が見られた。人間のため、あるタスクのためを考えれば、特定の技術に固執することよりも、複数の技術や方法論を必要に応じて使い分けた方が良いことが少なくない。

さらに、機械に対する人間の関係というものは必ずしも明らかになっていない。認知科学的、あるいは、社会科学的な分析も必要であろう。最後に、このような新しいインターフェイスの研究においては、遊び心や柔らかい発想というものが必要である。ガリガリに目的を絞った研究よりも、遊び半分での研究から創発性が生じる可能性もある。

5 今後の課題と方向性

マルチモーダルインターフェイスは、それ単独でも研究の意味は十分ある。しかしながら、マルチモーダルがこれから大きく伸びてゆくのは、その単独の技術というよりも、他の関連技術や関連分野との関係においてである。その意味で、これから、マルチモーダルインターフェイスの研究において、次の研究パラダイムとの関連が重要になってくる。

- だれでもが使える (障害者や高齢者にも)
- どこでも使える (ユビキタス性)
- 代理人として使える (エージェント性)
- 離れていても使える (ネットワーク性)
- 現実世界に意味をもつ (双対空間性)

6 おわりに

マルチモーダルインタラクションが、本当にGUIに代わる次世代インターフェイスになり得るかは、だれか偉い人の予言を待つような話ではなく、マルチモーダルインタラクションに興味をもつあらゆる研究者が自ら新しいインターフェイスを構築していくことしかないのは確かである。

参考文献

1. ビジュアルインタフェース調査研究 WG, ビジュアルインタフェースの研究開発報告書, 日本情報処理開発協会, 1995.
2. 安村通晃, 今野 潤, 八木正紀, マルチモーダルプラットフォーム MAI の構築に向けて, 日本ソフトウェア科学会 WISS'94, 1994.
3. 安村通晃, 伊賀聡一郎, マルチメディアからマルチモーダルへ, 日本ソフトウェア科学会 WISS'93, 1993.
4. Laurel, B., Ed, 人間のためのコンピューター, アジソンウェズレイ, 1994.
5. Wellner, P., et.al. Eds, Computer-Augmented Environments: Back to the Real World, *Comm. ACM*, Vol.36, No.7, 1993.

音声言語情報処理 7－16－5
（1995．7．21）

コミュニケーションにおけるマルチモーダルインタラクション

中津 良平
nakatsu@mic.atr.co.jp
ＡＴＲ知能映像通信研究所
〒619-02 京都府相楽郡精華町光台2-2

1．はじめに

我々は映像・音を中心としたマルチメディアを駆使した新しいコミュニケーション・通信の方式を開発することをめざしている。この立場からマルチモーダルインタラクションについて考えて見たい。コミュニケーションの立場からすると、マルチモーダルインタラクションすなわちコミュニケーションであるということができる。そこで、ここでは原点に立ち返って、コミュニケーションは本来どうあるべきかを考える。それによって、マルチモーダルインタラクションのあるべき姿と今後の研究方向が見えてくると思われる。

2．種々の観点からのコミュニケーションの考察

2．1 コミュニケーションと通信

本来、コミュニケーションは、人間同士が向かい合って、音声・身振り・手振りなどを用いて自分の感情・意思を相手に伝える全感覚的なものであった。すなわちマルチモーダルなインタラクションであったといえる。これに対し通信は、対面型のコミュニケーションの持つ距離・時間に関する制限を克服しようという立場で技術開発が進められてきた。印刷術の発明と電話の発明は通信の歴史における革命的事件であり、これによって、空間・時間の克服がある意味で実現できた。しかしながら、反面こられの技術はコミュニケーションに大きな制限を課することになった。すなわち、印刷・電話の導入は本来の全感覚的なコミュニケーションを文字もしくは音声という単一の言語メディアを用いた言語中心のコミュニケーションに制限してしまったのである。最近、マルチメディア通信が叫ばれながら革新的なサービスが現われていないのは、従来の言語メディアのみを用いた通信のスタイルに我々が大きく影響を受けているからとも考えられる。

2．2 マルチモダリティ

従来から、人間の音声の中に含まれる意味情報を自動的に認識しようとする研究や、人間の表情から感情を読み取ろうとする研究などが進められてきた。これは個々のメディアからすべての情報を取り出そうとする立場である。しかしながら本来のコミュニケーションでは、表情・身振り・手振り・対話者のおかれた環境を有効に利用し、さらにはこれらを組み合わせることにより、意思・感情を伝達し相互理解を図ろうとしている。これは、１つのメディアにすべての情報をのせるのではなく、各メディアに各々が得意とする情報を少しずつ乗せ、受け手側はこれらを統合することにより総合的な情報の送受を行っているのである。このようなモダリティ統合の立場が従来の研究では不十分であった。

2．3 インタラクション

従来、コミュニケーションにおいては、情報が送り手から受け手に一方向的に送られるというシャノン流の考え方が暗黙の前提となっていた。インタラクションを扱う際も、この暗黙の前提の元で研究が進められてきたきらいがあった。しかしながら、コミュニケーションにおけるインタラクションは以下のような特徴を持っており、その観点から今後のインタラクションの取り扱いを進めるべきであろう。

（1）送り手から受け手への一方向的な情報の流れがあり、単に流れの方向が順次切り替わることがインタラクションであるという考え方は不十分である。むしろ、コミュニケーションの参加者およびそれら間の情報の流れが渾然一体となった系全体の動作をインタラクションとして取り扱う必要がある。

（2）さらには、コミュニケーションの参加者間のインタラクションのみならず、個々の参加者はその環境とも常にインタラクションを行なっていると考えられる。したがって、このようにコミュニケーションの参

加者およびその環境が一体となった系の動作全体をインタラクションとして扱う必要がある。

3．今後のコミュニケーション研究

3．1　コミュニケーションのモデル

従来はＡＩの分野を中心として主として言語によるコミュニケーションを扱ってきた。これに対し図に新しいコミュニケーションのモデルを示す。言語を用いたコミュニケーション機能の内側には、相手との意思疎通を図るための基本的な動作を行なうべき機能（これをインタラクション層と呼ぶことにする）があると考えられる。例えば仲間の鳴き声に対し鳴き返したり、人間が相手の話しに相槌を打ったりする機能がこれに当たると考えられる。さらにその内側に、本来動物が持っている外界の刺激に反射的に反応する機能を持った部分（反射層）があると考えられる。このような複数の層の機能の複合作用の結果として我々のコミュニケーションの総体機能が創発してくると考えるのが自然である。マルチモダリティ、インタラクションもこのようなモデルの上で考えるべきであろう。

3．2　コミュニケーション研究の進め方

上記のような観点からコミュニケーション研究においては以下のようなアプローチが重要である。

（1）言語を用いたコミュニケーションの他に、インタラクションや反射運動といったより基本的なレベルでの機能を取り扱う必要がある。

（2）個々の機能の詳細な研究以外に、モダリティ統合、環境と一体となったインタラクションなど総合的な機能の取り扱いが必要となる。

4．おわりに

コミュニケーションの観点からマルチモーダルインタラクションを考察した。その結果、マルチモーダルインタラクションはコミュニケーションそのものであること、また従来これらの分野がコミュニケーションの一面しか取り扱っていないことを示した。これらの考察の上に立って、反射層、インタラクション層などの原始的もしくは動物的な機能の上に言語をあやつる層がのった創発モデルで表されるコミュニケーションのモデルを提案した。このモデルに基づくと、マルモーダルインタラクションを考える際、単に言語によるコミュニケーションのレベルでのみこれらを考えるのではなく、より基本的な層におけるそれらの機能を明確にし工学モデルを実現すると共に、それらの上に言語レベルのマルチモーダルインタラクションの機能を乗せるという立場での研究が重要であるといえる。もちろんこれらをすべて実現することはかなりの時間をかける必要がありすぐに実現できるとは考えられない。また、我々の日常のコミュニケーションが言語を極めて重視したものになっていることも事実である。したがって言語に主眼をおいてマルチモーダルインタラクションにアプローチすることは間違っているわけではないが、常に上記のような観点を念頭に置くこと、また、コミュニケーションの基本層におけるマルチモーダルインタラクションの研究を並行して進めることも極めて重要であることを指摘しておきたい。

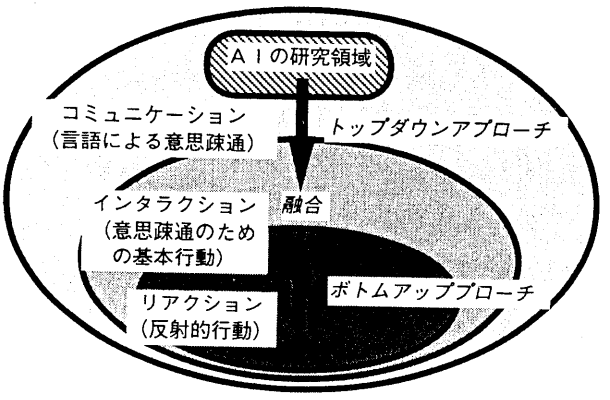

図　コミュニケーションのモデルと
アプローチ