情報メディア　5－2
（1992. 3. 2）

音声と映像コマンドを統合したマルチメディアインタフェース

間瀬健二　　福本雅朗　　末永康仁

NTT ヒューマンインタフェース研究所

概要

優れたインタフェース実現のために，言語に加え，顔表情，ジェスチャ等の非言語メディアの利用が有効であるという概念は多く発表されているが，実際に動くシステムを組んで実験を行なっている例は少ない．本文では，コンピュータの知的インタフェースをいかに作るかという問題に対して、通訳者のメタファを使って機能を整理する。さらに人物像の認識と音声認識を利用したマルチメディアインタフェース “Human Reader” の概念を紹介するとともに，既存のコンピュータを利用してその一部の実現を図った例を報告する。

A Multi-media Interface Based on Human Images and Speech

Kenji Mase, Masaaki Fukumoto and Yasuhito Suenaga

NTT Human Interface Laboratories

1-2356 Take, Yokosuka, Kanagawa 238-03, Japan

Abstract

Many conceptual works have been already reported that non-verbal media, such as facial expressions and gestures, are important as well as verbal language for better man machine interfaces. However, only a few system have been developed to display its effiencies. In this paper we introduce a framework of multi-media interface, named “Human Reader”, and an experimental system implemented with conventional computers.

本論文の１部は情報処理学会第９回グラフィクスとCAD シンポジウム (1991.11) で発表した内容である。インタフェースを中心に構成を変更しまとめた。

1　まえがき

キーボードやマウスは，現代のコンピュータ利用における標準的なインタフェースデバイスとなっている．通常の仕事をするうえでは大変便利であり，今後よほどの技術革新がなされない限り，当分その優位は続くであろう。しかし，人間の立場から見た場合，生活のあらゆる場面において上記の形態のみで充分であるということはない．人間は移動しつつ行動し，作業が続けば疲労し，誤りを犯し，飽きるからである．使いやすいインタフェースにするには、デバイスを自然な形態にして、インタフェース自身が少し知的になる必要がある。

本文では、コンピュータの知的インタフェースをいかに作るかという問題を整理して、その1例として手振りの映像と音声を使ったコマンドでコンピュータと対話するシステムの実現例を紹介する。すなわち、インタフェースをコンピュータと人間の間をとりもつ言語通訳者としてとらえてその役割を考察する。またインタフェースが扱うデータをメッセージとしてとらえ、メッセージの種類を調べる。また、人物像の認識と音声認識を利用したマルチメディアインタフェース“ヒューマンリーダ”の枠組みを紹介するとともに，既存のコンピュータ技術を利用してその一部の実現を図った例を報告する。具体的には“ヒューマンリーダ”のサブシステムである、“ヘッドリーダ”,“ハンドリーダ”,“ボイスリスナ”を結合した知的インタフェースの例を紹介する。

2　知的なインタフェース

人間と人間のコミュニケーションがうまくいかないことを、我々はしばしば体験する。特に、言語や文化が異なる人とのコミュニケーションは難しい。そのような人同士がコミュニケーションをするためには、共通の言葉をさがすか、通訳者に間に立ってもらうか、あるいはどちらかが相手の言語や文化を理解して同じ土俵に上がる必要がある。コンピュータと人間の関係もこれによく似ている。これまでのコンピュータと人間のコミュニケーションでは、人間の側がコンピュータに歩み寄るしか方策がないために、なれない言語や操作法を習得する必要があった。我々が目指す知的なインタフェースとは、人間とコンピュータのそれぞれ異なる言語と文化の両方を理解して、的確に通訳をしてくれるエージェントの介在を前提とする。

人間の通訳者の役割を考えると、知的インタフェースが所有すべき機能や性質が明確になると考えらる。すなわち、通訳者は目と耳と直感と相手に対する知識を総動員して一方の言葉を他方に翻訳している。主な点を列挙すると通訳者は、次のような作業を行なう。

1. 自然言語を理解して翻訳する
2. non-verbal な言語（動作、そぶり、表情、視線）などを理解して言葉を補う
3. 音、映像などいろいろな情報を統合する
4. わからないことを尋ねるなどして、情報を確かめる
5. 伝えた相手の理解を確かめる
6. 双方の背景となる知識を持ち、言葉と補強・修正する

これらと同じような機能をコンピュータと人間とのインタフェース・エージェントが有すれば、コミュニケーションがスムーズになると考えられる。例えば、入力デバイスからきた低レベルの信号をそのまま使うだけではいろいろな

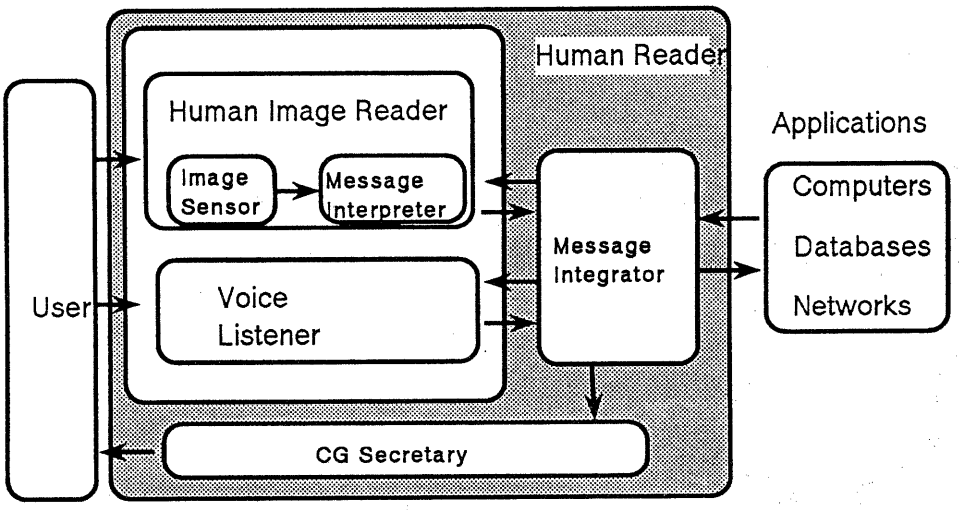

図 1: ヒューマンリーダ (Human Reader)

種類のメッセージを伝達するのに、多くの入力キーが必要となってしまう。また、ある信号は、状況によっていろいろなメッセージに解釈できることがある。このような複雑さをなくし、あるいは多義性のあるメッセージの曖昧さをなくすには、信号の解釈の仕方を、前後の状況や文脈によって何通りかにできるようにする必要がある。ここで、文脈解釈や知識ベースの利用が必要になる。人間がメッセージを解釈するように、必要な信号を的確に処理することが必要になる。また、メッセージに曖昧さがあったときには、自ら曖昧さを補間したり、相手に再度尋ねるような行動をおこすことも、知的インタフェースが備える機能である。また，システムが必要としているメッセージを分析してそれをユーザに要求するのも本来はインタフェースが備えている機能であろう.

この考え方に基づいて我々は “Human Reader” と呼ぶインタフェースの枠組みを提案している[1]。Human Reader は電磁センサやデータグローブのような特殊なデバイスを装着しない，より自然なインタフェースの実現を目指す枠組であり、コンピュータとの対話に限定されない，屋内，屋外すらも問わない，人間の様々な活動場面への適用を前提としている．特に本文では、視覚と聴覚を利用する Human Reader の具体例として、(1)Human Image Reader, (2) Voice Listener, (3) Message Integrator, (4) CG Secretary の 4 要素から構成される枠組について具体的システムを構成し、実験を行なう。

3 Human Reader の枠組み

ここでは，我々人間が発する様々な信号（図2を参照）を検出し，解釈してコンピュータで利用可能な形態に変換する Human Reader の枠組みを紹介する。Human Reader は、センサ部 (Image Sensor, Voice Sensor) と解釈部 (Message Interpreter) を有する視覚（画像，映像）情報認識用の枠組みを Human Image Reader と呼び，同じく聴覚（音声）情報認識用の枠組みを Voice Listener と呼ぶ．図1に示すように，Human Reader は，上記 2 つの枠組の他にメッセージ統合部 (Message Integrator)，および

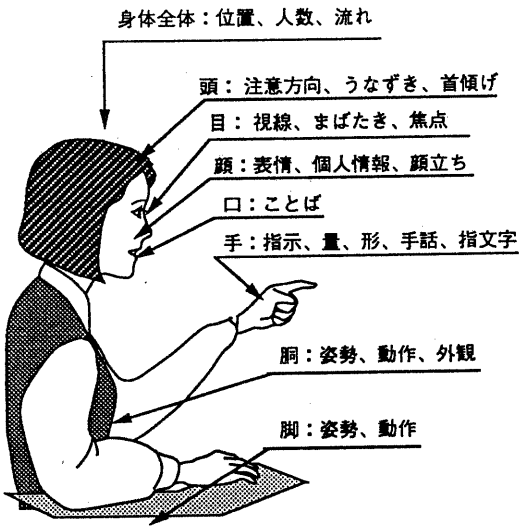

図 2: 人間が発するメッセージ

CG Secretary を構成要素としてもつ. Human Reader は知識の利用を前提とする総合的な知的インタフェースである. コンピュータのもつ知識が異なれば, たとえセンサ部から同じ出力を受けても, 解釈部や統合部では異なる出力を与える. これら特徴センサ部と解釈部の構成と機能はそれぞれの目的によって異なるが, これらは既存のグラフィカル入力デバイス (以下, 単にデバイスと呼ぶ) との対比で考えることができる.

以下では, Human Image Reader の説明のために, デバイスの分類を行なっておく. 聴覚情報認識用枠組である Voice Listener については, 現時点では既存の音声認識技術を利用することとし、ここでは説明を省く.

4 人間が送るメッセージの入力

グラフィカル入力デバイスはワークステーションやグラフィックディスプレイを操作する上で, 指示をしたり操作をする必要に応じて開発されてきたものであり, コンピュータとのインタラクションをする上で必要性が高いメッセージを運んでいる. したがって, 重要性の高いメッセージの種類を歴史の中から知ることが可能である. また, これらを代用できるより快適なデバイスが開発できれば, すぐに応用する分野が存在している.

一方, これらのグラフィカル入力デバイスで分類できないメディアについては 2 つの解釈ができる. すなわち, これらのメディアが運ぶメッセージはコンピュータインタフェースにおいては必要性が低いか, あるいは, 必要性は高いがこれまで技術的理由により実現が困難であったものである. 我々は後者についてよく吟味する必要がある.

現在用いられる, よく知られたデバイスと, そのメッセージ伝達機能, および, 同等のメッセージの人間からの発生源を整理すると表 1のようになる. このように、入力デバイスを一般化しておいて、本来の伝達手段としての音声と動作ではどのようにメッセージ伝達できるかを調べていく。

キーボード： カナやアルファベットなどの文字入力機能をもつデバイス. 人は“あ”と声をだしたり, 人差指をたてて指文字を作って“1”と示す.

セレクタ： 体系化されたカタログから目標を選ぶ機能を提供するデバイス. それぞれのカタログの項目 (アイテム) は別の意味をもつ機能や実体に対する識別子に割り当てられている. プルダウンメニューやリモコンのいろいろな機能ボタンが例としてあげられる. 人は“やま”という言葉で「山」, 顔写真で「 A さん」, 親指を立てて「男の人」を示すようなメッセージを送ることができる.

表 1: 情報入力用デバイス、機能と、人間のメッセージ発出メディア

デバイス	例	機能	人間からのメッセージ発出メディア
キーボード (keyboard)	キーボード	文字 (symbol) 入力	音声 (ことばの発声), 指文字, 他
セレクタ (selector)	ボタン, メニュー等	メニュー選択	ことば, 手話, 在不在, 個人特徴, 他
ロケータ (locator)	マウス, タブレット等	位置指定	指による指示, 視線, 頭部回転, 自分の位置, 他
バリュエータ (valuator)	ダイアル等	量的指示	手の回転, 両手による大きさの指示, 音声, 他
サンプラ (sampler)	TV カメラ,A/D 等	データの標本化	ものまね
イメージャ (imager)	特徴抽出器等	概念, 感性入力	声色, 手の形, 表情, 姿勢, 外観, 顔立ち, 他

ロケータ： 位置を指示するデバイス．マウスやタブレットで 2 次元の xy 座標を指定するように，人は指で方向や場所を特定する．また，視線や頭の向きからも方向を指定することがある．マウスの出力は基本的にロケータの機能であるが，他のデバイスの代用（セレクタやバリュエータなど）もできるような仮想デバイスが考案されている．

バリュエータ： ダイアルなど連続的な数値を入力するデバイス．ダイアルを回すようなしぐさで手を回転させたり，両手を広げて大きさを伝える動作がこれに含まれる．

以上の 4 種類が現存するグラフィカルデバイスから容易に類推できるのに対し、次の 2 種類はほかの非言語メッセージや実在する計算機インタフェースから当てはまるものを検討した結果考案したデバイスである。

サンプラ： パターンをあるがままに入力するデバイス。スキャナや TV カメラなどの入力装置からの類推である．人がこれに相当するメッセージを送る例は，少ない．あえて“ものまね”で，言葉や，振りを伝えることが分類できる．また、シーンや動作の記述などもこれに相当すると考える。

イメージャ： 概念や心象にかかわるメッセージの伝達デバイス。人は場所や個人の名前といったメッセージのほかに，感情や雰囲気あるいは形などといった，概念や心象に分類できるメッセージを伝達している．このようなメッセージを伝達できる独立した入力デバイスはいまのところ存在しない．ここでは，このような概念や心象にかかわるメッセージを伝えるデバイスを**イメージャ (imager)** と呼ぶことにする．たとえば，顔はロケータやセレクタにあたるメッセージより，イメージャに相当するメッセージを送るのが得意である．また，現存するデバイスがないため，映像で表現される顔（顔画像）をメディアとするイメージャの実現が必要と思われる．

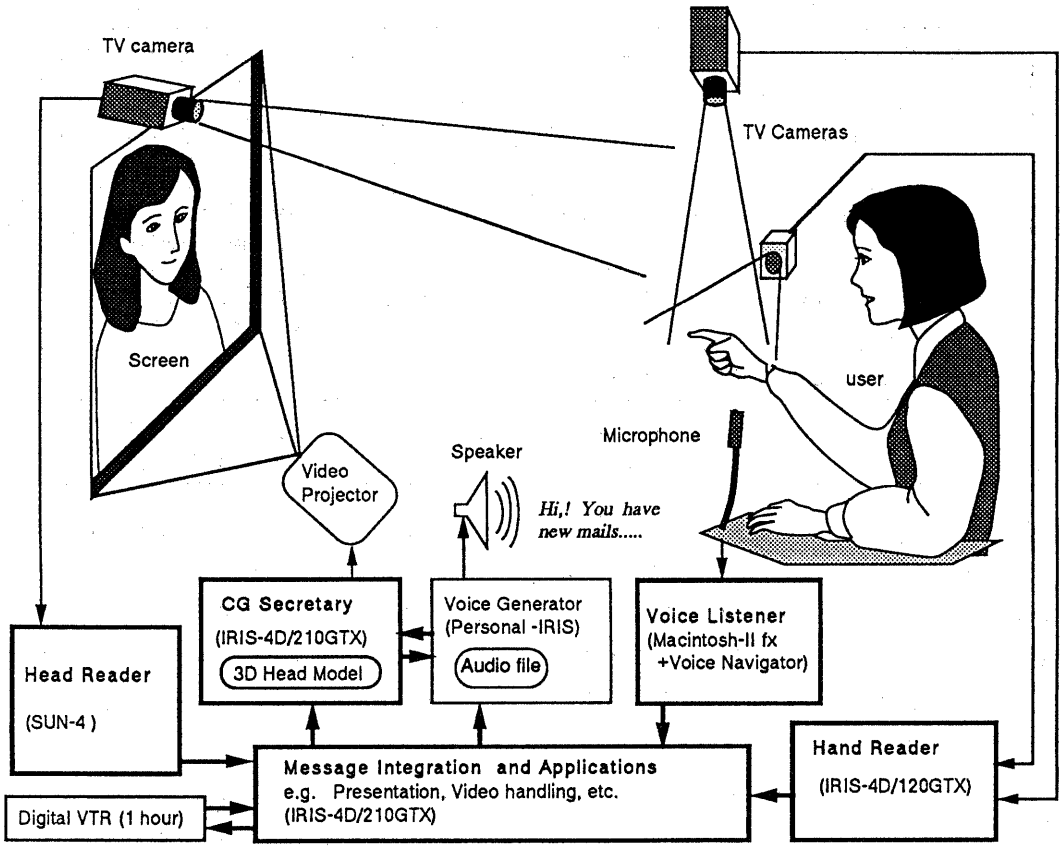

図 3: Human Reader の実験システム

さきに述べたように，キーボードとマウスが入力デバイスとして多用されている状況では，人間の動作を介したキーボードデバイスとロケータデバイスを実現することは，大きな意義がある．現存するキーボードとマウスに加えて，動作を介した入力デバイスを提供することで人間のメッセージ伝達手段の自由度が増し，それが，自然なインタフェースの実現につながる．手の場合を例にとると，ロケータデバイスは指先の一点の3次元座標を測定できれば，容易に実現できることがしめされている[2]．また，キーボードデバイスは指文字を認識することで可能である[3]．手話や動きを使ったセレクタデバイスやバリュエータデバイスを構築するには，動画像処理やパターン認識が必要となり難易度が増す．

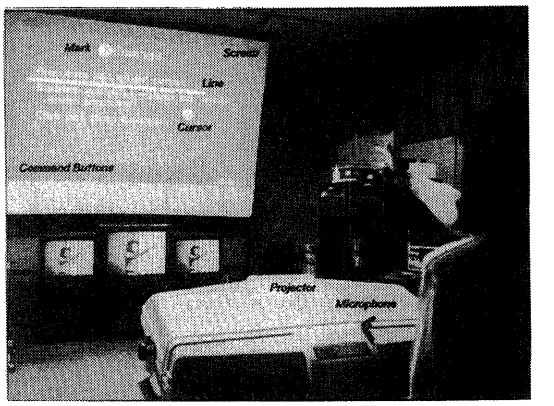

図4: プレゼンテーション実験システム (FP+)

5 実現例

図3はHuman Readerにおける各サブシステムのセンサの配置と実際のサブシステムの装置と接続状況を示している．Head Readerは前方のカメラを，Hand Readerは上と横のカメラをセンサとして，画像処理を別々のワークステーションで行なう．Voice Listenerはマイクの音声信号を単語音声認識装置で処理して登録キーワードの認識を行なう．なお、指向性マイクロホンを使用することにより、ヘッドセットの着用をさけ、非接触で自由なインタフェースを構築することを心がけている。

この実験システムにおいて、次に示すプレゼンテーション用アプリケーションについて実験プログラムを作成し、総合的なシステムとしての動作と効果の確認を行っている。

5.1 電子プレゼンテーションへのハンドリーダの応用

スクリーン上にスライドやビデオを表示してプレゼンテーションを行うアプリケーションである。Hand Reader, Voice Listener, Digital VTR, Message Integratorがプレゼンテーションを行うアプリケーションプログラムの中で結合して動作している。現在、マーカを打つ、下線を引く、ページをめくる、スクリーン上のメニュー選択によるVTRの操作などが可能である（図4を参照）。例えば、指でさしながら、「ここから」「ここまで」と発話することで、下線を引くことができる。使用しているVoice Listenerが音声入力から認識結果出力まで1秒弱かかるため、現在の統合方法では、キーワード認識が完了するまで指差しを持続しなければならないという問題はある。しかしHuman Image ReaderとVoice Listenerを組み合わせた何も装着しない快適なマルチモーダルインタフェースを実現している。

6 まとめ

本文では知的インタフェースを実現するシステムとしてヒューマンリーダの枠組みを紹介した.

感性に関するメッセージの応用としては,「すき／きらい」といった感情を伝える必要のある分野を考えることが必要であろう. 一例として、「困った／疲れた」というメッセージはシステムのインタフェースとしては, フェイルセーフ／フォールトトレラント／ガイダンスオリエントなシステムにとって重要であると考えられる.

文 献

(1) 末永康仁, 間瀬健二, 福本雅朗, 渡部保日児: Human Reader: 人物像と音声による知的インタフェース, 信学論, **J74-D-II** (1992-02).

(2) 福本雅朗, 間瀬健二, 末永康仁: 動作と音声を統合したマルチメディアインタフェース, 91 秋季信学全大 (1991).

(3) 福本雅朗, 間瀬健二, 末永康仁: 動画像処理による非接触ハンドリーダ, 第 7 回 HI シンポジウム (1991).