論 文

Technical Papers

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

An Interface Agent System Employing an ATMS-based Multimodal Input Interpretation Method

河野 恭之
Yasuyuki Kono
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
yasuyuki.kono@toshiba.co.jp

屋野 武秀
Takehide Yano
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
yano@krl.toshiba.co.jp

池田 朋男
Tomoo Ikeda
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
tommy@krl.toshiba.co.jp

知野 哲朗
Tetsuro Chino
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
chino@krl.toshiba.co.jp

鈴木 薫
Kaoru Suzuki
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
suzuki@krl.toshiba.co.jp

金澤 博史
Hiroshi Kanazawa
(株) 東芝 関西研究所
Kansai Research Labs., Toshiba Corp.
kanazawa@krl.toshiba.co.jp

Keywords: human interface, human-computer interaction, multimodal interface, multimodal input interpretation, reference resolution, interface agent, ATMS.

Summary

Two requirements should be met in order to develop a practical multimodal interface system, i.e., (1) integration of delayed arrival of data, and (2) elimination of ambiguity in recognition results of each modality. This paper presents an efficient and generic methodology for interpretation of multimodal input to satisfy these requirements. It is able to integrate delayed-arrival data well, and is able to efficiently interpret multimodal input that contains ambiguity by regarding the multimodal interpretation process as hypothetical reasoning and formalizing the control mechanism of interpretation on the basis of the ATMS (Assumption-based Truth Maintenance System). The proposed method is incorporated into an interface agent system that accepts multimodal input consisting of voice and direct indication gesture on a touch display. The system communicates to the user through the interface agent's 3D motion image with facial expressions, gesture, and synthesized voice.

1. ま え が き

計算機の能力の向上と情報化の進展により，従来のように人が人に接するのではなく計算機をその内部に持つ機械に人が接する機会が極めて多くなってきている．しかしながら現在計算機のヒューマンインタフェース (HI) の主流である GUI(Graphical User Interface) は，ユーザがその利用方法にある程度習熟することが必要であり，そのレベルに達していないユーザに対して逆にストレスを生じさせることになっている．ユーザが人に対するのと同じように計算機に接することのできるような自然でロバストな「人にやさしい」 HI が求められている．

人間は対面する人に対し，言葉や身振り，手振り，表情といった様々な伝達手段，すなわちモダリティを利用して意図を表現し，効率的に伝達している．この

ことから，人間のコミュニケーションは本質的にマルチモーダルであると言うことができる．計算機のHIにも，利用者が入力方法などを意識せずに自然に，すなわち人に対するのと同様に複数のモダリティを用いて計算機に意図を伝達でき，また，複数のモダリティを効果的に用いてわかりやすく自然な表現を利用者に提示できるマルチモーダルなものが求められている[Maybury94]．その意味で実用的なマルチモーダルインタフェース(MMIF)の構築はHI技術の発展に大きく寄与すると考えられ，Put-that-there[Bolt80]をはじめとして様々なMMIFが試作されてきた[Maybury93]．

一般にMMIFは，音声やポインティングデバイスによる直接指示といった個々のモダリティの認識モジュールから非同期に与えられる認識結果を受け入れて互いに関連付け，その統合・解釈を行い，利用者が様々なモダリティを通じて入力した個々には断片的な情報の集合から，全体として利用者が伝達を意図した内容を推定する．マルチモーダル入力の統合・解釈処理（MM入力統合・解釈）を行える実用的なシステムの構築のためには，(1)認識結果の曖昧性の取扱い，(2)遅着データの統合といった課題の解決が必要である．これらの課題を効率的に扱える汎用の枠組を構築するために，我々は一貫性管理機構ATMS[deKleer86a, deKleer86b]をMM入力統合・解釈に適用しその制御メカニズムを定式化した．

更に計算機システムがユーザに情報を伝達するためのコミュニケーションチャネルである出力においてもマルチモーダル化の検討を行った．そして音声とタッチジェスチャを組み合わせたマルチモーダル入力を用いてユーザが意図を伝達することができ，擬人化エージェントが自然言語合成音声だけでなく，表情や身振りといったノンバーバルメッセージを利用して効果的にシステムの状態を伝達することができるインタフェースエージェントシステムを試作した．本システムの構築と試用，及び展示を通して我々はマルチモーダルHIに関して様々な知見を得た．

2. マルチモーダルインタフェース

2•1 マルチモーダル秘書エージェントシステム

既に述べたように我々は，自然でかつ効率的に情報を授受することができるHI技術の確立を目指している．複数のモダリティを用いて情報を授受するMMIFと従来のHIを比較した場合，次のような項目がより自然なコミュニケーションを実現するためのポイントであると我々は考えている．

メディア統合　ユーザが様々なモードを使って伝達してきた情報を統合し，一つのメッセージとして解釈する．これにより人間が普段自然に行っているコミュニケーションに近いHIを可能にする．

メディア間補間　どのメディアについても全く誤りや曖昧さのない認識処理を実現することは困難であるが，相互に補間しあうことで各メディアの認識の不完全さを補う．

図1　マルチモーダル秘書エージェントシステムの画面例

メディア割り当て　授受する情報の種類，量や相手などによって適切なコミュニケーションメディアは異なるため，状況に応じてメディアを切替えることでコミュニケーションの質を向上させる．

ノンバーバルメッセージ処理　人間は他人とコミュニケーションを行う際，言語で表現可能な情報だけでなく，身振り，表情など文字として記述できない種類の多数のノンバーバルメッセージを交換している[*1]．人間と計算機システムのコミュニケーションにおいても，ノンバーバルメッセージは重要なチャネルになり得る[*2]．

前2者は主に入力の解釈に関する項目，後2者は主に出力の生成に関する項目である[*3]．これらの項目はいずれもその処理のための方法論が確立されていない困難な課題であり，これまでに提案されてきたMMIFシステム（[長尾96, Bolt80, Kobsa86, Stock91, Koons93]など）では，ごく限られたドメインについて特定の入出力モダリティのみを考慮した作り込みのシステムを構築することにより，上記のいくつかの項目の効果を計るというアプローチがなされてきた．実用的なHIの構築のためには，より多くの項目について汎用な処理方式を開発する必要がある．

本研究において我々は，メディア統合とメディア補間を効率良く実現するマルチモーダル入力統合・解釈方式を開発し，その試作システムとしてマルチモーダル秘書エージェントシステムを構築した[金澤97, 中山

*1　ノンバーバルメッセージによって伝達される情報の量は，通常の言葉によるもの（バーバルメッセージ）のおよそ倍であると言われており，10倍以上であるとする文献もある．

*2　我々はこれまでに，ハンバーガーの販売という限定されたタスクについて音声を用いてユーザと対話できる音声自由対話システムTOSBURG-IIを開発してきたが，CGによる店員の表情というノンバーバルメッセージを用いて音声認識の失敗等のシステムの状態をユーザに伝達することがコミュニケーションの維持に有用であることが実験的にわかっている．

*3　もちろん後2つは入力においても重要な項目であるが，その実現のための困難度は極めて高くなる．

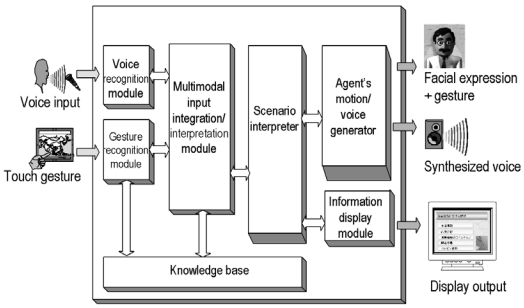


図 2 マルチモーダル秘書エージェントシステムの構成

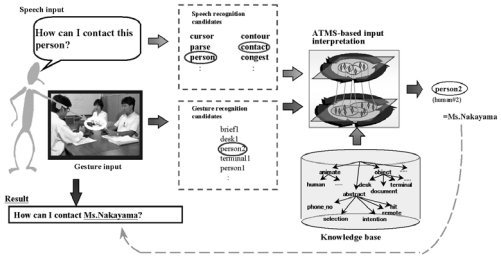
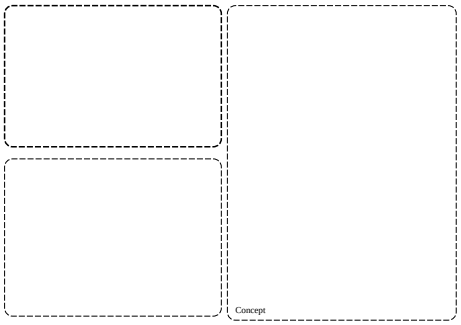


図 3 マルチモーダル照応解決処理の概要



97b]．図 1に示すように，本システムはインタフェースエージェント [Maes97] の一種であると言え，知識情報共有を促進するためのオフィス業務支援 [中山 97a] をアプリケーションとし，オフィス業務におけるノウハウを持った人間本人に代わって擬人化エージェントが問い合わせに答えるシステムである．本システムの利用者は，タッチパネルを用いたポインティング及びサークリングジェスチャ入力と，マイクからの音声入力を併用したマルチモーダル入力を行うことができる．システムは文字画像情報に加え，合成音声による音声出力，画面左下部に表示される男性あるいは女性の 3 次元 CG エージェントの身振り，表情といったノンバーバルメッセージを用い，オフィス業務ノウハウとシステムの状態をユーザに伝達する．エージェントは合成音声に同期して口を動作させる．

ユーザは文書名のリスト上で指示したい文書付近をサークリング / ポインティングしながら「この議事録を見せて」などとシステムに話しかけることができる．同じジェスチャ入力でもユーザの発話内容によってシステムのリアクションは変わる．例えば図 1の画面中にユーザのサークリングの軌跡が薄い円形で示されているが，ここでユーザの発話が「この議事録を見せて」でありその発話が正しく認識されるとシステムは「技術仕様検討会議議事録」を表示する．しかし，ユーザの発話が「この報告書を見せて」と認識されると，システムは「第二次市場調査結果報告書」をユーザに提示する．

図 2に本システムのモジュール構成を示す．マイクからの音声入力，及びタッチパネルへのジェスチャ入力は，それぞれ音声認識モジュール及びジェスチャ認識モジュールにおいて認識処理が行われ，その結果はマルチモーダル統合・解釈部（MM 統合・解釈部）に非同期に到着する．MM 統合・解釈部では与えられた認識結果を知識ベースの内容を参照しながら入力の統合・解釈結果を算出し，得られたユーザの要求をノウハウ検索システムに入力として与える．ノウハウ検索システムは与えられた要求を基にその返答を構成し，その情報を文字画像情報表示モジュールとエージェント制御モジュールに送付する．エージェント制御モジュールは，与えられた出力テキスト情報とその付加情報から合成音声とエー

指示語（連体詞），形容詞，名詞] からなる名詞句と [場指示語，名詞] からなる名詞句である．前者の例は「この赤い車」であり，後者の例は「この人」である．音声認識モジュールから音声認識結果が到着するとMM入力統合・解釈部はまず，各認識候補文から上記の条件に合致する名詞句を探索する．そのような名詞句が認識候補中にある場合，ジェスチャ認識結果と統合して照応解決を行おうとする．

照応解決処理は図4に示すような意味ネットワーク形式の知識ベースを参照しながら行われる．図4の左下の領域には表現に関する知識（言語知識）の意味ネットワークが配置され，右の領域には概念に関する知識（概念知識）が表示されている．これらの領域内の楕円は知識のクラスを表現しており，楕円内のシンボルはその要素を表している．そして言語的，あるいは概念上の知識が，これらクラスの要素間に張られたリンクとして記述されている．図4では，この図の左上にあるようにユーザがガレージの中に青い車と赤い車が描かれている絵をサークリングしながら「この赤いやつ」と発声した場合の処理過程を示している．このようなマルチモーダル入力が与えられると，MM入力統合・解釈部はこの図の太い破線で示しているように，その入力から導かれるリンクを辿りながらリンクの性質等により課せられた制約条件の中で解を探索することで問題解決を行い，右側の赤い車が照応解決結果として得られる．

3. 仮説推論に基づくMM入力統合制御

3•1 課　　題

汎用で実用的なMMIF開発のためにMM入力統合・解釈モジュールが持つべき能力として次の2点が挙げられる[*4]．

認識結果の曖昧性の取り扱い　一般に各モダリティにおける認識処理では100%の認識率は期待できないため，図3にあるようにMM統合・解釈処理モジュールには各モダリティにつき複数のスコア付けされた認識結果候補が入力として与えられる．入力モダリティが一般に複数となるMM入力の候補数はそれらの組合せとなり，膨大な数となることが多い．MM統合・解釈モジュールは，曖昧性を持った各入力要素について尤もらしい候補を効率的に探索する必要がある．

遅着データの統合　各入力モダリティでの認識処理に必要な計算量の差に起因して，あるモダリティからの認識結果が統合・解釈処理の開始後にMM統合・解釈モジュールに到着することがしばしばある．このような遅着データを取り込み，効率的に再統合して解釈する枠組が必要である．

[*4] これらの課題についてはMayburyが指摘している [Maybury94]．

汎用なMM統合・解釈アーキテクチャがあれば，MMIFのモダリティ拡張は容易になる．これまでに自然言語解析等で培われてきた構文解析技術は，MM解析にも有効な基盤を提供している．実際，汎用性を目指した研究成果には，MM-DCG[島津94] やKoonsらの枠組 [Koons93] をはじめとして，自然言語の構文解析技術を応用したものが多く見られる [Cohen94]．しかし，単純にこれらのシステムで曖昧性の問題を扱おうとすると，解釈の対象となるMM入力候補の数だけ解析処理を行うことになり非効率である．また，遅着データを扱おうとすると，解析処理をほとんどはじめからやりなおす必要が生じる．

遅着入力の再統合を行うことができるMM入力統合・解釈手法として難波らがドリップドロップモデルを提案しているが [難波97]，解釈時に必要なデータの抽象化のための汎用の枠組が定義されておらず，アドホックな処理を必要とする．また，入力の曖昧性の処理については上記の他の枠組と同様，候補の組合せの回数だけ解釈処理が必要になる．以上の議論から，上に挙げた課題を扱えるMM統合・解釈アーキテクチャを設計する際には，「他の候補の解析結果のうち，再利用できる部分は再計算しない」「遅着したデータに関係のない解析の途中データを再利用する」，といった問題解決の制御を実現する必要がある．

3•2 ATMSベースのMM統合・解釈方式

3•1の課題を満たすMM統合・解釈メカニズムの基盤として，ATMS[deKleer86a, deKleer86b] を我々は選択した．ATMSは多重世界問題を取り扱う問題解決器に対する知的キャッシュであり，複数の世界間で情報を最大限に伝達して効率よく問題を解決する上で有用な機能を提供する．

MM統合・解析問題をATMSベースの問題解決の制御問題として捉え，その問題解決プロセスを定式化することで，3•1で挙げた課題，すなわち各モダリティの認識結果の曖昧性と遅着データの統合の課題に対処し，効率的に適切な解釈結果を導く制御メカニズムを構築する，という方法論が本研究の柱の一つとなっているアイディアである．

図5にMM入力統合・解釈部から見たMM入力統合・解釈の枠組の構成を示す．音声モード認識部やジェスチャモード認識部といった各モダリティの認識モジュールは，利用者の音声入力やジェスチャ入力があるとその認識を行い，認識結果が得られると入力時刻情報とユニークなIDを付与してMM入力統合・解釈部に認識結果を出力する．これをMM入力要素と呼ぶ．MM入力要素の表現形態は各モダリティによって異なり，例えば音声モード認識部は文節ラティスを，ジェスチャ認識部は予め与えられた指示対象の中から認識結果対象候補を出力する，というように複数の認識結果候補が含まれる．

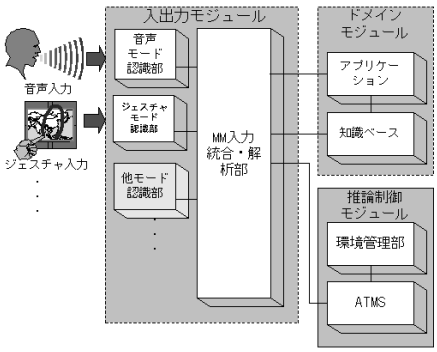


図 5　MM 入力統合・解釈部の構成

MM 入力統合・解釈部は各モダリティの認識部から非同期に MM 入力要素を受け取り，統合して解釈すべきひとかたまりの MM 入力要素集合（MM 入力）を決定する．これを MM 入力統合処理と呼ぶ．次に MM 入力統合・解釈部は，MM 入力要素それぞれについて候補を一つずつ選択した MM 入力候補の集合を生成する．そしてそれぞれの MM 入力候補について，知識ベース中のドメイン知識を参照して MM 入力の解析を行い，利用者の MM 入力内容を同定する．これを MM 入力解析処理と呼ぶ．MM 入力内容が同定されれば，それをアプリケーションに送付する[*5]．

MM 入力統合処理，及び MM 入力解析処理の過程は逐一 ATMS に通知され，ATMS はそれを記録する．MM 入力解析処理の失敗などにより ATMS に矛盾の発生が通知されるか，又はある MM 入力候補の解析処理が終了するなど ATMS が管理するデータが予め与えられた状態に達すると，環境管理部が問題解決のための新たな環境を生成し，ATMS にその環境への遷移を指示する．MM 入力統合・解釈部は新たな環境の下で問題解決を続行する．

MM 入力統合処理ではまず MM 入力要素全集合と呼ばれる集合 $\mathcal{S}$ の初期値を空とする．そして，(1) MM 入力解析処理の成功，(2) 十分長い間どのモダリティの認識結果も到着しない（タイムアウト）状態，のいずれかになるまで次の操作を繰り返す．

(1)　何れかのモダリティの認識結果が伝達されると，その ID を $\mathcal{S}$ に追加する．

(2)　$\mathcal{S}$ の任意のサブセット $\mathcal{SS}$ の解析処理を未だ行っていなければ，$\mathcal{SS}$ を MM 入力として仮定し MM 解析処理を行う[*6]．

(3)　タイムアウトすると $\mathcal{S}$ を空にする．MM 入力解析処理が成功すると，$\mathcal{SS}$ を出力すると共に $\mathcal{SS}$ を $\mathcal{S}$ から除去する[*7]．

MM 入力統合処理において ATMS に通知される仮定は以下の 2 種類のいずれかに属する．

integrate($\mathcal{SS}$, Mmi#)　$\mathcal{SS}$(第 1 引数) をまとまりとして統合し MM 入力とすることを仮定する．この統合と時間的に近い MM 入力統合について同じ MM 入力 ID である Mmi#(第 2 引数) が割り当てられる．

no_omission(modality, $\mathcal{SS}_m$, Mmi#)　MM 入力 ID が Mmi#(第 3 引数) の MM 入力統合に，modality（第 1 引数）の入力要素が $\mathcal{SS}_m$(第 2 引数) のみ含まれることを仮定する．システムに接続している各入力モダリティ毎に仮定される．

ここで，音声入力 vIn とタッチによる指示入力 gIn のみを扱う MMIF を例にとる．システムに対して利用者が音声とタッチ入力の両方を組み合わせて入力を行ったが，gIn の解析処理に時間がかかり vIn の解析結果である文節ラティス V#1 のみが先に得られたとする．この場合，システムは V#1 のみを統合された MM 入力とし，図 6の最上面左側に示すように

integrate([V#1], MMI#1)
no_omission(vIn, [V#1],MMI#1)
no_omission(gIn, [], MMI#1)

の 3 つの仮定を生成して特性環境に追加する．更に MM 入力 MMI#1 を表す導出データが

integrate([V#1],MMI#1)
& **no_omission**(vIn,[V#1],MMI#1)
& **no_omission**(gIn,[],MMI#1)
⇒ *integrated_input*([V#1],MMI#1)

という支持理由により導かれ，MM 入力解析処理に制御が引き渡される[*8]．

MM 入力解析では MM 入力候補集合から順に MM

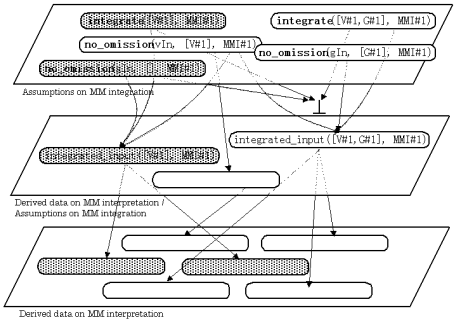


図 6　MM 入力統合・解釈処理の過程

*5　既に述べたようにマルチモーダル秘書エージェントシステムにおいてアプリケーションはノウハウ検索システム [中山 97a] である．

*6　時間的に近い入力要素を優先的に統合する等のヒューリスティックスを用いることで，MM 入力 $\mathcal{SS}$ の生成・テストを効率化することができる．

*7　解析に使われなかった入力要素（一般にノイズであることが多い）が $\mathcal{S}$ に堆積するのを防ぐため，一定時間以上離れた入力要素も $\mathcal{S}$ から除去される．

*8　実際には入力時刻情報も付加されるが，ここでは簡単化のため記述しない．

入力候補を選択して解析するが，その際に入力要素中の各要素について仮定が生成され，更にこれらの仮定から導出がなされるという形で解釈が進行し，その過程が逐一ATMSに蓄積される．紙面の都合上詳しい記述は省略するが，おおむね各モダリティの認識結果候補を一つの仮定として生成する．更に音声モダリティの文認識結果候補については，選択された候補文の仮定からその文に含まれる単語とその並びに分解して解析を進める[河野97]．これにより音声認識結果に関してはより小さい単位でMM入力解析過程がATMSに蓄積されるため，他の似た候補（例えば一つだけ単語が異なる文候補）の解析の際に以前の解析処理の途中過程をより多く再利用できる．

ひとつのMM入力候補の解析が終了し，次のMM入力候補が選択されると，環境管理部は新たなMM入力候補の解析処理に必要な環境を設定する．このとき，以前のMM入力候補の解釈過程でこれらのデータから導かれていたデータが参照できるようになる．このようにして過去の問題解決の途中過程を可能な限り生かしながら，適切な解を導ける候補の集合を探索することができる．

MM入力解析処理中にあるモダリティ，例えばジェスチャモダリティgInから統合すべき遅着入力G#1が届くと，新たに

integrate([V#1,G#1], MMI#1)
no_omission(gIn, [G#1], MMI#1)

の2つの仮定が生成される．このとき，

integrate([V#1,G#1],MMI#1)
& **no_omission**(gIn,[],MMI#1)
⇒ ⊥

integrate([V#1],MMI#1)
& **integrate**([V#1,G#1],MMI#1)
⇒ ⊥

の2つの矛盾が導かれる．この矛盾を解消するために環境管理部は，**no_omission**(gIn,[],MMI#1)と**integrate**([V#1],MMI#1)を含まず，**integrate**([V#1,G#1],MMI#1)を含む環境を決定し，遷移を指示する．この環境遷移により，図6の網掛けされたデータが自動的にコンテクストから除去されると共に，新たなコンテクストに含まれるデータ（**no_omission**(vIn,[V#1],MMI#1)）のみから導出されているデータ等については再計算することなく利用できる．このように進行中の処理の再利用可能な推論データの状態を保存しながら，遅着データを取り込みMM解析を再開できる．

4. 考　　察

4•1 CGエージェントインタフェース

我々は，6万人を越える来場者を数えた「**TOMMOR-ROW21** 東芝技術展」をはじめとする社内外の展示会等においてマルチモーダル秘書エージェントシステムのデモンストレーションを行った．また本システムの構築に先だってよりリアルなエージェント映像がリアルタイムで動作し喋るシステムを作成し，展示している[鈴木96]．それらの機会を通じて総合的にはエージェント映像には以下の効果が認められた．

Relaxation　エージェントの登場により，ユーザの心理的抵抗感を和らげる効果が認められた．

Gaze Control　エージェントが画面に登場すると，ユーザは誰と対話すべきなのかを認知でき，キャラクターを見つめながら話すことができる．

また，以下のポイントが重要であることがわかった．

Life-likeness　エージェントが生きているように見えることが重要である[*9]．生きているように見せかけるためには，エージェントにランダムな動きを加えるのが効果的である．特に視線の揺れ，首の振れなどが大きく影響した．

Balance Control　エージェントの見かけとその他の要素とのバランスが重要である．リアルな容貌にはリアルな声，リアルな動き，高い知的有能性が期待される．過度な期待を避けるため我々は，秘書エージェントにマンガ的な容貌を与えた．

また音声認識に失敗した場合，エージェントは「はい?」等と合成音声で話すと同時に，困惑したような表情と手の動きを見せるが，そのような直観的にわかりやすいノンバーバルメッセージの利用はシステムの状態伝達に効果的であることがわかった．

4•2 MM入力統合・解釈方式の効率と汎用性

曖昧性処理に関する本方式の効率向上の程度は各認識モジュールから得られる認識結果の数に依存し，一般に認識結果候補数が多く曖昧さの程度が高いほど枝刈りの効果が大きい．図1で示した例の場合，照応解決時のルールの発火数は本方式を利用しない場合に比べて約1/3程度であり，サークリングで囲まれる範囲がより大きくなると枝刈りの効果が増す．このような統合・解釈処理効率の向上はATMS利用のオーバヘッドを上回っており，本エージェントシステムのターンアラウンドタイムの短縮に寄与している．

本稿で提案したMM入力統合・解釈方式では，枠組み自体はドメインに対して汎用となるよう設計されている．すなわち，知識ベースに格納された（照応解決に必要な）ドメイン知識を差し替えれば，多様なドメインに対応できるようになっている．マルチモーダル秘書エージェントシステムでは，複数の画面においてマルチモー

*9 例えば，本システムではマイクから声が入力されはじめても，音声認識結果が得られるまでエージェントは何の反応も示さないが，それに対して「聞いているのか聞いていないのかわからない」という見学者の反応が多く見られた．

ダル入力が受理できるようになっているが，画面の切替の際に照応解決知識ベースを切替えることでこれを実現している．また，サークリングによる指し示しを伴う地図上での対象検索（例：「このへんの新しいホテル」）のドメインのシステムを試作して汎用性を検証した．その結果，知識ベースのドメイン知識，音声認識部に設定する音声認識対象語彙セット，ジェスチャ認識部に与える直接指示対象位置の認識対象候補集合の 3 種類のデータを差し替えるだけで，本システムは上記の両方のドメインにおいて MM 入力を解析することができた．

ただし，ジェスチャ認識部が出力する認識結果のスコアリングアルゴリズムについてはドメインによりある程度変更した方が結果がよくなることが実験的にわかっている．また，各モダリティについて仮定のグレインサイズの設定，及びその制御方法についてはモダリティやその組合せに依存する部分が多い．図 1の例にもあるように，今回の試作システムではタッチジェスチャの認識結果よりも音声認識結果に重点を置くような問題解決制御が働いている．この結果，タッチジェスチャで指示する場所についての自由度が高くなり，指示したいオブジェクトそのものをユーザが指さなくても MM 入力統合・解釈によりそれを得ることができるようになっている[*10]．その意味で本システムではジェスチャ認識結果に対して他メディアが補間することで，ジェスチャ認識単独ではより下位の候補を救い出すことができる．しかしながら音声認識が誤ると期待される結果が得られないことが多くなり，メディア間補間の効果がメディアにより異なっている．

4•3 時間の取り扱い

ユーザとコミュニケーションをとりながらタスクを遂行してゆくシステムでは，そのタスクに関連するデータは刻々更新されることになる．しかし後の問題解決で過去のデータを参照する場合もあり，時間の経過に従って単純にデータを消去したり無効にするわけにはゆかない．ATMS は推論順序に関係なくコンテクストを構築 / 再現できるため，上記のような要求に枠組的に対応が可能ではあるが，コンテクストの管理が複雑となる．

このような問題に対し，本稿で提案した枠組に次のように知識ベースのキャッシュ機能を持たせることで対応できると考えている．

(1) MM 入力統合・解釈部から知識ベース部への問い合わせは環境管理部を通して行う．

(2) 環境管理部は知識ベース部の回答に入力時刻情報を付与した ATMS の仮定を生成し，そのノード ID を内部履歴に記録する．

(3) 知識ベースが更新されると，知識ベース部は影響を受ける過去の問い合わせを検索し，その問い合わせに対する回答が無効である旨伝達する．環境管理部は対応する履歴レコードに無効フラグを付与する．

(4) 環境管理部は，MM 入力統合・解釈部からの問い合わせに対し，まず内部履歴を検索する．有効な同じ問い合わせが履歴中にあれば，対応する ATMS ノード ID を返答する．履歴中に有効なものがない場合にはじめて知識ベース部に問い合わせる．

このような仕組みにより，更新前の知識ベースに基づく推論結果を参照しながら，最新の知識ベースに基づく問題解決が可能になる．また，各時点での問題解決のスナップショットを再現できるようになる．この仕組みに加えて話題等の適切な時間の区切りを検出し記録する枠組が実現できれば，「〜の時の〜」という形態の参照が可能になると考えられる．

5. む す び

本稿では，MMIF のための仮説推論に基づく入力統合・解釈手法を提案した．この手法では，遅着データ，認識結果の曖昧性，時間経過によるドメインの変化に対応し，更に再計算の処理を大幅に減少させることができる．また本手法を用いた MMIF を複数のドメイン / アプリケーションで試作し，その汎用性を確認した．更に本研究では，提案した MM 入力統合・解釈手法を用いたインタフェースエージェントシステムを試作し，定性的ではあるがノンバーバルメッセージの効果を確認した．

今後はこの枠組を基に，各モダリティにおいて受理できる表現の拡張，及び入力モダリティの追加を行ってゆく．また，本研究で得られた知見を基に実用的で使いやすい MMIF の構築を進めてゆく予定である．

謝 辞

本エージェントインタフェースのバックエンドシステムである知識情報共有システムをはじめとして，多大な協力をいただいた (株) 東芝 研究開発センター 情報通信システム研究所 ヒューマンインタフェース技術センターの諸氏に感謝します．また，音声認識エンジン，及び音声合成エンジンを提供いただいた (株) 東芝 関西研究所の諸氏に感謝します．

◇ 参 考 文 献 ◇

[Bolt80] Bolt, R.A.：Put-that-there: Voice and gesture at the graphic interface, Computer Graphics, Vol.14, No.3, pp.262-270 (1980).

[Cohen94] Cohen, P.R.：Natural language techniques for multimodal interaction，信学論，J77-D-II，No.8，pp.1403-

*10 例えば図 1において画面上の適当なところをポインティングしながら「この議事録の全文を見せて」とユーザが発声しその通りに認識されれば，ポインティングされた座標に最も近い領域を占める「（会議）議事録」がユーザに提示される．

1416 (1994).
[deKleer86a] deKleer, J.D.：An assumption-based truth maintenance system, Artificial Intelligence, Vol.28, pp.127-162 (1986).
[deKleer86b] deKleer, J.D.：Back to backtracking, controlling the ATMS, In Proc. AAAI-86, pp.910-917 (1986).
[Finin97] Finin, T., et al.：KQML as an agent communication language, In Bradshaw, J. (Ed), Software Agents, MIT Press (1997).
[金澤 97] 金澤博史，知野哲朗，河野恭之，屋野武秀，池田朋男，鈴木薫，福井美佳，真鍋俊彦，竹林洋一：マルチモーダル秘書エージェントシステムの開発，情報処理学会研究報告，97-HI-74 (1997).
[Kobsa86] Kobsa, A.：Combining deictic gesture and natural language for referent identification, ACL, Proc. COLING86, pp.356-361 (1986).
[河野 97] 河野恭之，屋野武秀，池田朋男，知野哲朗：仮説推論に基づくマルチモーダル入力統合方式，インタラクション97論文集，情報処理学会，pp.33-40 (1997).
[Koons93] Koons, D.B., Spaarrell, C.J, and Thorisson, K.R.：Integrating simultaneous input from speech, gaze, and hand gestures, In Maybury, M.T. (Ed), Intelligent Multimedia Interfaces, pp.267-276 (1993).
[Maes97] Maes, P.：Agents that reduce work and information overload, In Bradshaw, J. (Ed), Software Agents, MIT Press (1997).
[Maybury93] Maybury, M.T. (Ed), Intelligent Multimedia Interfaces, MIT Press (1993).
[Maybury94] Maybury, M.T.：Research in multimedia and multimodal parsing and generation, Journal of Artificial Intelligence Review, Vol.8, No.3 (1994).
[長尾 96] 長尾：インタラクティブな環境を作る，共立出版 (1996).
[中山 97a] 中山康子，真鍋俊彦，竹林洋一：知識情報共有システム (Advice/Help on Demand) の開発と実践 - オフィス知識ベースとノウハウベースの構築-，インタラクション97論文集，情報処理学会，pp.103-110 (1997).
[中山 97b] 中山康子，金澤博史，竹林洋一：情報の発信と共有を促進するマルチモーダル秘書エージェントシステム，東芝レビュー，Vol.52，No.5，pp.25-28 (1997).
[難波 97] 難波康晴，田野俊一，絹川博之：マルチモーダルデータ特有属性の融合性を利用した意味解析，情報処理学会誌，Vol.38，No.7，pp.1441-1453 (1997).
[奥乃 91] 奥乃博：ATMSの高速化技法とその応用，人工知能学会誌，Vol.6，No.1，pp.24-37 (1991).
[島津 94] 島津秀雄，高島洋典：マルチモーダル definite clause grammar (MM-DCG), 信学論, Vol.J77-D-II, No.8, pp.1438-1446 (1994).
[Stock91] Stock, O.：Natural language and exploration of an information space: ALFresco interactive system, Proc. IJCAI91, pp.972-978 (1991).
[鈴木 96] 鈴木薫，山口修，福井和広，田中英治，倉立尚明，松田夏子：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)-，情報処理学会研究報告，96-HI-69，pp.47-53 (1996).
[竹林 94] 竹林洋一：音声自由対話システム TOSBURG-II，信学論，J77-D-II, No.8, pp.1417-1428 (1994).

〔担当委員：××○○〕

19YY年MM月DD日　受理

著者紹介

河野　恭之（正会員）

kono@krl.toshiba.co.jp

1989年大阪大学基礎工学部情報工学科卒業．1994年同大学大学院基礎工学研究科博士後期課程修了．同年(株)東芝入社．現在，同社関西研究所に勤務．博士(工学)．知的CAIやマルチモーダルインタフェースなど，ヒューマンインタフェースに関する研究に従事．情報処理学会，ACM各会員．

屋野　武秀

yano@krl.toshiba.co.jp

1994年神戸大学工学部電子工学科卒業．1996年同大学大学院自然科学研究科博士前期課程修了．同年(株)東芝入社．現在，同社関西研究所に勤務．ヒューマンインタフェースに関する研究に従事．電子情報通信学会会員．

池田　朋男

tommy@krl.toshiba.co.jp

1989年東京工業大学工学部情報工学科卒業．1991年同大学大学院理工学研究科博士前期課程修了．同年(株)東芝入社．現在，同社関西研究所に勤務．グループウェアおよびヒューマンインタフェースに関する研究に従事．情報処理学会会員．

知野　哲朗

chino@krl.toshiba.co.jp

1988年横浜国立大学工学部卒業．1990年同大学大学院修士課程修了．同年(株)東芝入社．以来，自然言語処理，談話理解処理，自然言語生成，文章要約処理の研究に従事．現在，同社関西研究所にて，音声言語理解，対話処理，ヒューマンインタフェースの研究に従事．情報処理学会，日本音響学会，日本ソフトウェア科学会各会員．

鈴木　　薫

suzuki@krl.toshiba.co.jp

1984年神戸大学工学部計測工学科卒業．1987年大阪大学大学院基礎工学研究科修士課程修了．同年(株)東芝入社．以来，文字図面認識の研究に従事．現在，同社関西研究所にてコンピュータグラフィックス，ヒューマンインタフェースの研究に従事．情報処理学会会員．

金澤　博史

kanazawa@krl.toshiba.co.jp

1984 年東京工業大学工学部機械物理工学科卒業．同年 (株) 東芝入社．
以来，音声認識，音声対話の研究に従事．現在，同社関西研究所に勤務．
1991 年〜 1993 年 (株) 日本電子化辞書研究所出向． 1993 年日本音響学
会技術開発賞受賞．電子情報通信学会，日本音響学会各会員．