仮説推論に基づくマルチモーダル入力統合方式

河野 恭之，屋野 武秀，池田 朋男，知野 哲朗
(株)東芝 関西研究所*

概要: 実用的なマルチモーダルインタフェースシステムを開発するために解決すべき必要な要素は，(1)遅着データの統合，(2)認識結果の曖昧性，(3)過去の問題解決データの参照，の3点である．本研究では，これらの課題を解くための汎用かつ実用的なマルチモーダル入力解析方式を開発した．本方式では，マルチモーダル入力解析過程を定式化し，システムの問題解決過程の管理に ATMS（Assumption-based Truth Maintenance System）を用いることで，遅着データをうまく統合し，また曖昧性を持ったマルチモーダル入力の解析を効率的に行うことができる．更に，ATMS に時間の概念を導入することで，問題解決データの一貫性を保ちながら過去の問題解決データを参照できる策組を提供する．ポインティングデバイスによる直接指示と音声を用いたマルチモーダル入力の解析を例に取り，提案する枠組みの動作を説明する．

A multimodal interpretation method based on truth maintenance

ABSTRACT Three requirements which should be met in order to develop a practical multimodal interface system, i.e., (1) integration of delayed arrival of data, (2) elimination of ambiguity in recognition results of each modality, and (3) treatment of past problem-solving data, are discussed. This paper presents an excellent 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). Furthermore, it is able to refer to past problem-solving data while maintaining the consistency of current data by incorporation of a time stamp management mechanism. The proposed framework is illustrated with an interpretation system that accepts multimodal input consisting of voice and direct indication by a pointing device.

1　はじめに

一般に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インタフェイス (MMIF) が求められている (Maybury, 1994).

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，

* 〒658 神戸市東灘区本山南町 8-6-26
Tel. 078-435-3104　Fax. 078-435-3162
E-Mail: {kono,yano,tommy,chino}@krl.toshiba.co.jp

全体的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 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)遅着データの統合，(2)認識結果の曖昧性，(3)過去の問題解決データの参照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 (de Kleer, 1986a; de Kleer, 1986b)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．

本稿では，MM 入力統合/解釈の制御メカニズムについて述べる．次にペン直接指示と音声によるマルチモーダル入力から利用者が参照している対象物を同定する照応解決を行うシステムを例にとり，その動作を述べる．

2 仮説推論に基づく入力統合制御

2.1 課題

我々は，汎用で実用的な MMIF 開発のために，

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．

認識結果の曖昧性: 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，MM 統合/解釈処理モジュールには各モダリティにつき一般に複数の認識結果候補が入力として与えられる．MM 統合/解釈部は，曖昧性を持った各入力要素について最も確からしい候補を探索しながら，効率的に解析処理を行う必要がある．

時間の取り扱い: 利用者とやりとりをし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．

という 3 つの課題の解決を MM 統合/解釈アーキテクチャ設計の目標とした[1]．

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，汎用性を目指したこれまでの研究成果には，MM-DCG(Shimazu *et al*, 1994)や Koons らのシステム(Koons *et al*, 1993)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(Cohen, 1994)．しかし，単純にこれらのシステムで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さなければならなくなる．また，各モダリティの認識結果の曖昧性を扱おうとすると，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．これらの課題を解決するためには，「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」，「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」といった問題解決の制御が必要となる．

2.2 ATMS ベースの問題解決制御

前節で挙げた課題を満足できる MM 統合・解析処理の制御メカニズムの基盤として，我々は ATMS(deKleer 1986a; 1986b)を選択した．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，複数の世界観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．

ATMS は問題解決器(PS)と独立に動作し，問題解決過程の一貫性を管理する．PS は対象知識・問題解決知識等を用いて，与えられた問題を解決するために推論を行う．PS はその推論過程を ATMS に通知する．PS が取り扱う全てのデータは ATMS が管理する．

PS が通知する情報は，$[N_1,N_2, ..., N_k \Rightarrow D]$ の形態をとり，データ D がデータの集合 $\{N_1,N_2,...,N_k\}$ から導出されたことを表す．$\{N_1,N_2,...,N_k\}$ を D の支持理由という．

PS で扱うデータは前提データ，仮定データ，導出データの何れかに分類される．前提データはいかなる状況でも成立するものとして定義される．仮定データは，他のデータに依存せずに成立すると仮定されたデータである．導出データは他のデータから推論規則により導出されたデータである．

導出データから支持理由をたどると，最終的には前提もしくは仮定データに到達する．このため，全てのデータに対してそれが依存する仮定の集合を計算することができる．この仮定の集合は環境と呼ばれる．各データについて PS から通知された支持理由を記録し，そのデータが成立する無矛盾な環境を計算することが ATMS の主要なタスクの一つである．矛盾の導出が通

[1] これらの課題の前 2 点については Maybury が指摘している (Maybury, 1994).

知されるとATMSは矛盾の環境を計算し，それを矛盾レコードに記録する．矛盾レコードに含まれる環境は許されない仮定の組み合わせと解釈できる．

推論のある局面はコンテクストと呼ばれ，その局面で成立するデータの集合により定義される．コンテクストに含まれる全てのノードを導出する環境をそのコンテクストの特性環境と呼ぶ．ATMSは，矛盾レコードを用いてPSの推論過程の無矛盾性を管理する．PSは矛盾レコードに記録された環境を包含しない新たな特性環境を選択し，ATMS に通知する．新たな特性環境が通知されると，ATMS はそれまでに通知された各ノードが新しい特性環境において成立する(in)か成立しない(out)かを決定し，inノードの集合により新しいコンテクストを構成する．そのコンテクストの下でPSは推論を続行する[2][3]．

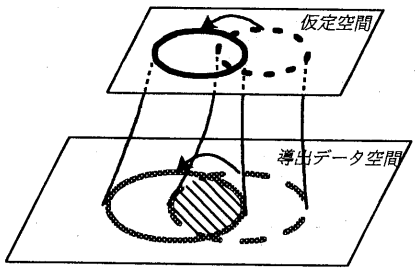

図1　ATMSベースの問題解決

図1にATMSベースの問題解決の制御プロセスを示す．上の平面に仮定の集合が，下に導出データの集合があり，上の平面の点線で示された環境から下の点線のコンテクストが導かれているとする．推論過程で矛盾が生じるなどPSの推論を制御する必要が生じると，環境管理モジュールが新たな環境を生成し遷移を通知する．ここで実線で示されている環境への遷移が通知されると，ATMS は新たな環境の下でのコンテクストを計算し，PSは新たなコンテクストの下で推論を続行する．この際，以前のコンテクストの下で行った推論で生成されたデータのうち，新たな環境においても導出可能なもの(図1の斜線部分)は自動的にコンテクストに含まれているため，PSは重複した推論を避けることができる．

特定のPSにATMSを導入する際の設計上のポイントとして，

1. データの分類（前提データ，仮定データ，導出データ）とその依存関係の設定
2. 矛盾解消ルールの定義
3. 矛盾解消器の設計

の3点が挙げられる．

2.3 MM統合/解析問題への適用

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，2.1節で挙げた課題の前2点，すなわち遅着データの統合と各モダリティの認識結果の曖昧性の課題に対処し，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．

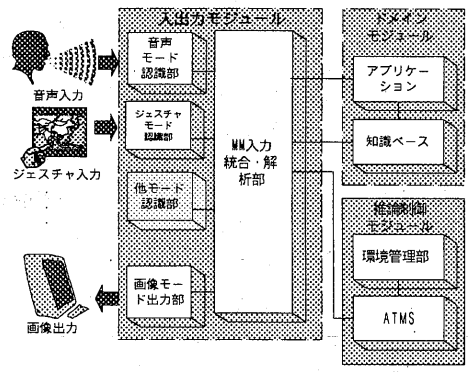

図2　MM入力統合システムの構成

図2にMM入力統合システムの全体構成を示す．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/解析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，というように一般に複数の認識結果候補が含まれている．

知識ベースにはMM入力統合・解釈に必要な知識，例えば参照の対象となる対象オブジェクトやそのクラス，対象オブジェクトが持つ属

---

[2] ここでは，並行型のBasic ATMS(de Kleer, 1986a)ではなく巡回型のDDB-guided ATMS (de Kleer, 1986b)を前提としている．

[3] ATMSの各ノードには，そのノードが成立する特性環境の最小サブセットの集合を表すラベルと呼ばれる情報が付帯している．あるノードのある特性環境における状態は，そのノードのラベルに特性環境のサブセットが含まれているかどうかを検証することで決定できる．ノードの状態検証，及び支持理由の通知によるラベルの更新はビット演算により高速化可能であり，その他様々な高速化の技法がこれまでに提案されている(奥乃 1991)．

性とその値，それらに対応する言語表現等，が意味ネットワーク形式で格納されている．

MM 入力統合/解析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/解析部は，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． また MM 入力統合/解析部は，アプリケーションからの出力要求を受け，出力モード部を通じて利用者へのフィードバックを行う．

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/解析部は新たな環境の下で問題解決を続行する．

2.3.1 遅着データと曖昧性の取り扱い

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 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 S に追加する．
2. S の任意のサブセット S_S の解析処理を未だ行っていなければ，S_S を MM 入力として仮定し MM 解析処理を行う[4]．
3. タイムアウトすると S を空にする．MM 入力解析処理が成功すると， S_S を出力すると共に S_S を S から除去する[5]．

MM 入力統合処理において生成される仮定は，以下の 2 種類のいずれかに属する．

integrate(*list of MMI element, MMI ID*): これまでに得られている入力要素集合のうち，*list of MMI element*（第 1 引数）をまとまりとして統合することを仮定する．この近辺の MM 入力統合について *MMI ID*（第 2 引数）が割り当てられる．

no_omission(*modality, list of MMI element, MMI ID*): *MMI ID*（第 3 引数）の MM 入力統合に，*modality*（第 1 引数）の入力要素が *list of MMI element*（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．

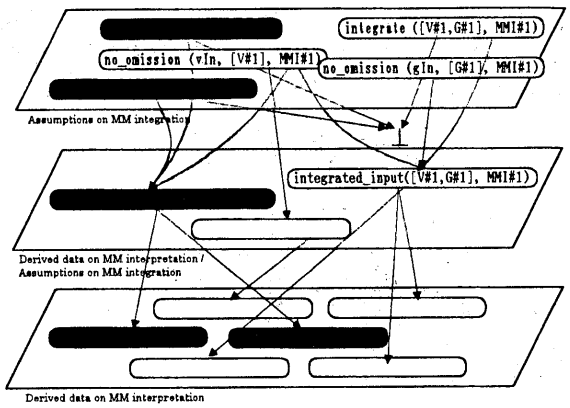

図3　MM 入力統合/解析処理の過程

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．この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 3 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[6]．

---

[4] 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 S_S の生成・テストを効率化することができる．

[5] 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 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 S から除去される．

[6] 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．

MM 入力解析では MM 入力候補集合から順に MM 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．この過程は 3 節で詳述する．

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 2 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．

2.3.2 時間経過の取り扱い

ATMS の発表以来，その汎用性と有用性から様々な分野の問題への適用が試みられてきた．HI の分野でも音声言語理解システムにおいて，曖昧性を持つ音声認識結果から言語情報を抽出する試み(Nishioka *et al*, 1991)等があるが，実用化には至っていない．

ATMS は推論順序に関係なくコンテクストを構築/再現でき，PS の推論順序の自由度を拡大する．しかしある程度の規模のタスクでは，既に更新された過去の推論結果を参照する必要が生じることがよくある．このような問題を仮説推論で扱おうとすると，コンテクストの管理が複雑となり，それが ATMS の実際的なタスクへの適用を妨げてきた．

本システムでは，次のように時間変化するデータを扱い，時間の取り扱いの課題に対処している．

1. MM 入力統合/解析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，ID を内部の履歴に記憶する．
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を伝達する．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/解析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 ノードの 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．

以上のようないわば知識ベースのキャッシュ機能を環境管理部が持つことにより，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できる．また，各時点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．

3 MM 照応解決問題への応用

ここでは，試作したタッチパネルによる直接指示と音声によるマルチモーダル入力の照応解決を行う MMIF を例にとり，MM 入力統合・解析過程を説明する．なお紙面の都合上，2.1 節で挙げた 3 つの課題のうち認識結果の曖昧性の取り扱いに関する処理過程を詳述する．

このシステムは，音声認識サーバ，ジェスチャ認識/出力サーバ，知識サーバ，照応解決クライアントの 4 つの 32bit プロセスで構成され，Windows95 もしくは Windows-NT3.5X, 4.0 上で動作する．これらのプロセスはサブネット上の他プロセスを自動的に探索し，知的エージェント間の知識交換言語である KQML (Finin *et al*, 1993) を用いて互いに通信することで，連携して照応解決を行う．

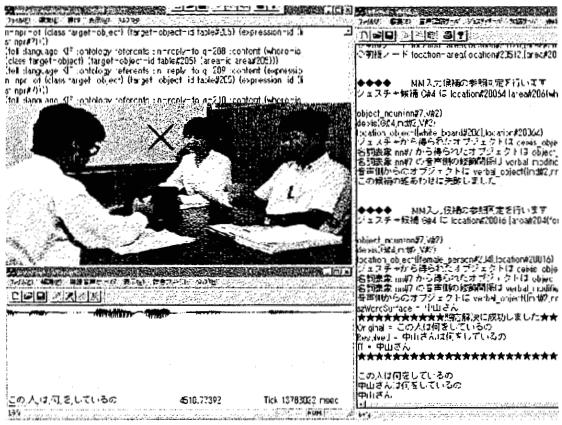

図4　MM照応解決システムの画面例

図4にシステムの出力画面の例を示す．この画面は，利用者がタッチパネルを用いて写真上の×印の部分[7]をポイントしながら「この人は何をしているの」と発声した場合のものである[8]．ここでのMMIFのタスクは参照表現を含むMM入力に対して参照物同定を行った上で参照表現を参照物名で置換し，例えばこの例では「中山さんは何をしているの」という文字列を生成してアプリケーションに送付することである．ここで図4にあるように，利用者がポイントした点に表示されている対象物は人物ではなく，別のもの（ホワイトボード）であるため，解である「中山さん」に相当する直接指示対象位置候補はジェスチャ認識結果の第1位ではなく，2位以下の候補となっている．このため，音声，ジェスチャの共に第1位の候補のみを解析しても解が得られず解析は失敗する．このため第2位以下の候補を探索する必要が生じる．このような場合のMM統合・解析過程は次のようになる．

MM入力統合処理では，2.3.1節と同様に

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

の3つの仮定が生成され，これらの仮定から，

integrated_input([V#1,G#1], MMI#1)

が導出される．

次にMM入力解析処理では，与えられたMM入力からMM入力候補を生成し，モダリティ毎に仮定を立てる．MM入力解析処理において生成される仮定及び仮定からの導出は入力モダリティによって異なる．この例では，文節ラティス V#1 の内容が [(この:md#2, 人:nn#7, は:unk#7, 何:unk#3, を:unk#6, しているの:unk#10), (この:md#2, 案内:nn#11, は:unk#7, 誰:unk#12, に:unk#13, 出せば:unk#14, いいの:unk#15)] で，直接指示対象候補集合 G#1 を [Location#20064, Location#20016, Location#20032] とする[9]．MM入力解析ではMM入力候補集合から順にMM入力候補を選択して解析処理を行うが，その際に入力要素中の各要素について仮定が生成される．この例では，最初のMM入力候補として [この:md#2, 人:nn#7, は:unk#7, Location#20064, 何:unk#3, を:unk#6, しているの:unk#10] の解析が行われるが，この場合，

vIn_phrase([md#2,nn#7,unk#7,unk#3,unk#6,unk#10],V#1)
gesture_location(Location#20064, G#1)
gestural_word(G#1, md#2)

の3つの仮定が立てられる．

更に仮定 **vIn_phrase**([md#2, nn#7, unk#7, unk#3, unk#6, unk#10], V#1)からは，

vIn_word(md#2,V#1),
vIn_word(nn#7, V#1),
vIn_word(unk#7, V#1),
vIn_word(unk#3, V#1),
vIn_word(unk#6, V#1),
vIn_word(unk#10, V#1)

の各々への導出がなされる．また，音声入力候補文の修飾関係を解析することで，同様に仮定 **vIn_phrase**([md#2, nn#7, unk#7, unk#3, unk#6, unk#10], V#1)から

modify(md#2, nn#7)
object_noun(nn#7, V#1)
verbal_modifier_noun([nn#7], V#1)

が導出される．次に知識ベースを参照しながらこの音声入力候補文の解析が進行し，

expression_class(nn#7, person)
& cc_relation(male_person,is-a, person)
& cc_relation(female_person,is-a, person)
& class_object(male_person,[male#202,male#203])
& class_object(female_person,[female#204])
⇒ object_of_noun(nn#7,[male#202,male#203,female#204])

vIn_word(md#2, V#1)
& vIn_word(nn#7, V#1)
& object_noun(nn#7, V#1)
& modify(md#2, nn#7)
⇒ object_singular(nn#7, V#1)

---

[7] 図4の×印は本稿の理解を容易にするために付けたものであり，実際の画面には表示されていない．

[8] 2.3節で述べたように，この枠組みでは遅着データを許容することができるため，厳密には2つのモードからの入力がほぼ同時である必要はなく，ある程度のタイムラグがあってもよい．どの程度のタイムラグを許容するかは設定可能である．

[9] 文節ラティス中の各単語は，"単語表象:単語ID"の形式で表現され，更に単語IDは"品詞#(品詞毎の)単語番号"という形式で表記されている．例えば，"人:nn#7"は表象「人」の単語が品詞nn(一般名詞表象)の7番目の単語であるということを表している．

vIn_word(nn#7, V#1)
& object_noun(nn#7, V#1)
& verbal_modifier_noun([nn#7], V#1)
& object_of_noun(nn#7, [male#202,male#203, female#204])
⇒ vIn_object(V#1, [male#202,male#203, female#204])

という導出過程を経て音声入力候補文のみから得られる参照物（3 人の「人」）が求められる．

次に，直接指示対象位置候補を表す仮定 **gesture_location**(Location#1, G#1)からの解析処理も，音声入力候補からの解析と並行して知識ベースを参照しながら行われ，

gesture_location(Location#20064, G#1)
& location_area(Location#20064, [area#206])
& area_object(area#206, white_board#1)
⇒ gIn_object(G#1, [white_board#1])

というように直接指示対象位置候補から得られる参照物（ホワイトボード）が求められる．これらの導出過程は逐一 ATMS に蓄積される．次に導出

integrated_input ([V#1,G#1], MMI#1)
& **gestural_word**(G#1, md#2)
⇒ deixis(G#1, md#2, V#1)

が行われ，これまでの結果を統合して参照同定結果を得ようとするが， gIn から得られる対象物が vIn から得られる参照物集合にないため，この MM 入力候補の解析は失敗し，

deixis(G#1, md#2, V#1)
& vIn_object(V#1, [male#202,male#203, female#204])
& gIn_object(G#1, [white_board#1])
⇒ ⊥

と矛盾が導かれる．この矛盾を解消するために解析中の MM 入力候補を表現する 3 つの仮定

vIn_phrase([md#2,nn#7,unk#7,unk#3,unk#6,unk#10],V#1)
gesture_location(Location#20064, G#1)
gestural_word(G#1, md#2)

が特性環境から取り除かれる．このため，これらの仮定から導かれていた全てのノードが out になりコンテクストから除去される．

次に，MM 入力候補集合から次の MM 入力候補である [この:md#2, 人:nn#7, は:unk#7, Location#20016, 何:unk#3, を:unk#6, しているの:unk#10]が選択され，

vIn_phrase([md#2,nn#7,unk#7,unk#3,unk#6,unk#10],V#1)
gesture_location(Location#20016, G#1)
gestural_word(G#1, md#2)

が仮定され，これらの仮定が特性環境に含まれるようになる．このとき，この MM 入力候補の音声入力候補文が前回の MM 入力候補と同じものであり，

vIn_phrase([md#2,nn#7,unk#7,unk#3,unk#6,unk#10],V#1)
gestural_word(G#1, md#2)

は前回の MM 入力候補解析時に仮定されている．このため，これらの仮定，すなわち音声入力候補文から導かれていた MM 入力解析の途中過程である

deixis(G#1, md#2, V#1)
object_singular(nn#7, V#1)
vIn_object(V#1, [male#202,male#203, female#204])

は自動的にコンテクストに含まれるようになり，この部分の解析処理を再び行う必要はない．

結果としてこの MM 入力候補の解析においては，直接指示対象位置候補を表す仮定 **gesture_location**(Location#20016, G#1)からの解析処理が行われ，導出

gesture_location(Location#20016, G#1)
& location_area(Location#20016, [area#204])
& area_object(area#204, female#204)
⇒ gIn_object(G#1, [female#204])

により直接指示対象位置候補からの参照物が得られる．そして，

integrated_input ([V#1,G#1], MMI#1)
& vIn_object(V#1, [male#202,male#203, female#204])
& gIn_object(G#1, [female#204])
& deixis(G#1, md#2, V#1)
& object_singular(nn#7, V#1)
⇒ referent_object([female#204] , MMI#1)

という導出により解の参照物[female#204（中山さん）]が得られる．これにより MM 入力候補の解析が成功し，図 4 の右側のウィンドウに示すように解である参照表現が置換された音声認識結果「中山さんは何をしているの」が得られ，アプリケーションに送付される．

上述の例では，最初の候補と音声入力候補文が同一の MM 入力候補で解析処理が成功し，効率のよい形で最初の候補の解析の途中過程を利用できた．ここまでに示してきたように，本方式では認識結果候補だけでなくより小さい単位で MM 入力解析過程が ATMS に蓄積され，より多くの場合に解析処理の途中過程を再利用可能となる．

4 考察

枠組の汎用性

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（参照同定に必要な）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．

本研究では，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」）のドメインと，ポインティングによる指し示しを伴う写真上での対象指示（例：「この人は何をしているの？」）の2種類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(3節におけるLocation)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に依存する部分が多い．

仮定の設定と矛盾の管理

過去の問題解決過程の再利用効率と仮定の状態制御に伴う副作用の大きさは一般にトレードオフの関係にある．すなわち，仮定がカバーする範囲を大きくしてデータの再利用効率の向上を図りすぎると，仮定の状態制御に伴う副作用が大きくなり有用な候補まで事前に枝刈りされてしまう．

また，試作システムを繰り返して使用しているうちにATMSノードの数及び矛盾レコードに記録される矛盾環境が増加し，ノードの探索及び矛盾の検証といったATMSを利用することによるオーバーヘッドが無視できなくなってくることも事実である．

これらの問題点はラベル及び矛盾の極小性と単調性というATMSが生得的に持つ性質に根ざしていると考えられる．人間が行う推論における「仮定」「矛盾」のグレインサイズは推論の局面に応じて変化する上に，忘却等により消滅することがある．このような状況に応じた「賢い忘却」のメカニズムがHIの高度化に伴い必要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．提案した枠組自体は汎用だが，制御方法はモダリティに依存する部分が多く，汎用の制御知識を分離する必要がある．今後は，試作した照応解決システムを基に実用的なMMIFを構築する予定である．

参考文献

奥乃 博 (1991). 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.

Cohen, P.R. (1994). Natural language techniques for multimodal interaction, *Trans. IEICE Japan*, Vol.**J77-D-II**, No.8, pp.1403-1416.

deKleer, J.D. (1986a). An assumption-based truth maintenance system, *Artificial Intelligence*, Vol.**28**, pp.127-162.

deKleer, J.D. (1986b). Back to backtracking, controlling the ATMS, In *Proc. AAAI-86*, pp.910-917.

Finin, T., *et al.* (1993). DRAFT Specification of the KQML, Agent-communication language, *The DARPA Knowledge Sharing Initiative*, http://www.cs.umbc.edu/kqml/kqmlspec.ps.

Koons, D.B., Spaarrell, C.J, and Thorisson, K.R. (1993). Integrating simultaneous input from speech, gaze, and hand gestures, In Maybury, M.T. (Ed), *Intelligent Multimedia Interfaces*, pp.267-276.

Maybury, M.T. (1994). Research in multimedia and multimodal parsing and generation, *Journal of Artificial Intelligence Review*, Vol.**8**, No.3.

Nishioka, S., Kakusho, O., and Mizoguchi, R. (1991). A generic framework based on ATMS for speech understanding system, *Trans. of IEICE Japan*, Vol.**E74**, No.7, pp.1870-1880.

Shimazu, H., Arita, S., and Takashima, Y. (1994). Multimodal definite clause grammar, In *Proc. COLING-94*, pp.832-836.