音声言語情報処理 17−20
(1997. 7. 19)

音声入出力、タッチジェスチャ入力、およびエージェントCG出力を持つマルチモーダル対話試作システム

○知野 哲朗　河野 恭之　屋野 武秀
池田 朋男　鈴木 薫　金沢 博史
{chino,kono,yano,tommy,suzuki,kanazawa}@krl.toshiba.co.jp
(株) 東芝 関西研究所
〒658 兵庫県神戸市東灘区本山南町 8-6-26

音声およびタッチジェスチャによる入力を受けつけ、合成音声出力と、顔表情と身振りを提示可能なCGエージェント出力を持つ、マルチモーダル対話試作システムについて報告する。特に、本システムの特徴である、仮説推論を用いたマルチモーダル入力統合方式、マルチモーダル照応表現解釈のための知識体系、知識交換言語による通信で連係動作するモジュール化されたアーキテクチャ構成、コンテンツ開発支援環境など、本システム実現の過程で開発した技術の概要を示す。

An Experimental Multimodal Interface System with Voice Input/Output, Deictic Gesture Input, and Agent CG Output.

Tetsuro CHINO　Yasuyuki KONO　Takehide YANO
Tomoo IKEDA　Kaoru SUZUKI　Hiroshi KANAZAMA
TOSHIBA Kansai Research Laboratories.
8-6-26 Motoyama-Minami-Cho Higashinada-Ku
Kobe, 658, JAPAN

In this paper, we discribe a multimodal interface system. The user can interact by voice input and diectic gesture via touch sensor on the screen. The system can reply by synthesized voice, and animated agent CG with facial expressions and hand and body gestures, and can provide visual information. We will focus on the following characteristics of our system developed through this research. A multimodal parsing method based on trouth maintenance, ontology for knowledge representation for multimodal reference resolution, archtecture of multimodal interfaces with high modulality, and support system for developping contents data.

1 マルチモーダルインタフェース

1.1 MMIF の技術課題

MMIF に固有の技術課題としては、以下の項目を挙げることが出来る。

メディア統合 ・・・複数のモードから別個に入力されるデータを、統合して一つのメッセージとして理解。文法だけでなく対象領域知識の表現 / 処理技術が必要。

メディア間補完 ・・・どのメディアに関しても、全く誤りを生じない解析手段を実現することは、非常に困難であるが、MMIF では複数のメディアからの認識結果を統合する際に、相互に補間し各々の不完全さを補間できる可能性がある。

メディア割当 ・・・授受する情報の種類、量、相手、あるは状況に応じて、最適なメディアを組合せて、コミュニケーションを行なう技術[1]。

ノンバーバルメッセージ処理 ・・・ノンバーバルメッセージ[2]の解釈 / 生成技術の実現は、人間同士のコミュニケーションに近い自然な HCI 実現のキーであり MMIF の利点が期待される領域の一つである。

これら課題に対し、本研究では、**メディア統合**および**メディア間補完**に関し、仮説推論に基づく MM 入力統合 / 解釈方式を提案し、この解釈処理過程で利用する知識体系とその管理モジュールを実装した。また、ノンバーバルメッセージ処理に関し、顔表情と身振りによる非言語メッセージを出力できる CG エージェントを開発した。

MMIF では、上述に加え汎用性と拡張性が要求される。従来の MMIF システム ([Bolt80] [Kobsa86] [Stock91] [Koons93] [長尾 96] など) では、特定の課題の極限られた領域だけを対象とし、かつ特定の入出力モードのみを考慮した作り込みによる実現がなされてきた。しかしこれは、実利用を目指す HI にとっては解決すべき問題点であるといえる。本研究では、この点に留意し、知識交換言語 KQML[Finin93] による通信で連係動作するモジュール化したシステム構成によって、高い拡張性を実現し、また、問題領域ごとのデータコンテンツ開発を容易とする支援環境を整備することで、高い汎用性を実現した。

[1]例えば、音声は位置関係など空間的な情報を伝えようとする場合には不適切であるが、多数の中から名前のわかっているものを指定するには最適なメディアの一つである。

[2]人間同士のコミュニケーションで授受されるメッセージのうちの、身振り、手ぶりなど、文字として記述できない種類のメッセージの総称

1.2 MM 照応解決問題

照応解決とは、言語などによって記述された対象を同定する問題である。ポインティングジェスチャ入力と音声入力とを統合する今回のシステムでは、この照応解決が第一に解決すべき課題である。

図 1は、本システムが行なうマルチモーダル照応解決処理の概要を示している。

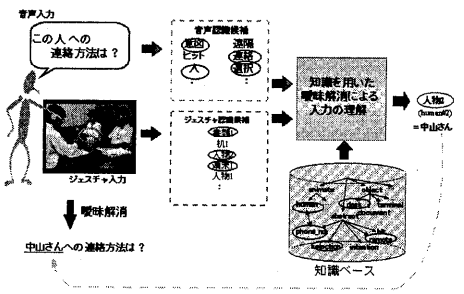

図 1: マルチモーダル照応解決処理の概要

ここでは、利用者が、画面に表示されたある人物の連絡先の提示を要求する場面での処理例が示されている。この音声とジェスチャを併用したマルチモーダル入力に対して、システムはそれぞれ音声認識処理とジェスチャ認識処理を実施する。しかしこういった認識処理では一般に解釈の曖昧性が発生し、一意の認識結果ではなく、スコア付けされた複数の認識候補が、それぞれの入力モードについて得られる。そして、音声とジェスチャという異なるメディアからの入力を統合するために、同一の表現レベルへと変換を行ない、知識との照らし合わせによる曖昧解消統合が行なわれる。この知識を用いた統合解釈によって、複数のモードからの入力の統合が実現するだけでなく、各モードの解釈結果によって相互に補完し合うことによって、

各モードを単独に利用した場合には解決不可能であった認識誤りの発生が抑制される[3]。

ここでは、各モード入力の解析結果の表現レベル統一と、領域知識の記述の為の知識体系が重要となる。

1.3 MM照応解決のための知識体系

本システムのMM照応解決のための知識は、意味ネットワークで表現される。

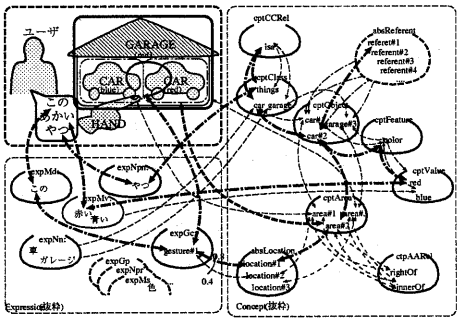

図 2: 知識体系の概要

図2は、左上の矩形領域はMM入力が行なわれている様子を表しており、また左下と右の矩形領域がMM入力の統合解釈のために設計した知識体系の概要を表している。

左下の領域には表現に関する知識(言語知識,cf. 辞書)が配置され、右の領域は概念に関する知識(概念知識)が配置されている。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。

例えば、言語知識に関しては、"expNpn"とラベル付けされた楕円は、代名詞表現のクラスを表現しており、また"expMv"とラベル付けされた楕円は、属性の値を規定する表現(一般には形容詞等)のクラスを表現している。また、一般には画面上の軌跡として入力されるジェスチャ表現に関してもそれぞれ個別の識別子が付与され、たとえばサークリングジェスチャ表現をあらわす"expGc"のクラスの要素として記述される。また、概念知識では、画面上に現れる指示対象を示す対象オブジェクト("cptObject")や、一般の事物を表現する対象クラス("cptClass")あるいは、対象クラス間の関係を表す対象クラス間関係クラス("cptCCRel")などが用意されている。

言語的あるは概念上の知識が、これらクラスの要素間に張られたリンクを用いて記述される。こういった知識を用いて、利用者が各モードを通じて入力されたマルチモーダル参照表現によって課せられる制約条件の中で、解を探索し照応解決を行なう。

この例では、音声入力された指し示しジェスチャの存在を表す言語表現「この」と、画面に示した軌跡を持つジェスチャ表現、さらに、音声入力された属性(eg. 色)の値(eg. 赤)を示す言語表現「あかい」、事物一般を指し得る言語表現「やつ」によって規定される参照対象の探索が行なわれ、結果として赤い車である"car#2"が得られることとなる。本研究で規定した知識体系は、現在17のクラスからなり、音声入力とジェスチャ入力によるマルチモーダル照応解決に関し、ほとんどの領域に適応可能な枠組を提供し、本システムの汎用性の実現に寄与している。

1.4 仮説推論に基づくMM入力統合解釈方式

我々は、汎用で実用的なMMIF開発のために、(1). **遅着データの統合**、(2). **認識結果の曖昧性**、(3). **時間の取り扱い**、の3つの課題の解決をMM入力統合解釈手法設計の目標とした。ここでは、「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。そこで、MM入力統合解析問題をATMS [deKleer86]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、上記の課題の前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した[河野 97]。

この枠組みにより、(1) 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、(2) 遅着データを取り込んでMM解析を再開することができるようになっている。

[3] 図1の例では、音声入力では第三位の認識候補「人」が選択され、ジェスチャ認識では第二位の認識候補「人物2」が選択され、正しい結果が得られている。

2 インプリメンテーション

2.1 マルチモーダル秘書エージェント

マルチモーダル秘書エージェント (図3) は、本人に変わって問い合わせに答えるコンピュータシステムである [福井 96]。

図 3: マルチモーダル秘書エージェントシステム

利用者は、従来のキーボードやマウスからの入力に加え、マイクからの音声入力と、タッチパネルを用いたポインティングおよびサークリングジェスチャ入力を併用したマルチモーダル入力を行なうことが出来、システムからは、文字画像情報に加え、合成音声による音声出力、画面左下に表示される男性あるいは女性の CG エージェントからの、身振り、手振り、顔表情 (の変化) といった非言語メッセージが提示される。さらに、CG エージェントの口は、合成音声に同期して動作する。

利用者は、これら入出力手段を用いて、ネット接続された各人の計算機を通して、オフィスデータベースやノウハウデータベース [中山 97] の情報を適宜授受する。これにより、各人が持っている有益な情報が、直接問い合わせなくても提供され、オフィスでの情報共有と円滑なコミュニケーションが実現される。

2.2 構成モジュール

図 4は、本システムの内部構成を表している。

各構成モジュールの機能の概要は以下である。

音声認識サーバ ・・・場面に応じてセットされる認識語彙について、連続出力分布型 HMM を用いたオートマトン制御による連続単語認識 [金沢 96] を行ない、認識結果と処理の成否に関する情報を、MM 入力統合サーバに通知。ノイズキャンセル、雑音免疫学習、HMM データの適応によって、高い耐雑音性を実現している。

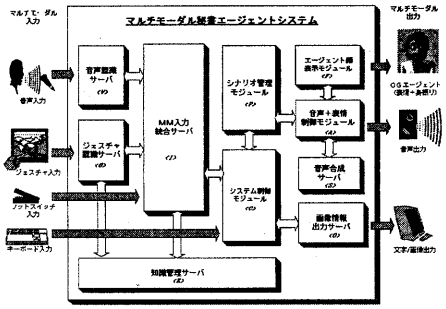

図 4: マルチモーダル秘書エージェント内部構成

ジェスチャ認識サーバ ・・・認識開始要求によって、利用者の入力ジェスチャを認識し、認識結果等を MM 入力統合サーバに通知。サークリングジェスチャと、ポインティングジェスチャを受け付け、ラフに入力されたジェスチャに対しても、少ない計算量で直観に合う認識結果を出力。

MM 入力統合サーバ ・・・音声認識サーバとジェスチャ認識サーバから非同期に伝達される音声入力とジェスチャ入力を、知識管理サーバの知識を参照して統合解釈し、解釈結果をシステム制御モジュールに伝達.

知識管理サーバ ・・・ MM 照応解決のために知識情報を管理し、他モジュールからの KQML による、要求に応じて保持している知識を編集あるいは提供。

画像情報出力サーバ ・・・画像情報の提示と、ジェスチャ対象となる画面上のオブジェクトの管理を行なう。

シナリオ管理モジュール ・・・シナリオデータに基づき利用者の入力に対するシステムの挙動の決定。

システム制御モジュール ・・・シナリオ管理モジュールの決定に従い、本システムの構成モジュールの動作を制御。

音声+表情制御モジュール ・・・合成音声出力機能とエージェント顔表示機能の同期を制御。

音声合成サーバ ・・・ユーザへの応答の音声波形を生成 [籠嶋 96] する規則音声合成器。

エージェント顔表示モジュール ・・・CG によって、エージェントの顔を表示。男性 / 女性について 6 種の表情を提

示可能で、合成音声と同期して口唇を動作。

3 コンテンツ開発支援環境

3.1 知識作成支援モジュール

本システムでは照応解決のための知識を用いた推論処理によって、複数モードからの曖昧性を持つ入力の統合解釈を行なっている。

知識データは、対象とする領域に強く依存しアプリケーションごとに用意する必要がある。また、照応解決のための知識データの他にも、利用者とのインタラクションの管理に用いられるシナリオデータや、利用者に提示する画像コンテンツ、ジェスチャ入力に対する認識候補の情報を含むシーン記述データ、場面に連動して切替えられる音声認識候補セットなど、多種多様のコンテンツが必要である。さらに、これらのコンテンツは、互いに依存関係を持ち、それが、複数のモジュールの連係した動作を可能としているため、コンテンツデータの整合性の維持が必須の条件であり、その重要性は非常に高い。

しかし、その依存関係は非常に複雑であり、その分量も膨大であるため、これを人手で管理することは不可能である。

そこで、本研究では、図5に概要を示したコンテンツ開発支援環境を開発した。

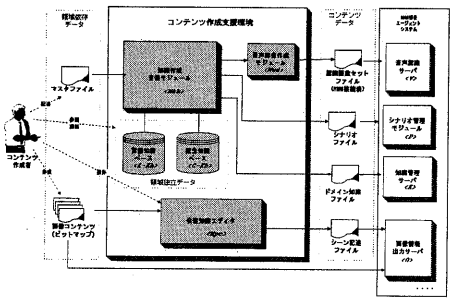

図 5: コンテンツデータ開発支援環境

コンテンツ作成者は、領域独立データを参照しながら、アプリケーションおよびドメインに依存する最小限の知識 / データだけを領域依存データとして追加するだけで、整合性が保証され本システムで実行可能なコンテンツデータのセットを得ることが出来る。

- 領域独立データとしては以下が用意される。

  概念知識ベース … 事物、属性、属性名、位置関係、階層関係、といった概念を表すクラスと、概念間の関係を表すクラス要素間などリンクを含む。

  言語知識ベース … 名詞、代名詞、形容詞、直示代名詞などの言語表現を表すクラスと、言語表現と概念を結びつけるリンク (辞書) を含む。

- 領域依存データとしては以下を用意する。

  マスタファイル … 状態遷移情報 (入力 / 応答, 遷移先, メディア割当, エージェント表情, 補助アプリスクリプトなど), 各場面の指示対象のクラス (概念) 名と属性情報,

  画像コンテンツ … 各場面で利用者に提示する画像情報 (ビットマップ)

- 自動生成されるコンテンツは以下である。

  ドメイン知識ファイル … 対象ドメインでの (マルチモーダル) 入力の統合解釈に必要な知識 (1.3節) が KQML 形式で記録されている。

  シナリオファイル … 利用者入力に対する本システムの挙動を記述。

  認識語彙セットファイル … 各シーン対応して切替られる音声認識候補のセット

  シーン記述ファイル … 各シーン対応する画像コンテンツ名、その上に存在するジェスチャ認識候補の位置および ID 情報など。

本開発支援環境によって生成されるコンテンツデータのうち、複数のモジュールから参照される知識 (*e.g.* ドメイン知識ファイル) は、知識管理サーバへのロードされ、他モジュールからの問い合わせに応じて適宜配布されることによって共有される。

4 考察

本試作システムの開発および試用によって、得られた知見は以下である。

まず、本システムは、音声とジェスチャによるマルチモーダル入力の解釈を可能としたが、この音声入力とジェスチャ入力は親和性の高い入力モードの組合せであり、双方の特徴を同時に活かしたマルチモーダル表現で、利用者に負担の少ない入力方式を実現出来ることを実証した。

また、エージェント顔表情を出力モードとしてシステムに加えたことの影響は予想外に大きく、非言語メッセージが利用者に与える効果が小さくないことが実証された。特に、謝り(PARDON)やおじぎ(BOW)の表情をいれたことが、マルチモーダル秘書エージェント全体をユーザフレンドリーなものとするのに大きく寄与したといえる。なお、HIシステムではその動作状態を利用者にフィードバックすることが重要であるが、そこでもエージェントの表情変化などの非言語メッセージが有用であると推察される。

さらに、知識データを人間可読な知識表現(**KQML**)で記述したことは、デバック効率を高めるという効果もあった。

今後の課題としては以下をあげることが出来る。まず、個別モジュールについては、各入力モードの受理可能表現の拡張と、各出力モードの表現能力の向上、認識結果の遅着を考慮した統合理解機能の実現などがある。また、非言語メッセージを利用したユーザへのシステム状態のフィードバック機能の実現、エージェントの提示する表情に応じた声質あるいは口調の変更、各モード入出力の制御機能を使った動的メディア割当の実現、インタフェース評価手法の開発および評価実施などがある。

5 まとめ

本研究では、(1). 仮説推論による入力統合手法と、(2). 照応解決のための知識体系によって、音声とジェスチャを組み合わせたマルチモーダル入力の統合理解技術を開発し、音声合成技術、および表情を提示可能なエージェントCG技術と統合することによって、「マルチモーダル秘書エージェント」システムを実現した。ここでは、MMIFの技術課題である、メディア間統合とメディア間補完を実現している。さらに、(3). モジュール化されたアーキテクチャ構成と、(4). コンテンツ開発支援環境とによって、MMIFに要求される、高い拡張性と汎用性も合わせて実現した。

参考文献

[Bolt80] R.A.Bolt, "Put-That-There": Voice and Gesture at the Graphics Interface Computer Graphics,14(3),pp.262-270, 1980.

[deKleer86] de Kleer, J., An assumption-based TMS, *Artificial Intelligence*, **28**, pp.127-162, 1986.

[Finin93] Finin, T., *et al*, DRAFT Specification of the KQML, Agent-communication language, *The DARPA Knowledge Sharing Initiative*, http://www.cs.umbc.edu/kqml/kqmlspec.ps, 1993.

[Kobsa86] Kobsa,A., Combining Deictic Gestures and Natural Language for Referent Identificaiton ACL, Proc.of COLING86, pp.356-361, 1986.

[Koons93] Koons, D.B., Spaarrell, C.J. & Thorisson, K.R., Integrating simultaneous input from speech, gaze, and hand gestures, In Maybury, M.T. (ed),*Intelligent Multimedia Interfaces*,pp.267-276., 1993.

[Stock91] Stock,O., Natural Language and Exploration of an Information Space: the ALFresco Interactive System Proc.of IJCAI91, pp.972-978, 1991.

[籠嶋 96] 籠嶋, 他, 高音質規則音声合成器のための有声音源生成法, 日本音響学会平成8年度春季研究発表会講演論文集I,pp.265-266, 1996.

[金沢 96] 金沢, 館森, 坪井, 竹林, 雑音免疫学習を用いたサブワードHMMに基づく雑音環境下での音声認識, 日本音響学会平成8年度春季研究発表会講演論文集,2-5-13,pp.83-84, 1996.

[河野 97] 河野, 屋野, 池田, 知野, 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.

[中山 97] 中山, 他, 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予稿集, pp.103-110, 1997.

[長尾 96] 長尾, インタラクティブな環境を作る, 認知科学モノグラフ(2), 共立出版, 1996.

[福井 96] 福井, 他, コミュニケーション支援のための個人情報公開システム(PIP)－音声とキー入力を用いたマルチモーダル対話の検討－, IPSJ-HI, HI-64-8, pp.43-48. 1996.