2004－SLP－50　(3)
2004／2／6

マルチモーダル対話の深化と記述言語の今後

新田 恒雄
豊橋技術科学大学 大学院工学研究科
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘１－１
E-mail:　nitta@tutkie.tut.ac.jp

あらまし　本報告は，マルチモーダル対話(MMI)研究に関する二つの話題を取上げて，今後の方向を検討している。MMI を持つ知的エージェントの研究が盛んになりつつあるが，全体の枠組みは必ずしも明確ではない。ここでは，Modality の処理から分割した U-M-A 三階層知能モデルを提案し，このモデルが知能エージェントの一連の MMI 処理機能を明確にすることを示す。MMI システムの開発では，比較的単純なものについてさえ設計方法論が確立していない。ここでは，MMI 設計に現れる幾つかのジレンマを考察し，その解決方法を提案する。

キーワード：　擬人化エージェント，マルチモーダル対話，知能モデル，対話記述言語

Consideration Towards Next-generation Multi-modal Interaction and MMI Description Language

Tsuneo Nitta
Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku-cho, Toyohashi-city 441-8580 JAPAN
E-mail:　nitta@tutkie.tut.ac.jp

Abstract: In this paper, two topics concerning Multi-modal Interaction are discussed and given their next generation schema. Research on intelligent agents has currently increased, however, we do not have the definite framework. Here, a three layer model of Unimodal – Multimodal- Amodal is proposed to clarify the functionalities in MMI between human and an intelligent agent. There are no methodologies in the development of MMI systems. Here, some dilemma that appear in MMI design are discussed, then a solution for the problem is proposed.

Keywords: Anthropomorphic Agent, Multi-modal Interaction, Model of Human Intelligence , MMI Description Language

1．はじめに

第三世代携帯 (3GPP) から家庭用ロボットまで，マルチモーダル対話の応用が始まろうとしている。しかし multi-modal interface (MMI) は, UI としての研究期間からみると 15 年に満たないため[1], [2]，それが本来どのように役立つもので, 実用化には UI 設計をどのように行っていけばよいか，といった基本的

な点も明確ではないのが現状である。

この手稿では，最初に MMI に実現して貰いたい（と著者が思っている）豊かな対話を幾つか説明した後，そこに示された機能仕様と実現への研究の枠組みを考察する。この中では，MMI を持つ知的エージェント研究の枠組みとして，Modality の処理から分割した三階層知能モデルを提案する。この知能モデルは，Uni-modal 処理の感覚レベル，Multi-modal 処理の認知レベル，そして A-modal（modality に依存しない）処理の意図レベルから構成され，全体として意図の理解－表出までを行うことを目指している。

次に，近未来に必要な MMI 設計に話題を転じ，特に対話記述の課題とそれらの解決方法を考察する。MMI は，GUI や音声インタフェースといった uni-modal interface に比べると，本格的な実用化には至っていない。実用化を阻害する大きな要因は，新手の UI に対するユーザの戸惑い以外に，3GPP が想定する比較的単純な MMI についてさえ設計方法論が未成熟なことが挙げられる。本文では，MMI 設計に現れる幾つかのジレンマとそれらの解決方法について考察する。

2．マルチモーダル対話(MMI)の深化

2.1 豊かな対話を与える MMI をめざして

図 1 は，エージェントが MMI を通して人間を含む環境から様々な情報を受け取りつつ，何らかのアクションをしようとしている場面を示したものである。このうち，1-A は講義・講演要約エージェントで，多くの入力情報（講演者の発話・表情・動作・スライド内容・ポインティング・デジタルインク描画など）から，講演内容の重要個所を抽出しつつ，マルチメディアの要約文書を作成している。

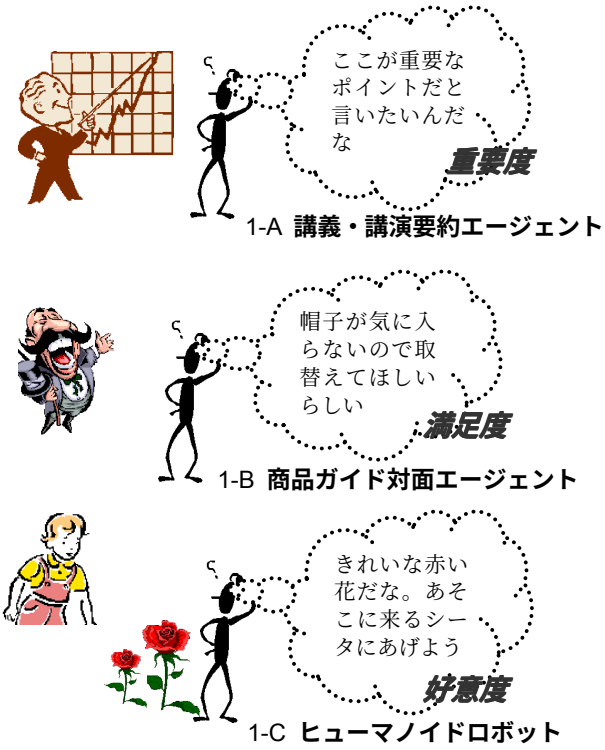

図 1 マルチモーダル対話の味わい

ここでは重要度抽出が大きな役割を持つ[3]。1-B は商品ガイド対面エージェントが顧客の苦情を確認している。ここでは入力情報（客の顔表情・発話・身体動作・商品など）から，商品（帽子）とそれに対する（不）満足度を推測し，応対（もしこちらに落ち度があるなら態度で示す必要も）を返すことが期待されている。次に 1-C では，ヒューマノイドロボットが赤いきれいな花を見つけ，そばにいる好意を持つ子に渡そうとしている（「天空の城ラピュタ」（宮崎駿監督）にこんなシーンがあった）。ここでは美的なものを識別し，生き物に対して好意度を計り，その結果，（自分の好きな）花を渡すことで好意を示すといった計画行動が必要になる。

2.2 Modality 処理の観点からみた知能モデル

2.1 に述べた例では，エージェント達が様々な情報を人間を含む環境から受け取り，加工

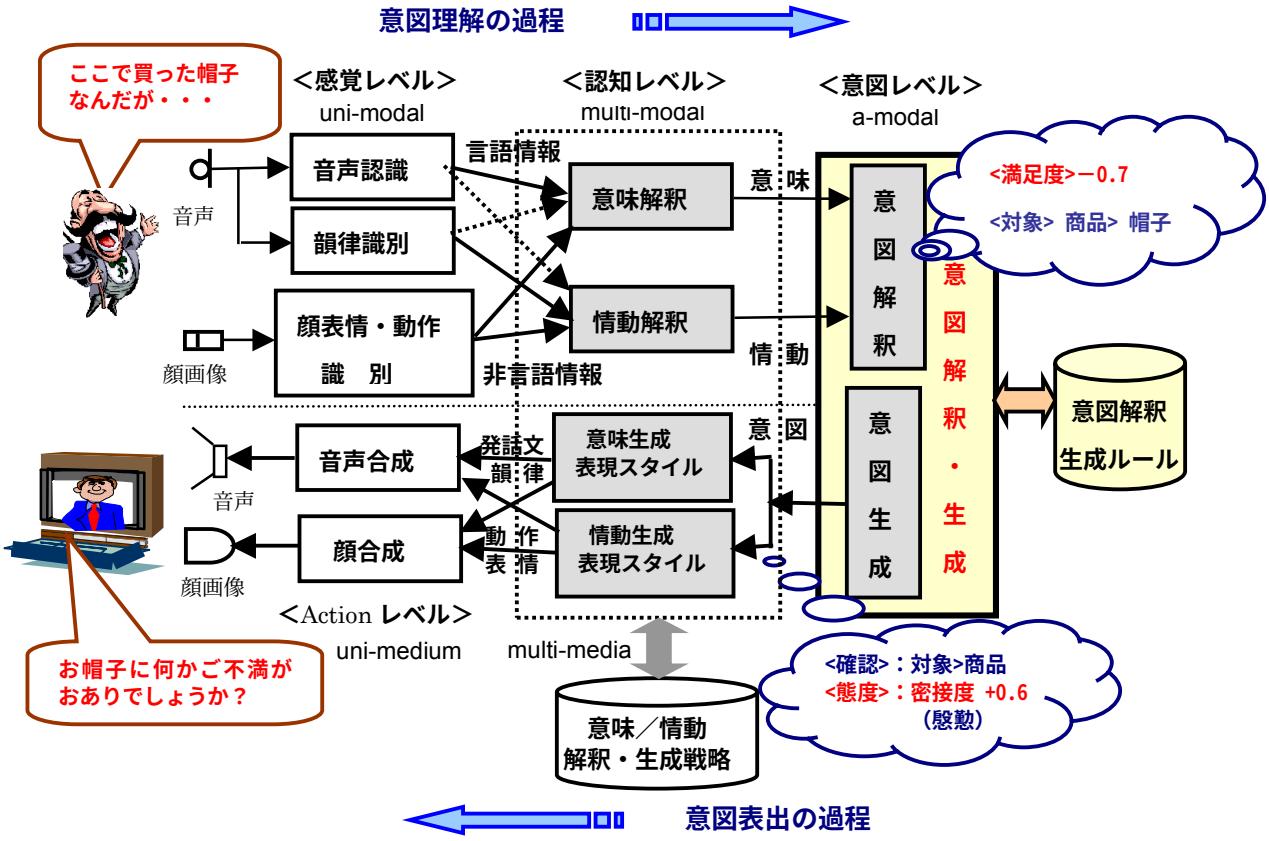

図 2 マルチモーダルコミュニケーションの例
：商品ガイド対面エージェント

し，必要なら環境へ働きかけを行っていた。これら一連の作業を，以下ではモダリティ処理の観点から考察する。図 2 は，商品ガイド対面エージェントの対話のやり取りを例に，内部処理を模式的に辿ったものである。まずエージェントからみて感覚レベルの入力情報（客の顔表情・発話・身体動作・商品など）が識別される。ここでは uni-modal 単位に処理が行われる[注]。次に，modality の統合解釈に進み，意味（<semantics type=? objct=商品(帽子)>）と情動（<emotion type=? value=快-0.8; 接近+0.2; 興奮+0.7>; value の内容については後述）が抽出される。これらは認知レベルの multi-modal 処理といえる。続いて，意味（商品（に関する発話））と情動（不快かつ興奮）を総合した意図解釈（<intension type=満足 value= – 0.7 objct=商品>）が行われ，必要なら対応する意図が表出される（<intension type=確認 $objct attitude=慇懃>）。これらは意図レベルの a-modal 処理である。意図は，認知レベルの処理段階で再び，意味生成（<semantics type=確認，objct=商品 “帽子”，mode=慇懃>）と情動生成（<emotion type=確認，value=快+0.4; 接近+0.2; 興奮 0.0>）分化される。続いて意味と情動は，これらを multi-media に分担させる戦略に基づき，発話文と韻律，表情と動作の表現スタイルに加工され，最後に Action レベルの働きかけが，uni-medium ごとの合成処理（音声合成と顔表情合成など）を通して行わ

れる。

以上を整理すると，MMI における内部処理の過程は，図 3 に示す感覚レベル／Action レベルー認知レベルー意図レベルの三階層のモデルで説明される。また同時に，説明の中では Modality/Media の係わり方の違い，すなわち Uni-modal/ Uni-medium 処理－Multi-modal/ Multi-media 処理－A-modal 処理に着目して知能モデルを再構成する考え方を示した。そしてこの考え方を導入することにより，エージェントが行う一連の MMI 処理の機能を一層明確にすることができることを示した。そこで，これを U-M-A 三階層知能モデルと呼ぶことにする。

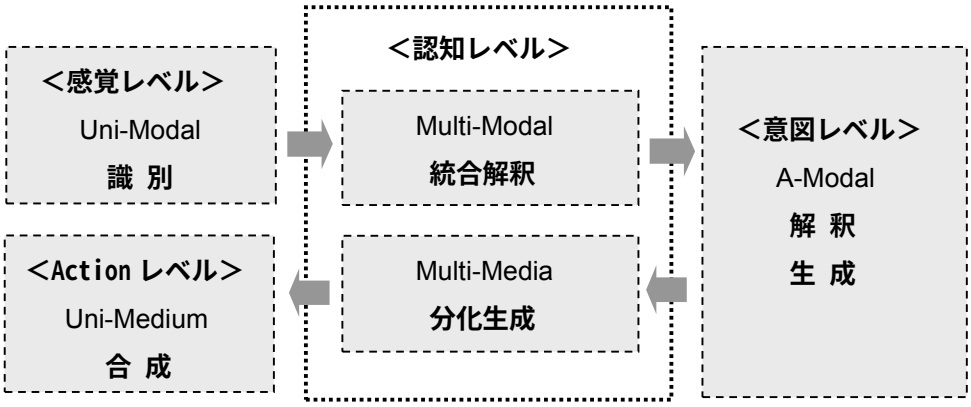

図 3　意図理解・表出の U-M-A 三階層知能モデル

（注）（感覚）モダリティは H.L.F. Helmholz (1821-94)やその師の J.P. Müller (1801-58) が最初に使用したようであるが，五感（視覚，聴覚，皮膚感覚（触覚，圧覚，振動感覚，温度感覚，痛覚など），味覚，臭覚），および運動感覚，平衡感覚，内臓感覚を指すのが一般的である[4]。これに対して，人間にはモダリティ毎に独立した認知記憶システムがあり，他のモダリティにも適用して可能であるとする立場も少数ではあるが存在する[5]。これらは人間の知覚からみたモダリティの使用であるが，擬人化エージェント（あるいはロボット）を考えると，機械は人間と比較してより広範なセンサーを持つ。彼らは GPS から位置を正確に「知覚」し，インターネットから多種多様な情報を（人間と違い電磁的獲得手段により直接）得ることもできる。そこで筆者は，エージェントを対象とする認知工学の研究では，人間を対象とする認知心理学の立場から離れ，modality を拡大して使用することが研究者に利益をもたらすと考えている。

2.3 U-M-A 三階層知能モデル研究の周辺

2.2 に説明した三階層のうち uni-modal/ uni-medium 単位の処理（識別・合成）は，音声・画像両分野で既に大きな研究者のコミュニティが形成されている。一方，multi-modal/ multi- media 処理（統合解釈／分化生成）と a-modal 処理（解釈・生成）は，研究が緒についたばかりである。これらの分野には，工学研究者（MMI 研究者，ロボット研究者ほか）だけでは手に負えない多くの課題が山積している。このうち対話の意味解釈・意図解釈は，言語処理に限定すればこれまで比較的研究集積があり[6], [7]，また音声と顔の動きを統合した意味解釈[8] や，情動を表現可能な音声合成・顔画像合成の研究も増えつつある[9], [10], [11], [12]。さらに，情動解釈結果を強化学習における報酬として利用することを仮定し，

エージェントの対話戦略と概念獲得を効率よく行う研究も始められている[13]。意味生成（ここでは意図から意味構造，さらに人間が了解可能な発話文生成までを含むとする）は，研究が余り進んでいない[14]。

以上に述べた分野と比較すると，情動を統合解釈する研究，意図から情動を生成する研究，および情動と意味を総合して意図を解釈する研究は非常に少ない。マルチモーダルコミュニケーションでは, 2.1 に例を示したように満足度，好意度，重要度といった情動解釈がキーとなって，意図理解が可能になるケースが少なくない。もちろん，言語情報だけでも多くの意図は伝えられるが,「赤は禁止の記号です[15]」といった短文でさえ，それが単なる説明なのかあるいは警告なのかは，意味と情動の二つの解釈を統合してみないと理解できない。

情動解釈の研究は，主に顔表情に対する観察から始められ（ダーウィンによる動物の顔表情研究が先駆とのこと), 情動を定量的に扱う試みも，顔表情から分類した情動の三次元モデルが古く提案されている[16]。この中で情動は, {快／不快, 接近（注意）／回避（拒否), 強さ（喚起レベル；他の表現として“睡眠／緊張”)} として構成される。例えばこれを正規化して表すなら，怒り{-1, 0, +1}, 嫌悪{-0.2, -1, 0}, 愛情{+1, +0.2, +0.2}, 驚き{+0.2, +1, +0.3} 等となる。我々は当面，情動を記録したコーパスを収集し，音声，顔画像，動作等を解析し，情動状態を統計的に解釈する研究を進める必要がある。

2.1 に示したエージェントの振る舞いは, 最終的に意味と情動から意図を理解する過程で決定される。すなわち，講義・講演要約エージェントにおける*重要度*，商品ガイド対面エージェントにおける*満足度*，ヒューマノイドロボットにおける*好意度*である。こうした計数値は，目の前で進行する個々のマルチモーダル対話と共に，エージェントが蓄えた体験や状況判断を含む推論から得られるものである。著者は, こうした研究を促進するために, U-M-A 三階層知能モデルに基づき階層間のデータ記述仕様を定めた後，MMI コーパス収集とアノーテーション作業を，研究者間の連携により進めたいと考えている。

3．MMI 記述の今後

3.1 MMI 記述の課題

MMI を設計する技術者は，多くのジレンマに直面する。具体的には，(1) visual interface(GUI) vs. speech interface に始まり, (2) タスク記述 vs.対話シナリオ記述，(3) 宣言的記述 vs.手続き的記述など, 互いに相容れない設計コンセプトの問題である。

音声対話の能力が，ユーザの意図を的確に掴み取れるなら別であるが，当面，そのようなことは期待できないとすると，上に挙げた問題の解決も限界を承知の上でのものとなる。(1)の GUI と音声 IF のジレンマでは, SALT[17] を含む多くのアプローチがそうであるように, GUI に音声 IF を同期させるのが現実的な解であるが，音声対話の自由度が大きく制限される。対話に自由度を与えつつ GUI とうまく同期を取る仕組みが当面の課題である。

(2)のタスク記述と対話シナリオ記述では，例えば slot filling ベースの記述は対話手順に影響されないものの，MMI による複雑な操作手順などは記述できない。逆に，対話シナリオを直接表現する記述では，複雑なタスクを書く際の労力が大きすぎる。開発者にとっては，記述量の少ないタスク記述ベースが魅力

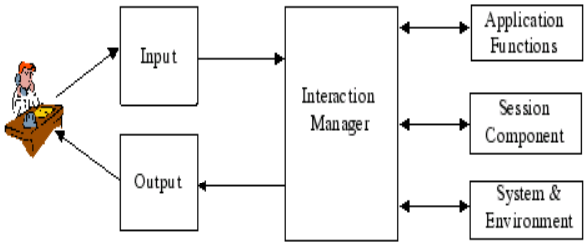

図 4 MMI フレームワークを構成する主要コンポーネント

的だが，その場合，MMI システム（の対話制御）に負担が掛かるため，対話制御を楽にする記述方法が課題である。

(3)の宣言的記述と手続き的記述では，前者は記述が短く保守性・再利用性に優れる反面，対話全体の流れが把握し難いという問題がある。他方，手続き的プログラミングスタイルでは，この逆の傾向がある。以上に説明した(1)から(3)の問題は，これらを一つの括りで解決することは困難であり，3.3 では階層に分けた記述方法を提案する。

3.2 W3C-MMI-WG の動向

W3C（World-Wide-Web Consortium）は，インターネットで利用する様々な技術の標準化を推進する団体である。この活動の中で，W3C-MMI-WG は MMI による Web アクセスを実現することを目的に結成された[18]。この WG では，(1) MMI 記述言語に対する要求仕様[19]，(2) MMI システムのフレームワークの二つを対象に主に提案・検討が行われている。図 4 に MMI フレームワークの構成を示した。このうちの入力部について，詳しい構成例を示したものを図 5 に示す。

MMI フレームワークでは，マルチモーダル入力データ(マルチメディア出力)を，EMMA (Extensible Multi-Modal Annotation language) 形式で統一することにより，モダリティに依存しない解釈を実現することを目指している

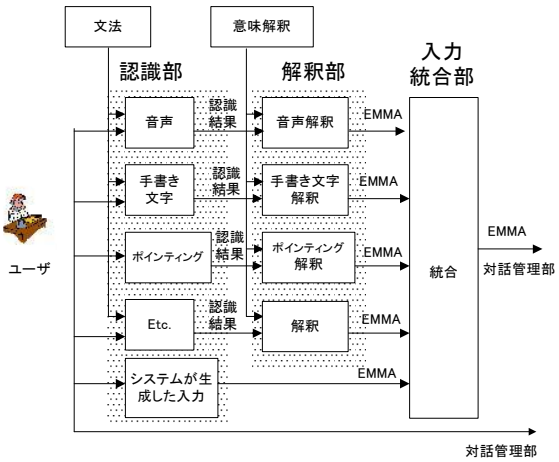

図 5 入力部を構成するサブコンポーネント

[20]。EMMA の記述例を図 6 に示す。この例は，音声認識エンジンの出力データをもとに，音声解釈部が生成した EMMA の記述例を示している。<emma: interpretation …….>は，入力情報をそのまま記述する”raw”と，入力情報を解釈した結果として生成される”int”から構成されている。またこの例では，タイムスタンプや確信度が同時に付加されている。EMMA 形式は，個別モダリティの解釈結果と複数のモダリティに対する統合結果の双方で使用される。一方，解釈方法については利用者に任されると考えられるが，このうち個別モダリティに対する EMMA を統合結果の EMMA に変換する文法ルールが討議されている。

このほか，各種デバイス間でやり取りされるイベントの処理方法，ペン入力 (ink) のための記述言語 InkXML，さらに図 4 に示されている複数ユーザの利用を想定した Session component，利用環境の違いに対応する System&Environment なども検討が行われている。

```
<emma: emma>
  <emma: interpretation  rdf: ID=”raw”>
       これを３こください
  </emma: interpretation>
  <emma: interpretation  rdf: ID=”int”>
    <order>
      <num>3</num>
      <goods>これ</goods>
    </order>
  </emma: interpretation>

  <rdf: RDF>
    <rdf: Description  rdf: about=”#raw”
         emma: medium=”audio”
         emma: mode=”speech_command”
         emma: start=”XXX”
         emma: end=”XXX”>
      <emma: confidence>0.9 </emma: confidence>
    </rdf: Description>
    <rdf: Description  rdf: about=”#int”>
      <emma: model

ref=http://myserver/models/shopping.xml/>
        <emma: derivation ref=”#raw”>
    </rdf: Description>
  </rdf: RDF>
</emma: emma>
```

図 6 RDF による EMMA 記述の例

3.3 改良版 XISL による MMI システム設計の提案

XISL（eXtensible Interaction Scenario Language）[21], [22]は，モダリティ記述に関する拡張性が高い, 対話シナリオを Web コンテンツから分離できるといった特徴を持つ。しかし XISL1.1 では，入力モダリティの指定と入力統合方法の二つを入力操作記述（<operation>タグ）内に直接書く必要があった。このため, 端末ごとにモダリティ（<input>タグの属性で指定）仕様が異なる際には，記述量が増えるという欠点があった。

現在，新しく設計を始めているXISL1.x では，こうした欠点を修正すると共に，3.1で考察したMMI設計上のジレンマを軽減することを目指している。基本的な考え方は，MMI記述を二つの階層に分離し，まずタスクをA-modalな対話手順に沿って記述することで，対話を簡潔にする（この部分が新しいXISL仕様の対象となる）。次に，個別モダリティ間の

```
<xisl>
  <head>
    <model id="Shopping">
      <instance>
        <SHOP>
          <SELECT>
            <GOODS/><PRICE/><NUM/><SUBTOTAL/>
          </SELECT>
          <LIST>
            <GOODS name="りんご" price="100"/>
            ・・・
      </instance>
      <bind ref="SHOP/SELECT/SUBTOTAL"
calculate="SHOP/SELECT/PRICE*SHOP/SELECT/NUM"/>
    </model>
  </head>
  <body>
    <dialog id="sample" scope="dialog">
      <begin>対話の導入処理記述部</begin>
      <prompt>
        <output type="agent" event="speech">・・・</output>
        <grammar rule="mmigram.xml#goods_num"/>
        ・・・
      </prompt>
      <action>
        <set_element var="vgoods"
            ref=" SHOP/SELECT/GOODS"/>
        ・・・
      </action>
      <exchange>・・・</exchange>
      ・・・
```

図 7　新 XISL による A-modal な対話の記述

同期制御を含むMulti-modal統合文法（新XISLから呼ばれる）を記述する。

このように対話記述を分離したことで，(1) GUI-音声IFの問題に対しては，これらを第二段のMulti-modal統合文法中で記述する（MMIをゼロから設計する場合），またGUIが既に存在する場合には，XISL中のモデルとHTMLをバインドさせることで，既存のGUIとのデータをやり取りできるため，簡単にシナリオを作成できる。

(2) のタスク – 対話シナリオの問題に対しては，対話手順をA-modalな記述としたことで，シナリオ記述量がタスク記述と同等になる。また(3) の宣言的 – 手続き的の問題は，第一段のA-modalな記述（手続き的記述）で入力手順を簡潔に書けるようにしたと共に，

```
<grammar id="goods_num">
  <operation comb="par">
      <input mode="touch" match="ols.htm#goods"/>
      <input mode="speech" match="spgram.gram#num"/>
  </operation>
</grammar>
```

図 8　マルチモーダル統合文法の例

第二段のMulti-modal統合文法を宣言的に書くことを可能とした。最後に，図7に新しいXISLによるA-modalな対話記述の例を，また図8にマルチモーダル統合文法の例を示す。

4．あとがき

Modality の処理から分割した U-M-A 三階層知能モデルを提案し，このモデルが知能エージェントの一連の MMI 処理機能を明確にすることができることを述べた。また，MMI 設計に現れる幾つかのジレンマを考察し，その一解決方法を提案した。本報告をまとめるにあたっては，認知心理学の専門家である本学知識情報工学系北崎充晃助教授から多くの示唆をいただいた。ここに厚く御礼申し上げる。

参考文献

[1] Taylor M.M., Neel F., and Bouwhuis D.G. Ed.: The Structure of Multimodal Dialogue, North-Holland (1989).

[2] 新田恒雄：GUI からマルチモーダル UI (MUI)に向けて，情報処理学会誌，Vol. 36, No. 11, pp.1039-1046 (1995.11).

[3] 山田，松田，田口，桂田，小林，新田：“講義再現システムにおけるスライド重要度抽出,人工知能学会誌 Vol.17，No.4，pp.481-489 (2002).

[4] 心理学辞典，有斐閣 (1999).

[5] 岩波講座認知科学第５巻　記憶と学習 (1994).

[6] 岡田直之：語の概念の表現と蓄積，電子情報通信学会 (1991).

[7] 石崎雅人，伝康晴：談話と対話，東大出版会 (1988).

[8] 藤江，江尻，菊池，小林：パラ言語の理解能力を有する対話ロボット，情処研報，SLP-48, pp.13-20 (2003).

[9] 都築，全，徳田，北村，Bulut, Narayanan：HMM に基づく感情音声合成に関する検討，音学全大，Ⅰ-1-8-30, pp.241-242 (2003-9).

[10] 川本，下平，新田，宇津呂，西本，中村，伊藤，森島，四倉，甲斐，山下，小林，徳田，広瀬，峯松，嵯峨山：カスタマイズ性を考慮した擬人化音声対話のソフトウエアツールキットの設計，情報処理学会論文誌，43, 7, pp.2249-2263 (2002. 7).

[11] 石塚：生命的エージェントによる感性的マルチモーダルコンテンツ記述と生成，音学全大，Ⅰ-2-8-10, pp.265-266 (2003-9).

[12] 三輪，伊藤，高信，高西：人間との円滑なコミュニケーションを目的としたヒューマノイドロボットの心理モデルの構築，人工知能学会 AI チャレンジ研究会（第 18 回）, SIG-Challenge-0318-7, pp.39-44 (2003-11).

[13] 田口，山本，桂田，新田： “Infant Agents 間相互対話による対話戦略の自動獲得”，人工知能学会研究会資料 SIG-SLUD-A302-04，pp.15-20 (2003-11).

[14] Young, S. J., and F. Fallside: "Speech Synthesis from Concept: A Method of Speech Output from Information Systems," J. Acoust. Soc. Am., 66(3):685-695 (1979).

[15] ロラン・バルト：モードの体系，みすず書房 (1972).

[16] Schlosberg, H.: Three dimensions of emotion, Psychological Review, 61 (1954).

[17] http://www.saltforum.org/

[18] http://www.w3.org/2002/mmi/

[19] 中村，桂田，山田，新田：MMI 記述言語の標準化動向と XISL の対応について，電子情報通信学会研報，SP-2002-160, pp.73-78 (2002-12).

[20] http://www.w3.org/TR/emma/

[21] 桂田，中村，山田，山田，小林，新田：MMI 記述言語 XISL の提案，情処論 Vol.44，No.11， pp.2681-2689 (2003).

[22] http://www.vox.tutkie.tut.ac.jp/XISL/XISL.html