音声言語情報処理 5－2
（1995. 2. 3）

Mutli-Modal Method - マルチモーダルインタフェース構築方法論 -

島津 秀雄　　高島 洋典
日本電気(株)　情報メディア研究所

本稿では、マルチモーダルシステム設計手順である Multi-Modal-Method (MMM) について述べる。MMM は、我々のマルチモーダルインタフェースシステムの開発経験をもとにして定義された。MMM は、1) タスク選択、2) モードとメディア選択、3) コーパス収集、4) コーパス分析、5) 解釈系レベル設定、6) システム構成設定、7) 文法記述、の 7 ステップの繰り返しから成る。MMM の文法記述のためには、マルチモーダル文法記述形式 Mutli-Modal Definite Clause Grammar (MM-DCG) をツールとして提供している。

Multi-Modal-Method: A Design Methodology for Building Multi-Modal Systems

Hideo Shimazu and Yosuke Takashima

Information Technology Research Laboratories
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Japan, 216

This paper describes Multi-Modal-Method, a design methodology for building multi modal systems. Multi-Modal-Method defines the procedure which interface designers may follow in developing multi-modal systems, and provides MM-DCG, a grammatical framework for multi-modal input interpretation. Multi-Modal-Method has been inductively defined through several experimental multi-modal interface systems.

1 はじめに

本稿では、マルチモーダルインタフェース構築方法論である Multi-Modal Method (MMM) について述べる。MMM は、我々のマルチモーダルインタフェース開発経験を通じて帰納的に定義された。

本研究の動機は2つある。1つは、マルチモーダルインタフェース方法論の確立は、今後のインタフェースを考える上で必須と考えられたからである。我々は、マルチモーダルインタフェースを Windows や X window のような GUI (Graphic User Interface) の次にくるインタフェースパラダイムの有力な候補と考えている。

第2に、マルチモーダルインタフェース構築の方法論が存在しなかったことである。近年、マルチモーダル研究は増加しているが、それらは特定の応用問題の解決手法に焦点があり、開発の方法論やそれを反映したマルチモーダルシステム記述言語について提案がされてこなかった。そこで、我々は、まず世界で初のマルチモーダル文法記述形式 MM-DCG を開発した[Shimazu et. al., 1994]。MM-DCG を使うことで、マルチモーダルインタフェースを形式的に構築できるようになった。さらに進んで、我々自身のマルチモーダル開発経験から帰納的にマルチモーダルシステム構築の方法論を定義することを試みた。本稿では、その試みから生まれた Mutli-Modal-Method について述べる。

2 GUI を超えるインタフェースとしてのマルチモーダルインタフェース

我々は、マルチモーダルインタフェースは、現在の GUI の発展の一形態であると考えている。現在の GUI は、オブジェクト指向とイベント指向アーキテクチャの2つのアイデアを基本にしている。オブジェクト指向の世界では、クラス / サブクラスの関係定義を使ってソフトウェアの部品をあたかも物理的なもののように扱うことが可能になる。プログラマは、簡単なオブジェクトを組み合わせてより複雑なオブジェクトを定義することが出来る。コンピュータの画面について考えた場合、画面上のウィンドウやその中のボタンやメニューのような部品がすべて統一的にオブジェクトとして扱えるのは、インタフェース設計者にとって非常に便利である。しかし、最近の GUI システムでは、従来の定義の範疇をこえるある種のオブジェクトが出現している。例えば、利用者のマウス操作やペンで絵を描くこと、ドキュメントを走査すること、あるいは、文を発話すること、などはオブジェクトとして扱うのは不自然であり、むしろ対象のオブジェクトに生ずる「イベント」としてとらえられるべきである。

本稿での我々の主張は、オブジェクト指向にイベント指向を付加することは、正しい方向ではあるが、それだけでは充分ではなく、イベントの並びであるイベント列をモデルの単位として考えることが重要であるということである。これは、とりわけマルチモーダルインタフェースに於て、一般的である。マルチモーダルインタフェースでは、同時に複数の異なる機器からイベントまたはイベント列が到着する。このようなインタフェースでは、マウスのクリック操作、利用者の発話、ペンでの描画、タイプされたコマンド、などは1つの入力に統合されなくてはならない。入力におけるイベントの順番は、それらの入力を理解する上で非常に重要である。そのようなイベント列を解析するためには、個別のアプリケーションプログラムで勝手に処理するのではなく、形式的な枠組みが用意されるべきである。

3 イベント列を理解すること

マルチモーダルインタフェースが扱う表現の処理の複雑さを例を使って示す。マルチメディア辞書があり、入力方法には、音声認識（自然言語処理）とマウスがあるとする。利用者が "Can this do this?" と、言いながら、最初の「この」のときに画面上の絵をマウスでクリックし、つぎに、2番目の「この」のときに、メニュー項目から1つを選択したとする。このシステムは、例えば、最初のマウス入力は特定の動物の絵のことであり、2番目の入力は、メニュー中の「fly」という項目を指していると解釈し、

データベースへの問い合わせ式を生成しなくてはならない。

オブジェクト指向の考え方でこれを処理しようとすると、中心的なオブジェクトがマルチモーダル統合処理を受持ち、画面、マウス、そして音声認識を処理するオブジェクトに、入力処理結果を渡すよう陽に指示するメッセージを送らなくてはならない。マルチモーダルシステムでは、しばしばモードの種類や数が増減するし、それに従って利用者が入力可能なマルチモーダル表現も変化する。このようなシステムでは、マルチモーダル入力を柔軟に取りだしたり、組み合わせることを簡単に扱えないため、入力モードの変化に対して迅速に対応するのは困難である。我々が提案する計算メタファは、オブジェクト指向ではなく、自然言語の文法定義とその処理のようにイベント列を処理する。

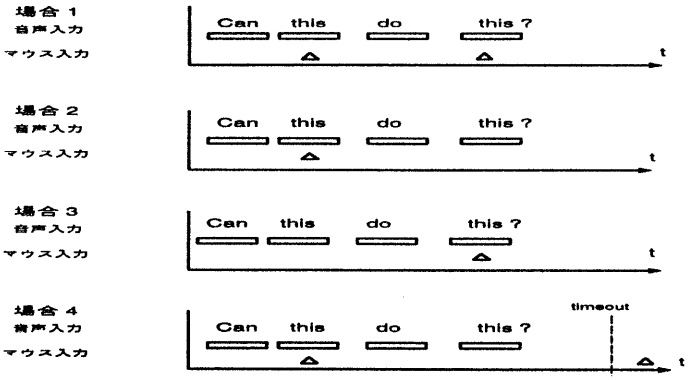

図 1: “can this do this” を表現する 4 つの入力タイミング

上の問い合わせ例 “Can this do this” を使って、マルチモーダル入力が複雑に処理される様子を説明する。図 1は、入力タイミングを 4 つの場合に分けて、 タイムチャートで表現したものだが、文脈や入力タイミングによってその解釈方法はそれぞれ異なる。

- 場合 1 は、 マウス入力が 2 つで、 しかもそれぞれマウス入力を対応する音声入力と同期して入力した場合である。 この場合は、 自然言語発話とマウス入力は簡単に対応づけられる。
- 場合 2 は、 マウス入力が 1 つで、 それが特定の動物を指示した場合である。マウス入力は、 組み込む先を示したものとして解釈され、 1 番目の “this” と対応づけられる。そして、 2 番目の “this” については、 文脈情報を使って直前に参照された動作を対応させる。
- 場合 3 は、 マウス入力が 1 つで、 それが特定の動作を指示した場合である。マウス入力は、 特定の動作を示したものとして解釈され、 2 番目の “this” と対応づけられる。そして、 1 番目の “this” については、 文脈情報を使って直前に参照された動物を対応させる。
- 場合 4 は、 マウス入力は 2 つだが、 2 回目のマウス入力を発話の (例えば)3 分後に行なった場合である。この場合、 2 回目のマウス入力を 2 番目の “this” と対応させるのは不自然であり、文法ルール中に前もって定義しておいたタイムアウトを過ぎているとみなされる。それゆえ、 2 回目のマウス入力は無効となり、マウス入力が 1 つしかなかったものとして、 場合 2 または場合 3 と同様に解釈される。

この例からわかるように、マルチモーダル入力を解釈するには、文法記述形式中に複数のモード入力の処理や時刻情報の処理を統合し、文脈情報と組み合わせて全体の解釈ができる必要がある。

4 MMM の設計手順

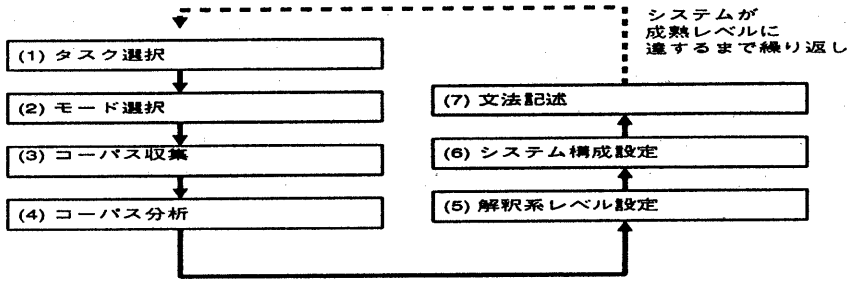

図 2: Multi-Modal-Method 設計の 7 ステップ

図2に示すように、MMM は、7つのステップからなる。

(1) タスク選択: マルチモーダルインタフェースに向いている分野と向いていない分野が存在する。向いている分野として比較的多くの研究が既になされている分野としては、デザイン / 編集、プレゼンテーション、情報検索、質疑応答 / 案内、遠隔教育システム、等がある。

(2) モードとメディアの選択: 一般的には、モードとメディアは 1 対 1 にはならない。メディアが異なっていても、モードとしての扱いは同じ場合がある。例えば、音声入力とキーボード入力は、メディアは、異なるが、モードの解釈処理は同一に処理する応用もある。

(3) コーパス収集: 自然言語処理システムの設計と同様、どのような対話が可能かを調べるため、マルチモーダルなコーパスを収集する。

(4) コーパス分析: 収集したコーパス中の個々の表現を、2 つの基準にもとづいて分析される。

経済性: その表現が、利用者の発話行為を本当に節約するかどうか調べられる。例えば、あるマルチモーダル表現が可能になることで、利用者の操作数や種類が少なくなれば、それは利用者の行為を節約したといえる。

有用性: その表現が、実際の利用形態の中で本当に使われるかどうかを調べられる。マルチモーダルインタフェースの文法記述は、シングルモーダルインタフェースの文法記述よりもはるかに多くの文法ルールを定義しなくてはならないので、非常に頻繁に使われるマルチモーダル表現のみが選択されるべきである。

ここで選択されたマルチモーダル表現が、設計中のマルチモーダルシステムの機能仕様の種となる。

(5) 解釈系レベル設定: レベルは 5 段階あり、[Suenaga et. al., 1992]らの分類と一致している。

レベル 1: シングルモード入力: マルチモーダルシステムといえども、利用者はいつもマルチモーダル表現を使うわけではなく、シングルモーダルで表現する事もある。

レベル 2: すべてのモードの入力が同一内容を示す場合: 異なるモードの入力がそれぞれ自己完結型に同一の内容を表現する場合。例えば、ある四角形を差し示したあとにメニューの「削除」を選択し、それと同時に、「その四角形を削除せよ」と発話する場合。

レベル3: 個々のモード入力は不完全だが、補間しあって全体として意味をなす場合: 個々のモードの入力からは部分的な意味しか解釈できないが、それらを持ち寄って合せると一意の意味をなす場合。例えば、ある物体を差し示しながら、同時に「削除」と発話する場合。

レベル4: 個々のモード入力解釈結果が矛盾する内容の場合: 個々のモードの入力解釈から生成された意味が相互に矛盾する場合であり、例えば、「その丸を削除せよ」と発話しながら、その丸の上に配置されている四角形を差し示す場合である。このような矛盾は、しばしば文脈解析結果を利用して解決される。

レベル5: すべてのモードの入力解釈を合せても意味を一意に決定できない場合: 例えば、「それをここに移動せよ」と発話しながら、特定の場所を差し示す場合。この場合、差し示された特定の場所は、「ここ」のことを意味し、「それ」で参照された物体は、直前に参照された物体を意味する。ここでも、解釈には、文脈解析の結果を利用しなくてはならない。

レベルが上がるにつれ、そのマルチモーダル表現の解釈が難しくなる。とりわけ、レベル4と5では、文脈解析との密接な統合が必要となるので、インタフェース設計者は、現在設計中のマルチモーダルシステムで真にレベル4、5が必要かどうかを熟考する必要がある。

(6) システム構成設定: 独立したモードの解釈処理を、独立のエージェントに割当てる事は容易なので、一般には、エージェント型のアーキテクチャになる。黒板モデルでは、黒板と呼ばれる共有メモリを介して情報交換するが、これは、文脈解析と統合して解釈しなくてはならない複雑なマルチモーダル表現を取り扱う必要な応用に向いている。サブサンプションアーキテクチャは、エージェント間の情報交換方法は限定されており、個々のエージェントが比較的独立に挙動し、比較的簡単で定型的なマルチモーダル表現の処理をする場合に向いている。

(7) 文法記述: 選択されたマルチモーダル表現は、対応する文法ルールを定義される。マルチモーダル表現を解釈する文法記述形式は、以下の機能を有する必要がある。

1. 文法ルール中で、すべてのモードを対等に扱えること: 従来の多くのマルチモーダルシステムでは、自然言語が主でそれ以外の入力、例えばマウス入力は従の関係だった。しかし、様々なマルチモーダル表現を解釈するためには、すべての入力モードは対等に扱われるべきである。
2. 文法ルール中で異なるモードの入力解釈の相互参照が可能であること: 個別のモードの入力は独立に解釈されるべきである。しかし、個々のモード間でその解釈を相互に参照できる必要がある。例えば、上の例における場合2と場合3では、異なるモードの部分的解釈を持ち寄り、文脈情報を参照して、全体の意味を決定している。
3. 文法ルール中で、時刻情報を扱えること: 入力間の到着時刻順や到着時間差は、解釈生成において重要である。例えば、図1の場合4のように、あるマウス入力の時刻がタイムアウトを越えるか否かで、文全体の解釈が異なることもある。

MM-DCGは、これらの要求を満足するように設計された。MM-DCGは、広く使われている自然言語処理の文法記述形式であるDCG (Definite Clause Grammar) [Pereira and Warren, 1980]の拡張であり、DCGの特徴をすべて引き継いでいる。MM-DCGの文法記述は、DCG同様、簡単な変換プログラムによってPrologの述語に1対1に翻訳される。MM-DCGがDCGに加えた拡張は、以下の3つである。

(1) 個別のモードごとに独立したストリームを割り当てたこと

DCGが1つの入力ストリームのみを受け付けるのに対し、MM-DCGでは、任意数の入力ストリームを受け付ける。個々のモードには、独立したストリームが対応づけられる。従って、1つの文法ルール中に、異なるモードの入力から構成される文法カテゴリを混在させ、それらを組み合わせた記述が可能である。

(2) 文法ルール中の文法カテゴリの生成時間を自動的に算出すること

個々の入力は、それぞれ入力開始時刻と終了時刻をタイムスタンプとして付与する事を要求されている。このタイムスタンプ情報を使って、MM-DCGは、解析中に生成されるすべての意味カテゴリーの入力開始時刻と終了時刻を自動的に計算する。実際には、MM-DCGからPrologへ変換する翻訳プログラムが、それを計算するコードを自動的に生成し、挿入する。

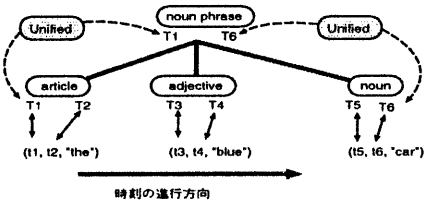

図 3: 生成された文法カテゴリの時刻計算

図3は、名詞句 (noun_phrase) 定義における時刻情報の引数間の関係を表している。例えば、"the blue car" という入力が、それぞれの開始時刻と終了時刻を伴って、[(t1, t2, "the"), (t3, t4, "blue"), (t5, t6, "car")] の形で入力された場合,T1 = t1, T2 = t2, ..., T6 = t6 と単一化され、noun_phrase は、t1 に入力が開始され、t6 に終了したと自動的に計算される。

(3) タイムアウトを表現できること

タイムアウトとは,ある入力とそれに続く入力の間の時間間隔に関する制約であり、時間間隔が文法ルール定義者の指定した間隔より大きくなれば,その入力ストリームは,まだ入力の残りがあるにもかかわらず,一時的に空であるとみなすことができる。

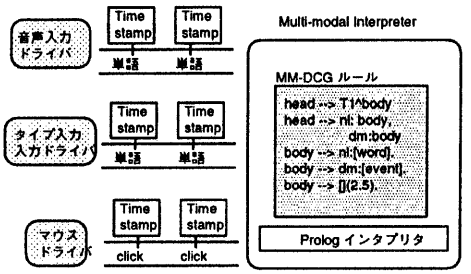

図 4: MM-DCG で記述したマルチモーダルシステム

マルチモーダルの文法定義作業について重要なことは、文法定義に必要な労力である。文法定義者は、マルチモーダルインタフェースの定義に必要な文法ルールの数は、シングルモーダルインタフェー

スに必要なルールの数に比べてはるかに多くのルールを定義する必要があることを理解しなくてはならない。もし、3つのモード M_1、 M_2、 and M_3、が存在し、個々のモードのみの入力解析に必要な文法ルールの数が G_1、 G_2、 G_3 で表される時、モード間の組合せを許すマルチモーダル文法ルールの数 G_{total} は、

$$G_1 + G_2 + G_3 \tag{1}$$

ではなく、個々のモードのすべての組合せごとの文法ルールの集合となる。それゆえ、マルチモーダル文法ルールの数 G_{total} は、

$$G_{total} = \sum_{M_1,M_2,M_3 \supseteq S} G_S \tag{2}$$

で表される。

以上の7ステップを繰り返すにつれて、開発するマルチモーダルシステムの性能が向上していく。システムの性能が充分成熟したレベルに達した時に、エンドユーザに提供される。

5 関連研究

本稿で述べたような複数のモードを組み合わせた入力を解釈するマルチモーダルインタフェース自体は、新しいものではない。既に1980年には、音声指示と動作情報を組み合わせた "Put-That-There" [Bolt, 1980] プロジェクトがこの分野の先駆けとなっている。マルチモーダルのアイデアは、その後、自然言語と指示入力の組合せの研究として発展した。例えば、Hayesらは、指示入力の解釈を自然言語処理における代名詞参照の問題 [Grosz, 1977] [Sidner, 1979] として扱った。最近のインタフェースの研究でも、様々なマルチモーダル入力を統合して解釈する研究は、数多く提案されている[Kobsa et. al., 1986] [Hayes, 1987] [Cohen et. al. , 1989] [Allgayer et. al., 1989] [Wahlster, 1989]。しかし、これらの多くは、マルチモーダル・インタフェースの応用システムに焦点があり、本稿で述べたようなマルチモーダルな入力を統合する文法記述形式については、多くは言及していない。

6 結論

本稿では、マルチモーダルシステム設計手順である Multi-Modal-Method (MMM) について述べた。MMM は、我々のマルチモーダルインタフェースシステムの開発経験をもとにして定義された。MMM は、1) タスク選択、2) モードとメディア選択、3) コーパス収集、4) コーパス分析、5) 解釈系レベル設定、6) システム構成設定、7) 文法記述、の7ステップの繰り返しから成る。MMM の文法記述のためには、マルチモーダル文法記述形式 Mutli-Modal Definite Clause Grammar (MM-DCG) をツールとして提供している。

参考文献

[Allgayer et. al., 1989] Allgayer, J., Jansen-Winkeln, R., reddig, C., and Reithing N.,

[Arita et. al., 1992b] Arita, S., Shimazu, H., and Takashima, Y., "Simple + Robust = Pragmatic: A Natural Language Query Processing Model for Card-type Databases", Proc. of the 13th Annual Conference of the Cognitive Science Society, 1992.

[Bellik and Teil, 1993] Bellik, Y. and Teil, D., "A Multimodal dialogue controller for multimodal user interface management system application: A multimodal window manager", Adjunct Proceedings of INTERCHI-93.

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

[Clocksin and Mellish, 1981] Clocksin, W.F. and Mellish, C.S., "Programming in Prolog", Springer-Verlag, 1981.

[Cohen et. al., 1989] Cohen, P.R., Dalrymple, M., Moran, D.B., Pereira, F.C.N., et al., "Synergistic Use of Direct Manipulation and Natural Language", Proc. of CHI-88, 1989.

[Cohen, 1991] Cohen, P.R., "The Role of Natural Language in a Multimodal Interface", 1991 International Symposium on Next Generation Human Interface, 1991.

[Grosz, 1977] Grosz, B. "The representation and use of focus in a system for understanding dialogs," *Proc. IJCAI 1977*, Boston, MA.

[Hayes, 1987] Hayes, P.J., "Steps towards Integrating natural Language and Graphical Interaction for Knowledge-based Systems", Advances in Artificial Intelligence - II, Elsevier Science Publishers, 1987.

[Hayes, 1988] Hayes, P.J., "Using A Knowledge Base To Drive An Expert System Interface With A Natural Language Component," in J. Hendler (ed.) *Expert Systems: The User Interface*, Ablex Publishing, 1988.

[Hiyoshi and Shimazu, 1994] Hiyoshi, M. and Shimazu, H., "Drawing Pictures with Natural Language and Direct Manipulation" Proc. of COLING-94, 1994.

[Kobsa et. al., 1986] Kobsa, A., Allgayer, J., Reddig, C., Reithing, Nl, Schumauks, D., Harbusch, K., and Wahlster, W, "Combining Deictic Gestures and Natural Language for Referent Identification", Proc. of COLING-86, 1986.

[Nigay and Coutaz, 1993] Nigay, L. and Coutaz, J., "A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion" Proc. of INTERCHI-93, 1993.

[Pereira and Warren, 1980] Pereira, F., and Warren, D.H.D., p"Definite Clause Grammars for Language Analysis - A survey of the Formalism and a Comparison with Augmented Transition Networks", Artificial Intelligence, vol. 13, no. 3, 1980.

[Shimazu et. al., 1992] Shimazu, H., Arita, S., and Takashima, Y., "Design Tool Combining Keyword Analyzer and Case-Based Parser for Developing Natural Language DataBase Interfaces", Proc. of COLING-92, 1992.

[Shimazu et. al., 1994] Shimazu, H., Arita, S., and Takashima, Y., "Multi-Modal Definite Clause Grammar" Proc. of COLING-94, 1994.

[Sidner, 1979] Sidner, C. *Towards a computational theory of definite anaphora comprehension in English Discourse.* TR-537, MIT AI Lab, Cambridge, Ma.

[Suenaga et. al., 1992] 末永、間瀬、福本、渡部, "Human Reader: 人物像と音声による知的インタフェース", 電子情報通信学会論文誌、 Vol. J75-D-II, NO 2, 1992

[Vo and Waibel, 1993] Vo, M.T., and WAibel, A., "A multi-modal human-computer interface: Combination of Gesture and Speech Recognition", Adjunct Proceedings of INTERCHI-93

[Wahlster, 1989] Wahlster, W., "User and discourse models for multimodal communication", in J.W. Sullivan and S.W. Tyler, editors, Intelligent User Interfaces, chapter3, ACM Press Frontiers Series, Addison Wesley Publishing, 1989.