マルチモーダル対話システムにおける複数モダリティの統合と解釈に関する一検討

賀川経夫† 阿部慎也† 遠藤勉†

小学校1年生の算数のドリルテキストを対象とした問題解決システムの構築に取り組んでいる．対話の際には，教師(人間)は音声発話とペンによるジェスチャ(動画像により獲得)という複数モダリティを利用する．本研究では，この複数のモダリティ間の相互補完的なインタラクションの実現に関する検討を行ない，意味レベルの統合について考察する．本稿では，対話中に行なわれる音声発話と操作を表現するためのマルチモーダル意味表現フレームを定義し，それを用いた音声／ジェスチャの解析手法に関して検討する．また，フレーム表現を利用することにより，動画像解析であるジェスチャの解析と音声認識処理の統合に関して検討を行なう．

A Method of Interpretation of Different Modalities in Multi-Modal DialogSystem

TSUNEO KAGAWA,† SHINYA ABE† and TSUTOMU ENDO†

Our aim of research is to build a problem solving system with dialogue. This system can deal with an arithmetic drill text of elementary school. When the system ask a question to solve a problem in the text, user answer the question using speech and gesture (with a pen motion). It is necessary to integrate each modality effectively to eliminate ambiguity in the both of a speech recoginition and a gesture analysys in our system. So we difinite a concept description *Action Concept* which represent manuplating figures in the drill text. This description is case structure. System can eliminate the meaning of user's utterance.

1. はじめに

通常の人間同士の対話においては，音声言語だけではなく，表情，身振り等の情報を用いてコミュニケーションを実現している．人間－コンピュータ間の対話においても，テキストのみではなく，グラフィックス，動画像，顔画像，自然言語，音声等といった複数のメディアを伝達手段として利用することにより，エラーや曖昧さを減少させ，対話の質の向上が期待できる．近年，このようなマルチモーダル対話システムに関する研究が多くなされている[1)~5)]．これらのメディアの解釈を行なうために，各々の解析処理において生じた曖昧性を相互補完するためのメディア間のインタラクションの実現が課題となる[6),7)]．

我々は，従来より，図と問題文が混合する小学校1年生の算数ドリルテキストを題材とし，問題解決能力の習得や知識の形成を行なうシステムの構築に取り組んでいる[8)~10)]．先に述べたように，身振り，音声，図形等の様々なメディアが教師－生徒間の対話において重要な役割を果たしていることを考慮し，ペンを用いたジェスチャと音声による自然言語を入力モダリティとするインタフェースに関する検討を行なっている．システムは教師（人間）からの音声発話とジェスチャの複合的な解釈処理を行ない，問題解決に必要な情報を抽出する必要がある．

本研究では，システムにおける問題解決コマンド処理との整合性を考慮し，フレームで記述された知識を用いて教師からのマルチモーダル入力の解釈処理を実現する．そこで，そのためにマルチモーダル意味表現を設定し，フレーム形式で記述する．このフレームは，本システムでの対話における各発話の意味的な中心をなす図形操作を中心として表現される．この意味表現フレームに対して音声発話解析処理とジェスチャ解析処理との双方から参照／書き込みを繰り返し行なうことにより，各モダリティ解釈における曖昧性の解消を行なっていく．すなわち，異種メディアの解釈における相互補完によるインタラクションの記号処理を用いた実現を試みる．

† 大分大学工学部
Faculty of Engineering, Oita University

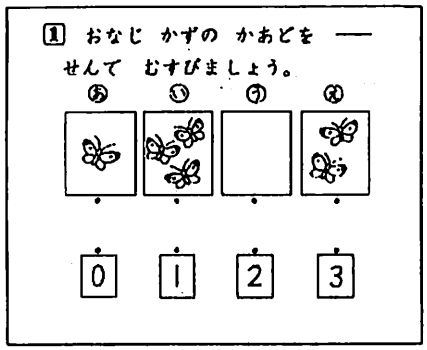

図1 ドリルテキストの例
Fig. 1 An example of drill texts

S1	:なんと おなじ かずですか.
T1	:この わくの なかの ちょうの かずです.
S2	:これを かぞえるのですね.
T2	:そうです.

図2 対話例
Fig. 2 A dialogue example

2. 問題解決システムの概要

2.1 コマンドによる問題解決処理過程

本システムにおける問題解決とは，"問題解決ならびに解答合成に必要なコマンド系列を組み立て実行する"ことであると考えている．各コマンドは，フレーム形式で記述されており，コマンドフレームの付加手続きが次々と実行されることで，対話に基づく問題解決処理が制御される．例として，図1に関する問題解決処理とその場合の対話処理の流れを考える．詳細は，文献8)~10)を参照されたい．

まず，問題文を含む問題図形(図形構造表現，ビットマップ)から，解答合成に必要なコマンド候補群が抽出される．

$$(eqn(かず,?), link(かあど,?))$$

ただし，$eqn(Num_1, Num_2)$ は，数 Num_1 と数 Num_2 が等しい事を表し，$link(Obj_1, Obj_2)$ は，図形オブジェクト Obj_1 と Obj_2 を線で結ぶ事を意味するコマンドである．? は未決定のパラメータである．

次に，コマンドの組立・実行の優先順位より eqn コマンドに着目する．計数の対象物である第2パラメータが未知であることから，「なんと おなじ かずですか?」と発話する制御が行なわれる．ここで，教師(人間)は，枠をペンで指示しながら，例えば，「このわくの なかの ちょうの かずです。」というような応答を行なう．

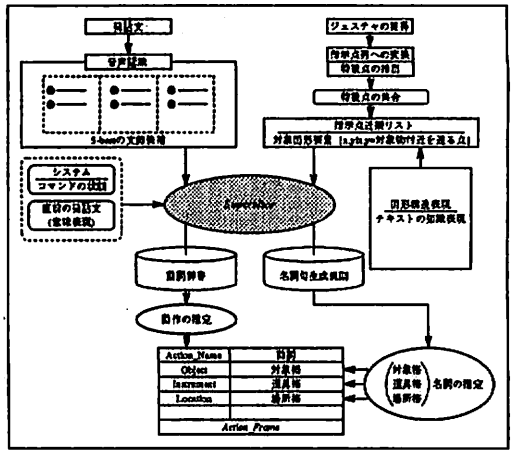

図3 マルチモーダル対話システム(入力解釈部)
Fig. 3 Multi-modal dialogue system (interpretation part)

表1 *Action Name*
Table 1 Action name

Action_Name	実際の操作	対応する動詞
Count	対象物を数える	「かぞえる」
Select	条件に合うものを選ぶ	「えらぶ」
		「みつける」
Write	対象物を描画する	「かく」
Link	対象物を線で結ぶ	「むすぶ」
		「せんをかく」
Enclose	対象物を囲む	「かこむ」
Paint	対象物を塗る	「ぬる」
		「ぬりつぶす」
Point	対象物を染める	指示操作

この応答に対する解析処理によりシステムが現在必要としている情報(計数の対象物)が獲得されたら，次のコマンドもしくはパラメータに着目し，問題解決処理が繰り返される．

以上のような処理により，未知のパラメータが全て決定された時に，最初に抽出されたコマンドが実行され，問題解決処理が終了する．この時の対話例の1部を図2に示す．

2.2 マルチモーダル対話システム

図3に教師からの入力を解釈する処理の概要を示す．解釈処理を行なう前に各々のモダリティに関して次に示す解析処理が行なわれる．

2.2.1 音声認識

音声認識においては，文節を単位とした処理が行なわれる．この処理結果として，各文節に対して最大5つの認識候補が得られる．認識の際の辞書においては，数種類のドリルテキストを用いた対話サンプルから出現が予想される文節の収集と分類を行うことにより構築を行なった．

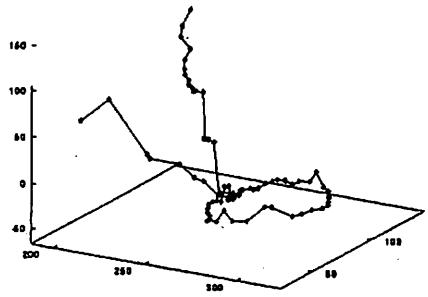

図4 ペン先の移動軌跡
Fig. 4 A trajectory of pen motion

認識辞書や発話文解析用辞書は，以下に挙げる品詞の組合せにより生成可能な文節によって構成される．

- 動詞:(表1に示す動詞)
- 名詞:(操作の対象となる具体物名，現場の事物名，概念名，位置関係を表現する名詞等)
- 指示代名詞:(ここ，これ，これら)
- 指示詞: (この，これらの，これらの)
- 助詞:(で，と，に，の，を)
- 助動詞:(です，だ，でない等)
- 形容詞:(おおきい，ちいさい等)

2.2.2 ジェスチャ認識

実際のテキストを対象とした対話において，教師は，ペン先を図形に近付けることで操作対象を明確にし，ペンを移動させることにより各種の操作を表現する．そこで，ペン先の動き，すなわち移動軌跡を用いてジェスチャの解析が行なわれる．

ペンの動きは，カメラを利用して連続画像データ (毎秒30フレーム) として獲得され，各フレーム画像からペン先の座標が検出される．また，本手法では，鏡を利用することにより，ペン先の3次元座標が得られる[8]．

次に，得られた点列を射影処理によりシステムの持つテキスト画像との対応づけが行なわれる．投影された各点を以後，指示点と呼ぶ．また，ジェスチャの始点と終点は，それぞれ，ペン先がテキスト上に出現した時，消失した時とし，出現ー消失間の指示点列が処理単位として扱われる．これらの移動軌跡，指示点を得る処理に関しては，文献8) に詳しく述べる．移動軌跡の例を図4に示す．

2.3 図形構造表現

問題図形は，図形構造表現と呼ばれる形式により記述され，テキスト中のの全ての図形要素と要素間の構造的関係が記述される．

$p\_form(INST, CLASS, LBL, STRUCT)$.

マルチモーダル意味表現 [[No, フレーム番号],[[
[*Agent*, 発話者 (システム／教師)],
[*Intention*, 発話意図]],
[*Action_Frame*,*Action_Frame*],
[*Command*, 着目されたコマンド]]].

図5 マルチモーダル意味表現
Fig. 5 Multi-modal meaning description

Action_Frame [[[*Action_Name*, 操作概念名],[
[*Object*, *Thing_Name*,*constraint*],
[*Instrument*,*Thing_Name*,*constraint*],
[*Location*, *Thing_Name*,*constraint*]]]]

図6 Action Frame
Fig. 6 Action frame

Thing_Frame [[*Thing_Name*, フレーム名],
[*class*, クラス名],[*instance*, インスタンス名],
[*word*, 名詞],[*indic*, 指示代名詞],
[$inside^+$, 図形名],[$inside^-$, 図形名]]

図7 Thing Frame
Fig. 7 Thing frame

$INST$, $CLASS$ は，図形要素のそれぞれインスタンス名，クラス名を表す．$STRUCT$ は，部分図形間の特徴的な空間的位置関係を表す．

また，図形構造表現と実際のテキスト中の線や点等の図形要素との対応づけを行なうために，各図形要素にラベルが付けられたラベル画像が利用される．図形構造表現中の LBL は，対応する図形要素に付与されたラベルである．

3. マルチモーダル意味表現

それぞれの入力モダリティ (音声とジェスチャ) に対して意味情報による統合／解釈処理を行なうため，マルチモーダル意味表現と呼ぶフレーム記述形式を導入する．この意味表現は，基本的に図5の形式に従って記述される．*Intention*(発話意図) は，システムの問題解決処理により，“情報の要求／提供”，“確認の要求／提供”，“(不) 同意表示” 等が記述される．

Action_Frame は，図6に示される格構造を用いて記述される．現在，検討を行なっている *Action_Name* を表1に示す．それぞれの格ラベルは，操作の対象物 (*Object*)，手段 (*Instrument*)，場所 (*Location*) を表す．各々の格要素として記述される *Thing_Name* は，“もの” を表す *Thing_Frame* の名前である．未知パラメータに対しては，*Thing_Name* を unknown とし，

優先的な決定が行なわれる．*constraint* は，格要素を決定する際の制約条件を表すものである．

Thing_Frame は，図 7 の形式により記述される．*class*，*instance* は，それぞれ“もの”のクラス名とインスタンス名を表し，辞書と図形構造表現を参照することにより決定される．*instance* には，*p_form* 名が記述される．*word* は，発話文中に出現した名詞であり，*Thing_Frame* に対応する．スロットには，他に *indic*(指示詞による修飾) や $inside^+$，$inside^-$ (図形間の包含関係) 等がある．

4. ジェスチャの解析

4.1 特徴点の抽出

ジェスチャ (テキストに関する指示点) は，以下のように得られる．

$$point(t) = \{\ (x, y, z), rgn, type\ \}$$

ただし，x，y，z は指示点の座標である．

rgn は，囲まれた領域の番号を表す．指示点によって囲まれた領域は「かこむ」操作や複数の図形を指示する操作において抽出される事が多い．この領域は，テキスト表面 (z < 一定値) で一旦離れて再び近付く指示点の組を検出し，その間の全ての指示点を含む最小の長方形として得られる．この領域に番号を付け，領域を構成する全ての指示点の rgn にその番号を設定する．

type は各指示点の特徴を表す．表 1 の各操作の判別を行なうために，実際のジェスチャのサンプルを分析した結果から，(1) 停留 (SP)，(2) ペンアップ (UP) ／ペンダウン (DP) という特徴を用いる．それぞれの特徴に関して以下に述べる．

(1) 停留 (SP)：ペン先が特定の点の付近で留まる状態を表す．指示操作の際に行なわれる．直前の指示点からの変化量 (距離) が閾値よりも小さい指示点の一定以上の個数である集合に対してその *type* を SP とする．

(2) ペンアップ (UP) ／ペンダウン (DP)：ペン先がテキスト表面に接近したり，離れたりする状態を表す．「むすぶ」操作の起点や終点，叩いて行なわれる指示等の場合に行なわれる．ペンのテキストからの高さ z の変化量の絶対値が閾値以上である場合，変化量の正負により，*type* を UP もしくは DP とする．

4.2 指示点近接リスト

指示点に対する特徴抽出後，各指示点と図形要素との対応付けにより，各図形要素の近傍の指示点の分布を示す指示点近接リストが作成される．このリストは，操作の対象や操作が行なわれた場所を判別する上で重要なものとなる．リストの作成は，以下のように行なわれる．

[1] 各図形要素の外接長方形を求め，その中に $rgn = 0$ である指示点を含むものを求める．囲み領域が存在する場合は，その領域内にある図形要素を *class* 別に分類して列挙する．この時，これらの指示点は，その図形要素に近接しているとみなす．

[2] 図形要素を基準として以下の形式で時系列順にリストを作成する．

図形要素名:(始点 t_1，終点 t_2):*rgn*:*score*

[3] 各図形要素に近接する指示点の *type* を用いて *score* を計算する．現在は，特徴点の個数をそのまま用いる．

5. 解釈処理

システムの「なにを かぞえるのですか？」という発話に対し，教師が「このわくの なかの ちょうを かぞえます」と応答した場合の例を用いながら解釈処理に関して述べる．

システムの発話文「なにを かぞえるのですか？」の生成時のマルチモーダル意味表現を以下に示す．

マルチモーダル意味表現 [[*No*,*n*],[[
[*Agent*, システム],[*Intention*, 情報要求]],
[*Action_Frame*,*Count*],
[*Command*,*count*(?,?)]]].
Action_Frame [[[*Action_Name*,*Count*],[
[*Object*,unknown,$constraint_n$]]]]

応答文「わくの なかの ちょうを かぞえます」に対して，音声認識の結果は表 2 のように得られた．

5.1 述語の決定

認識結果の 5-best の文節候補から，述語となる文節の決定を行なう．そこで，まず，最終文節 (文節 5) に着目し，システムの発話意図に応じて以下のような処理が行なわれる．着目した文節での選択に失敗した時は，直前の文節に着目して同様のことが行なわれる．

(1) 情報要求：教師はその情報を提供する応答を行うと推定される．[1] 動詞句，[2] 図形クラス名，指示代名詞 (+ 助動詞) を含む文節候補のうち 5-best の最上位にあるものが述語として選択される．

(2) 確認要求：教師は同意／不同意の発話を行なうと推定される．「そうです」等の同意や「ちがいます」等の不同意を表現する語の検出が行なわれる．

(3) 言い直し要求：システムに蓄積された発話履歴を参照し，言い直しを要求する発話の発話意図に応じて (1) または (2) の処理が行なわれる．

例の場合は，システムの発話意図が“情報要求”であることより，第 5 文節の「かぞえます」(最上位) が選択される．動詞辞書により，動詞と *Action_Name* の

表2 音声認識の結果(5-best)
Table 2 Voice recognition result

候補	文節1	文節2	文節3	文節4	文節5
音声	この	わくの	なかの	ちょうを	かぞえます
	この	わくの	なかの	いろを	かぞえます
	こまを	なかまを	なかまを	とどを	つばめの
	ねこの	なかまの	やかんの	よっとを	かずの
	だいこんを	さかなを	ぶたの	ちょう	はなの
	だいこんの	こっぷの	やかんを	ねこを	かぞえません

対応づけが行なわれ，解釈のためのマルチモーダル意味表現が生成される．

マルチモーダル意味表現 [[No,n+1],[[
[$Agent$, 教師],[$Intention$, 情報提供]],
[$Action\_Frame$,$Count$]]].
$Action\_Frame$ [[[$Action\_Name$,$Count$],[
[$Object$,unknown,$constraint_n$]]]]

5.2 格要素の決定

$Action\_Frame$ の格要素とそれに関する修飾語の推定が行なわれる．そこで，名詞辞書と名詞句生成規則が利用される．

名詞辞書は，名詞とその意味属性が意味素性を用いて記述されたものである．また，発話文中の名詞と図形構造表現中の図形要素との対応づけにも利用される．例えば，「ちょう」に関する記述は以下の通りである．

(名詞) ちょう:(クラス名)$butterfly$：
(図形操作関係属性):$+Count\_obj$,$+Select\_obj$,...

名詞句生成規則は，「この ちょう」や「わくの なかの ちょう」等の名詞間の関係を抽出するための規則であり，以下の形式により記述される．

(対象となる名詞 [の属性],(助詞,
(意味的に接続可能な名詞, 代名詞 [の属性])),
(操作))

格要素の選択とそれに関する修飾関係の推定は以下のように行なわれる．
[1] 述語が選択された文節より前にある全文節候補から，格要素の $constraint$ と合致する名詞を含む文節が検出される．
[2] 選択された格要素候補に対して名詞句生成規則が適用され，文節間の修飾関係が抽出される．

例の場合，文節4に関して「いろを」以外全て「かぞえる」の対象の候補として選択される．ただし，$constarint$ は，次のようなものである．

表3 指示点近接リスト
Table 3 Points adjacent objects list

図形要素名	(t_1,t_2)	rgn	$score$
p_form(中黒点3)	(20,23)	0	1
p_form(中黒点7)	(23,27)	0	1
p_form(番号2)	(30,33)	0	1
p_form(番号3)	(35,37)	0	1
p_form(枠2)	(40,45)	0	1
p_form(枠3)	(44,48)	0	5
p_form(枠4)	(49,65)	1	8
p_form(枠7)	(70,84)	0	5
p_form(枠8)	(85,92)	0	3

($+Count\_obj \cap$「名詞(を)」)

5.3 指示点近接リスト／ジェスチャの利用

次に，ジェスチャ解析によって得られた指示点近接リストを参照し，名詞と図形構造表現との対応づけが行なわれる．この処理により，音声発話の解析で得られた対象が実際の指示対象と同定される．例の場合(図4)の指示点近接リストを表3に示す．

$score$ の閾値を2とすると，ジェスチャ解析により抽出された対象図形要素の $class$ は，$rectangle$(わく)，$butterfly$(ちょう)，$number$(かず) となる．ただし，ちょうとかずは，枠の中にあるために選択されている．ここで，上記の文節4の候補のうち，「ちょうを」(「ちょう」は統一される)が選択され，「ちょう」に相当する $Thing\_Frame$ が生成される．他の候補に関しては，この時点で棄却される．

抽出された文節4より前に未処理の文節が残っているので，名詞句生成規則を参照し解析が続行される．文節3からは，「なかの」のみが選出される．名詞句生成規則の適用により，文節2から「わくの」が抽出される．「なかの」に関しては三項関係の規則が適用される．

(「なかの」,(「の」 ($+inside \cup +indic$),
(「」,($Count\_obj \cap Select\_obj$ ...))),
($Thing\_Frame$ を生成))

次に「わくの」に関して，文節1から「この」が選択される．適用された規則により，「わく」の $Thing\_Frame$ が生成され，$inside^+$ に「ちょう」

マルチモーダル意味表現 [[No,n],[[
[$Agent$, 教師],[$Intention$, 情報提供]],
[$Action\_Frame$, $Count$],
[$Command$,$count$(?,?)]]].
$Action\_Frame$[[no,0],[[$Action$,$Count$],[
[$Object$,$thing_4$,$const$]]]].
$Thing\_Frame$[[$Thing\_Name$,$thing_4$],
[$word$, 「ちょう」]].
$Thing\_Frame$[[$Thing\_Name$,$thing_5$],
[$word$, 「わく」],[$inside$,$thing_4$],
[$indic$, 「この」]].

図8 解釈結果
Fig. 8 Interpretation

の $Thing\_Nrame$ 名が記述される (「ちょう」を表す $Thing\_Frame$ の $inside^-$ も同様).

ここで，未処理の文節が無くなったので解析処理が終了する．最終的な解釈結果を図8に示す．

6. 検　　討

本手法において，例えば，言語のみを用いた対話システムと比較して，利用者の発話が(人間どうしの対話に近い形で)簡潔に表現可能である，各モダリティの曖昧性の解消が容易となる等の効果が得られた．フレーム形式の意味表現を用いた統合／解釈処理を行なうことにより，音声，ジェスチャ(動画像)といった異種メディアの解釈処理における両者のインタラクションの実現が容易となった．また，記号表現を利用することにより知識の体系化された記述とそれによる拡張性の向上も期待される．その反面，タスクや領域に大きく依存することになるため，一般的な利用が困難となるといった欠点も存在する．本稿での解釈手法は，タスクに大きく依存しており，ドリルテキストの変化には対応できるが，他の対象への適用が困難である．

7. おわりに

マルチモーダル対話システムにおける異種モダリティ間の意味表現を用いた統合と解釈のメカニズムに関して述べた．実際には，指示操作の対応づけだけではなく，「むすぶ」操作等に関しても解釈処理の実現も行なっている．

現在は処理過程において利用される様々な背景知識，特に教師の状態を推定するためのユーザモデル，文脈情報として利用するための対話モデル(対話履歴，スクリプト等)を明確にし，そのモデルの構築に取り組んでいる．また，本システムで利用される知識について意味ネットワーク等を用いた体系化の検討も行なっている．

本研究では，入力に対する解釈処理のみではなく，システムからの出力に関しても検討中である．システムからの出力としては，音声発話とCGアニメーションによるジェスチャが生成されるが，ここにも同様にマルチモーダル意味表現を導入し，入出力間のインタラクションの検討を行なっていく．

今後の課題として，(1)本手法の整合性及び出力を含む本システムの評価法の確立．(2)文脈情報や背景知識の明確な記述．(3)音声認識，ジェスチャ解析の精度の向上．等が挙げられる．

参 考 文 献

1) 速水, 竹澤: マルチモーダル情報統合システムの研究動向, 人工知能学会誌,Vol.13 No.2 (1998)
2) 河野, 屋野, 池田, 知野, 鈴木, 金澤: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, 人工知能学会誌,Vol.13 No.2 (1998).
3) 神尾,松浦,正井,新田:マルチモーダル対話システムMulticsDial, 信学論,Vol.J77-D-II,No.8 (1994).
4) J.W.Sullivan,S.W.Tyler:Intelligent User Interfaces,ACM Press,(1991).
5) M.T.Maybury:Intelligent Multi-Media Interfaces,AAAI Press (1993).
6) 田村編: ヒューマンインタフェース, オーム社 (1998).
7) 平川, 安村:bit別冊ビジュアルインタフェース, 共立出版,(1996).
8) Endo, T.and Kagawa,T.:Cooperative Understanding of Utterances and Gestures in Dialogue-Based Problem Solving System, Proc. the Conference of PACLING, pp.113-123, (1997).
9) 賀川, 高津, 尾上, 遠藤: マルチモーダル対話を用いた問題解決システムにおける音声発話文解析処理, 信学技報 TL98 (1998).
10) 賀川, 阿部, 吉岡, 遠藤: マルチモーダル対話のおける発話文との統合を考慮したジェスチャ解析に関する一考察, 信学技報 PRMU98(1998).
11) P.Cohen : Natural Language Techniques for Multi-modal Interaction , Trans. of IEICE,Vol.J77-D-II,No.8 (1994).
12) R.Bolt : The Integrated Multi-Modal Interface,Trans. of IEICE,Vol.J70-D,No.11 (1987).
13) J.Neal,Z.Dobes,K.Bettinger and J.Byoun : Multi-Modal References in Human-Comouter Dialogue,Proc. of AAAI,Vol.2 (1988).
14) 安藤, 北原, 畑岡：インテリアデザイン支援システムを対象としたマルチモーダルインタフェースの評価, 信学論,Vol.J77-D-II,No.8 (1994).
15) 小1算数5分間トレーニング, 教学研究社 (1990).