

# 音声とポインティング・ジェスチャの統合意味解析

水梨 豪 ローケン・キム キュンホ 森元 逞

ATR音声翻訳通信研究所

{mizu, kyungho, morimoto}@itl.atr.co.jp






## 1 はじめに

マルチモーダル・インターフェースは、次世代のマン・マシン・インターフェースとして注目を集めており、それに関してはこれまでさまざまな研究や開発が行なわれている[1]。そこでの課題のひとつに、入力されたさまざまなモダリティ間の意味的な対応をどのようにとるかという事項がある。本稿では、音声入力とタッチスクリーン上のポインティング・ジェスチャ（以下、単にジェスチャと呼ぶ）入力を受け付けるマルチモーダル・インターフェースを取り上げ、そのふたつの入力の意味的な対応をとり、両者の統合的な意味を同定する手法について述べる。さらに、このマルチモーダル・インターフェースを利用した問い合わせシステムを簡単に紹介する。

## 2 マルチモーダル環境におけるジェスチャ

われわれはこれまで、マルチモーダル環境における対話実験を幾度か重ね、そこで用いられた発話とジェスチャの収集と分析を行ってきた[2]。その結果、地図を用いた対話で使用される主なジェスチャの種別とそれらの主な指示内容(表1)が明らかになった。

表1 ジェスチャの種別と指示内容

種別	形状	指示内容
サークリング		もの
ポインティング		もの
マーキング		位置
スクランプリング		位置
ラインドラッキング		経路

## 3 発話とジェスチャの統合的な意味

これらのジェスチャは、ある特定の語句（主に指示語を含む語句）とともに使用され、その語句（以下、ジェスチャ対応語句と呼ぶ）に対して指示内容を意味的に補填する形で対応し、この結果、発話とジェスチャ両者の統合的な意味が同定される。例として、「このホテルがいちばん安いです」という発話とともに地図上のXホテルを丸で囲む場合を考える。最終的な「Xホテルがいちばん安い」という意味は、そのサークリング・ジェスチャがXホテルという「もの」を指示しており、さらにその指示内容は「このホテル」に対して補填されるということがわかって初めて同定される。「このように行ってください」という発話と、経路を表わすラインドラッキング・ジェスチャの組み合わせの場合も同様で、「このように」という語の意味内容として、「ジェスチャの線で表わされる具体的な経路」という指示内容が補填される。以上からわかるように、発話とジェスチャの統合的な意味を同定するためには、

- 1) ジェスチャの指示内容の同定
  - 2) ジェスチャが対応する発話中の語句の同定
- という作業が重要である。

## 4 ジェスチャと発話の意味統合システム

前章で述べた発話とジェスチャとの意味統合の枠組みを実装した。音声は、マイクを通して、ジェスチャはタッチパネルを通して入力される。システム構成を図1に示し、各モジュールについて説明する。

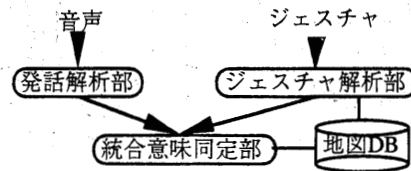


図1 システム構成

### 4.1 発話解析部

発話解析部では、音声認識[3]、言語解析[4]を行い、図2のような発話の意味表現を出力する。現在、道案内に関する約50タイプの発話を解析できる状態である。これらの発話は主に、実験で収集した、ものの名前や道順、ホテルの予約状況などを尋ねる発話からなる。

```
[[RELN *CONFIRMATION-QUESTION*]
```

```
[AGEN *SPEAKER*]
```

```
[RECP *HEARER*]
```

```
[OBJE [[RELN *行く*]
```

```
[SUBJ *HEARER*]
```

```
[DEST [[RELN *ホテル*]
```

```
[RESTR [[RELN *この*]]]]]
```

```
[MANN [[RELN *このように*]]]]]
```

図2 「このホテルへはこのように行くんですね」の意味表現

### 4.2 地図データベース

ユーザがジェスチャを行なう地図としては、京都市内の道路地図(図3はその一部分)を用いた。この地図上にある建物や道路などの「もの」(以下、オブジェクトと呼ぶ)に関する情報は、地図データベースに記述してある。地図データベース中のオブジェクトに関する主な情報は、1)ID番号、2)名称(左京区、京都駅、都ホテル、三条通り…)、3)地図上の位置(矩型の対角点の座標)、4)種別(地域、駅、寺、ホテル、道路…)である。

### 4.3 ジェスチャ解析部

#### 4.3.1 ジェスチャの指示内容の表現

2章で述べたジェスチャの指示内容を、次のように表現する。

##### 1) 「オブジェクト」の指示

```
[MOBJ [[ID n1][ID n2]…]]
```

n1などはオブジェクトのID番号

##### 2) 「位置」の指示

```
[MPART [[ID n][MAREA [[X1 x1][Y1 y1][X2 x2][Y2 y2]]]]]
```

(x1,y1),(x2,y2)はサークリングジェスチャの外接矩型の対角点で、全体として「IDがnのオブジェクトのうち、サークリング・ジェスチャで囲まれた領域」を示す。

##### 3) 「経路」の指示

```
[[START [MOBJ n1]] #経路の起点
```

```
[MOVE [[FROM [MOBJ n2]]
```

```
[TO [MOBJ n3]]
```

```
[ALONG [MOBJ n4]] #経路
```

```
[PASS [MOBJ n5]] #経路上のオブジェクト
```

```
[TURN [[AT [MOBJ n3]] #進路の転換点
```

```
[DIRECTION 270]] #進路の転換方向(角度)
```

```
[MOVE [[FROM [MOBJ n3]]
```

```
[TO [MOBJ n6]]
```

```
[ALONG [MOBJ n7]]]
```

```
[END [MOBJ n6]]] #経路の終点
```

#### 4.3.2 ジェスチャの指示内容の同定

ジェスチャの形状を認識[5]した後、サークリング・ジェスチャ

ならば、サークルに内包されている、または触れているオブジェクトを抽出し、それらのオブジェクトに関連した指示内容（「オブジェクトそのもの」が指示されているのか、「オブジェクトの中の位置」が指示されているのか、あるいはその両方か）をすべて生成する。

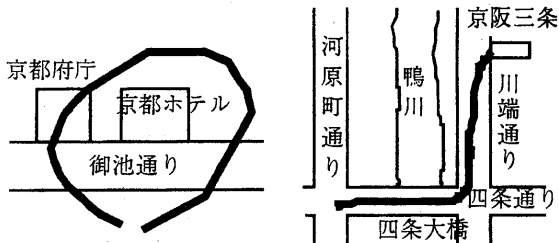


図3 サークリングとラインドラッキング

たとえば、図3のジェスチャの場合、  
 [MOBJ [[ID n 京都ホテル][ID n 京都府庁]]  
 [MOBJ [[ID n 御池通り]]  
 [MOBJ [[ID n 中京区]]  
 [MOBJ [[ID n 地図]]  
 [MPART [[ID n 中京区][MAREA [[X1 x1][Y1 y1][X2 x2][Y2 y2]]]]  
 [MPART [[ID n 地図][MAREA [[X1 x1][Y1 y1][X2 x2][Y2 y2]]]]  
 という指示内容の候補が生成される。ここで、「n 京都ホテル」などは、「京都ホテル」のID番号を示す。このあと、統合意味同定部で発話との関連を考慮してこれら候補から最終的に指示内容が同定される。ポインティングなどの他のジェスチャの場合も同様に処理される。

ラインドラッキングの場合は、線の曲折点でジェスチャを断片化し、それぞれの部分に対して始終点、経路、通過点を抽出し、次のような指示内容を生成する。  
 [[START [MOBJ n 四条河原町]]  
 [MOVE [[FROM [MOBJ n 四条河原町]]  
 [TO [MOBJ n 四条川端]]  
 [ALONG [MOBJ n 四条通り]]  
 [PASS [MOBJ n 四条大橋]]  
 [TURN [[AT [MOBJ n 四条川端]]  
 [DIRECTION 270]]  
 [MOVE [[FROM [MOBJ n 四条川端]]  
 [TO [MOBJ n 京阪三条]]  
 [ALONG [MOBJ n 川端通り]]  
 [PASS NULL]]  
 [END [MOBJ n 京阪三条]]

#### 4.4 統合意味同定部

##### 4.4.1 サークリング・ジェスチャ

発話の意味表現から、ジェスチャ対応語句になりうる語句、すなわち、命題内容の述語の引数中の、指示語を含む語句（図2では「このホテル」と「このように」）や固有/普通名詞を抽出する。次に、それらの語句のそれぞれについて、4.3.2で述べたようなサークル・ジェスチャの各指示内容候補との関連尤度を計算する。今回は、指示語の有無、語句と指示内容の種別の一致度、語句が対応すべき指示内容の形式（「オブジェクト(MOBJ)」か「位置(MPART)」かなど）の一致度をもとに関連尤度を計算し、それがもっとも高い語句と指示内容の組を、それぞれジェスチャ対応語句とジェスチャ指示内容として認定した。図2の発話例と4.3.2のサークル・ジェスチャ指示内容候補例の場合、「このホテル」がジェスチャ対応語句となり、指示語「この」があること、「ホテル」という種別が一致すること、また、「ホテル」はMPARTではなくMOBJに対応するのが最も適切ということ considering、最終的に、「このホテル」がジェスチャ対応語句、[MOBJ [[ID n 京都ホテル]]]がジェスチャ指示内容として認定される。ポインティング、マーキングなども同様の処理を行う。

##### 4.4.2 ラインドラッキング・ジェスチャ

発話意味表現中の命題内容の述語（図2ならば「行く」）と、4.3.2で述べた、断片化されたラインドラッキング・ジェスチャの指示内容中のラベル(START, MOVE, TURN, END)との関連性、ならびに、命題内容の述語の引数（図2ならば「ホテル」）と指示内容表現中のラベルの引数(FROM, TO, ALONG, AT, DIRECTION など)との関連性をもとに尤度を計算し、発話意味表現とジェスチャ指示内容の対応を同定する。具体的には、「行く、歩く、進む」のDESTの値（マテ格）は「MOVE」のTOの値に、また、「曲がる」のLOCTの値（テ格、ヲ格）は「TURN」のATの値に、最も対応する可能性が高いというルールのもとで、対応関係を同定する。

「ここで左に曲がって、川端通りを真っ直ぐ進め」という発話と4.3.2のジェスチャ指示内容との対応同定を例にとると、発話情報中の「曲がる[LOCT \*ここ\*][OBJE \*左\*]」に対応するものとして、ジェスチャ指示内容の「TURN[AT 四条川端][DIRECTION 270]」との対応尤度が最も高くなるので、「ここ」が「四条川端」と同定される。同様に、次の「進む」に関しても、「MOVE」が対応することがわかり、発話とジェスチャの統合的な意味が得られる。

#### 5 マルチモーダル問い合わせシステム

以上で述べたインターフェースを用いた問い合わせシステムを試作した。このシステムは、地図上のオブジェクトや道順に関する、ユーザの数十種類の問い合わせ発話と5種類のジェスチャの統合的な意味を同定した後、それに対する適切な応答を生成し、合成音声と地図上のサークルやラインドラッキングの描画によってユーザに情報を提供する。例えば、「ここここではどちらが安いですか」という発話とともに、AホテルとBホテルにサークル・ジェスチャを行うと、統合的な意味を同定した後、データベースのホテルの値段を検索して、「こちらのホテル（このときたとえばAホテルが丸で囲まれる）が安いです」という応答を返す。

#### 6 おわりに

地図上のジェスチャと発話の統合的な意味を同定する手法を提案し、その手法を用いた問い合わせシステムを試作した。この手法では、ジェスチャ側の情報は、発話の意味解析後にはじめて考慮されて意味統合される形をとっている。しかしながら、ジェスチャの情報は、意味解析以前の言語解析、すなわち、構文解析、形態素解析、音声認識のレベルで解析精度などの向上に貢献できる可能性も考えられるので、今後はその点について検討してゆく。

#### 謝辞

システム構築において多大な支援を頂いた田中吾郎氏(株式会社CSK)に感謝致します。

#### 参考文献

[1] M. Blattner and R. Dannenberg (Ed.): "Multimedia Interface Design," ACM Press, 1992  
 [2] K. Loken-Kim, et al: "Verbal-Gestural Behaviors In Multimodal Spoken Language Interpreting Telecommunications," Proc. of EUROSPEECH'95, 1995  
 [3] T. Shimizu, et al: "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. of ICASSP'96, 1996.  
 [4] 田代他: "音声言語処理のための構文解析ツールキット," 情報処理学会研究会報告NL-106-12, 1995  
 [5] ローケン・キム他: "ATRにおけるマルチモーダル翻訳通信研究 -マルチモーダル・ユーザ・インターフェースの実装-, " 電子情報通信学会技術報告書SP95-64, 1995