マルチモーダル観光案内対話システムへの擬人化エージェントの実装とその評価

傳田 明弘　伊藤敏彦　小暮 悟　中川聖一

E-mail address : {akihiro, itoh, kogure, nakagawa}@slp.tutics.tut.ac.jp

豊橋技術科学大学 情報工学系
〒441 豊橋市天伯町字雲雀ヶ丘 1-1

本研究室では、富士山周辺の観光案内をタスクとする「富士山観光案内日本語音声対話システム」に「タッチ入力」及び「対話の途中経過の画面表示」の機能を付加した、マルチモーダルインタフェース化の開発を行なっている。さらに、実顔画像／アニメーション、及び、実音声／合成音声を用いたエージェントインタフェースをシステムに実装し、被験者によるタスク遂行及びアンケート調査の評価実験で、インタフェース及びシステム全体についての評価を行なった。実験では、ユーザは「機械らしく」「首尾一貫している」エージェントを好むことが分かった。また、本システムのマルチモーダルインタフェースの有用性を十分に示すことができ、旅行案内のタスクにおいてマルチモーダル対話システムが持つ可能性を見い出すことができた。

Implementation of Anthropomorphic Interface on Multi-Modal Sightseeing Guidance Dialogue System and Evaluation of the System

Akihiro DENDA, Toshihiko ITOH, Satoru KOGURE and Seiichi NAKAGAWA
E-mail address : {akihiro, itoh, kogure, nakagawa}@slp.tutics.tut.ac.jp
Department of Information and Computer Sciences
Toyohashi University of Technology
Tenpaku-cho, Toyohashi-shi, Aichi-ken, 441, Japan

In our laboratory, we have developed the multi-modal interface with speech input/output, graphical output and touch input for our spoken dialogue system; "Mt. Fuji Sightseeing Guidance System by Spoken Japanese". Furthermore, we implemented an agent interface with real face image / animation and real speech / synthesized speech to the system and carried out evaluation experiments which consist of task completions and questionnaires to evaluate the interface and whole system. The results indicate that users prefer "mechanical/artificial" and "consistent" agent. And they indicate the usefulness of the interface and potentialities of our multi-modal dialogue system about sightseeing guidance.

1　はじめに

音声認識技術、及び、これを支援する言語処理技術の向上により、ユーザがより自然な言い回しで対話を行なえる音声対話システムが実現されてきている。また、音声対話システムの対話を支援し、ユーザに使い勝手のよいマン-マシンインタフェースを提供することを目的として、「音声入出力」以外に「ポインティングジェスチャ入力」や「画像による出力」などの複数の入出力手段（モダリティ）を相補的に統合し、同時あるいは逐次に用いるマルチモーダルインタフェースの研究も近年盛んに行なわれるようになってきている [1, 2, 3, 4, 5, 6] 。

我々の研究室では、「富士山観光案内日本語音声対話システム」[12, 13] の開発を行なってきた。しかし、マン-マシンインタフェースを音声のみで提供していた従来のシステムでは、システムからの応答が音声のみで行なわれるために、ユーザに不安や負担を与えることがあった。そこで、「システムとの対話の途中経過表示」や「タッチ入力、及び、指示詞や指示代名詞（以降は、これらをまとめて「指示語」と呼ぶ）を含んだユーザ発話の許可」といった、マルチモーダルインタフェースの実現を試みてきた [18, 19, 20] 。

また最近では、計算機と人間との（face-to-face）インタラクションを、対話システム上でエージェントという形で実現しようとする研究も盛んに行なわれている [7, 8, 9, 10, 11] 。顔は、非言語的（ノンバーバル）な情報をも含む様々な情報を相手に効果的に伝達できるメディアであるとされ、他のモダリ

ティと相補的に扱われることで、人に計算機に対する親近感を与え、人と計算機との対話をよりスムーズにするといった効果も確認されている。

本稿では、音声入出力、タッチ入力、テキスト／画像出力、及びエージェントインタフェースによるマルチモーダルインタフェースを備えた観光案内対話システムについて、及び被験者を募って行なったインタフェースの評価実験について述べる。

2 富士山観光案内システム

「富士山観光案内日本語音声対話システム」[12, 13] は、富士山周辺の観光案内をタスクとしており、ユーザの発声する音声を入力とし、その発話内容に対する観光案内を合成音声で応答する。現在のシステムでは、普段我々が使用しているような話し言葉に近い『自然な発話（Spontaneous Speech）』を理解することが可能になっている。ここでいう『自然な発話（Spontaneous Speech）』とは、発話文中に「関投詞」「未知語」「助詞落ち」「言い淀み・言い直し」「倒置」といった話し言葉特有の現象を含んだ発話のことである。

「富士山観光案内音声対話システム」は、「入力音声認識部 [14] 」「対話理解・管理部 [12, 13] 」及び「応答生成部」の 3 つの部分に、以下で述べる「マルチモーダルインタフェース」を付加した構成になっている [19] 。

2.1 マルチモーダルインタフェース

観光案内のような検索システムのインタフェースが、音声のみで実現されている場合には、従来のシステムでは、様々な問題点が生じてくる [19] ため、場合によってはユーザに不安や負担を与えかねない。そこで、問題点の一つの解決策として本システムでは、以下に示すようなマルチモーダルインタフェースを採用している。

2.1.1 ディスプレイ上への情報表示

システムからの応答や過去の対話で得られた情報を記憶し、ユーザが必要とする情報を、対話の途中経過表示として、以下の 4 種類の手段によって画面上に表示する。

- 現在の対話内容に対応する場所の地図（富士山、河口湖、山中湖、西湖、精進湖、本栖湖のいずれか）及び現在のトピックを表示。
- 現在の対話内容の対象が観光地や観光施設である場合、その場所の写真画像を表示。
- システムからの応答文が多く（現在は 4 個以上）の項目を含んでいる場合に、これらをメニューとしても表示。
- システムとの対話から得られた情報の内、各観光施設の（名称、種類、料金や食事、駐車場の有無といったその他の）情報は、対話履歴として随時表示。

これらをディスプレイ上に表示した画面の例を図 1に示す。

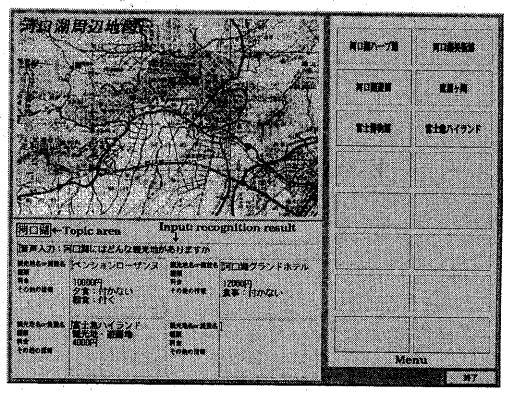


図 1: 画面表示の例

（左上：地図, 左中央：現在のトピックと入力音声の認識結果, 右：応答文のメニュー, 左下：対話履歴）

2.1.2 タッチ入力処理部

ユーザがシステムからの応答文の一部を聞き逃してしまった場合にも、以降の対話で聞き逃した観光地名等を対象とする質問を行なえるように、ディスプレイ画面上にはメニューが表示される。このメニュー内の項目、若しくは、地図上の位置や地名等に指を触れながら（タッチ入力）、これを「こ」系列の指示語で言い表すことによって、システムとの対話を行えるタッチスクリーン入力法を本システムでは実現している。「こ」系列の指示語を含んだユーザの発話とタッチ入力を組み合わせたシステムへの入力は、例えば次のようになる。

[例 1.]

ユーザ動作： 地図上の湖（例えば、河口湖）の位置にタッチしながら

↓

ユーザ発声： この周辺にはどんな観光地があるんですか。

[例 2.]

ユーザ動作： メニュー内の項目「ハイランドホステル」にタッチする。

↓

ユーザ発声： ここの入場料金はいくらかかりますか。

3 マルチモーダル音声対話システムの改良点

これまで行なってきた予備的な評価実験[19, 20]では、満足のいく「認識速度」と「ユーザ発話文認識率」が得られない、タッチ入力の精度が悪い、一部の語が省略された文に対する認識結果を意味ネットワークに変換する生成規則が不十分である、といった問題点が浮き彫りにされたため、今回の実験の前に、以上の問題点を解消する改良をまず行なった。

3.1 「入力音声認識部」の改良

システムとのリアルタイムの対話を実現するために、以下のような変更、改良を行なった。まず、認識用計算機をHP Model C-160から、HP Model J280に変更することによって、認識の高速化をはかっている。さらに、ビームサーチにおける探索時に単語の予備選択処理を行なうアルゴリズムの改良も行なっている[15]。認識に使用しているCFGについても、これまでの予備評価実験では、文法記述の不十分さがあったため、新たな文法記述を追加し、併せて、単語辞書にいくつかの単語も追加した。これにより、単語辞書の語彙数は292から373に、パープレキシティは、これまでの103から126に増加している。

これらの改良によって、現在のシステムの認識処理にかかる時間はリアルタイムの1.1〜2倍弱である。

3.2 タッチ入力装置の変更

これまで十分な精度が得られていなかった「タッチ入力」の性能を向上させるために、新しいタッチスクリーンディスプレイCV213PJ（TOTOKU東京特殊電線株式会社製）をシステムに導入している。主な仕様を以下に示す.

タッチスクリーン方式：	アナログ静電容量方式
分解能：	1024×1024
応答速度：	最大15ms

3.3 「対話理解・管理部」の改良

「対話理解・管理部」で使用している形態素解析システム“JUMAN”のバージョンを0.8から3.2にあげた。これに伴って、「対話理解・管理部」に、形容詞や副詞の理解を可能とし、また次のような省略語の補完処理を加えた。この改良によって以下のような対話が可能になった。

USR: 本栖湖にはどんな旅館がありますか。
SYS: 本栖湖山荘が本栖湖にはあります。

USR: ホテルはありますか。
SYS: 富士本栖湖ホテルが本栖湖にはあります。

例では、ホテルについてのシステムへの問いかけで、「本栖湖には」が省略されていると解釈し補完を行なっている。

3.4 エージェントインタフェース

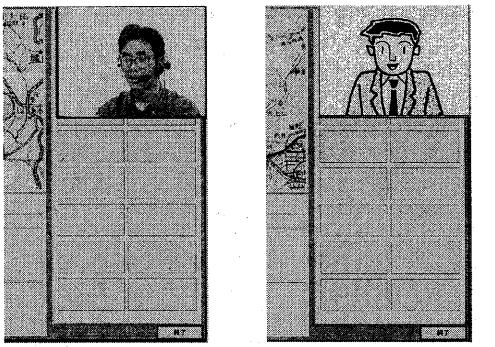
本システムでは、システムからユーザへの決まり文句、例えば

- 「データベースに登録されておりませんので、お答えできません」
- 「もう一度おっしゃって下さい」

といった応答についてはエージェントインタフェースによってユーザに提供している。本システムで実装しているエージェントインタフェースは以下の3種類である。

- 実画像 & 実音声 のエージェント
- 実画像 & 合成音声 のエージェント
- CG 画像 & 合成音声 のエージェント

実際にエージェントが画面上に表示された様子を図2に示す。



(a) 実画像　(b) CG

図 2: エージェントの出力例（画面の右半分を表示）

1つ目の、「実画像 & 実音声のエージェント」は、ビデオカメラで撮影した人物映像を切り出して、MPEG形式の動画像に変換したものである。

2つ目の、「実画像 & 合成音声のエージェント」では、システムからの大部分の応答で使用している合成音声（富士通研究所（株）提供の音声合成ソフトウエアによるもの）とエージェントの応答音声の声質を同一にしたものである。MPEG実画像に付随している実音声と、基本となる合成音声との間で、DPマッチングを行ない、実音声の発声速度に、合成音声の発声速度を合わせたものを使用している。同時に、合成音声のピッチも実音声にできるだけ近付くように調節を行なっている。

3つ目の、「CG 画像 & 合成音声のエージェント」は、Macintosh上で取り込み、加工を行なった、口唇の動きを表した9枚の人物静止画像から、QuickTime形式→ MPEG形式へと変換したCG画像に、合成音声を組み合わせたものである。CGのエージェントとしては、今回我々が使用したような2次元画像のCGエージェントの他に、3次元画像のCGエージェントがある[8, 9, 11]。我々は、3次元

CGエージェントは、「リアル過ぎてシステムに馴染まない，違和感がある」「開発に時間がかかり過ぎる」といった理由から、2次元CGエージェント[2]を今回採用した。

4 評価実験

本節では、マルチモーダルインタフェースを備えた「富士山観光案内日本語音声対話システム」について行なった2種類の評価実験、システムの「マルチモーダルインタフェースの評価実験」及び「エージェントインタフェースの評価実験」について述べる。

4.1 マルチモーダルインタフェースの評価

4.1.1 実験の形式

この実験は、マルチモーダルインタフェース化を行なった「富士山観光案内日本語音声対話システム」について、「システムの使い勝手の良さ」や「マルチモーダルインタフェース化の効果」に着目して行なった。

実験は被験者らに、条件付きの「2泊3日の富士山周辺への旅行計画」を、システムとの対話で得られる情報を元に立ててもらうという形式で行なった。決定してもらう内容は、1、2日目の目的地とそこでのプラン（どこに行くかor何をするか)、及び、1、2日目夜の宿泊先の施設の所在、種類、料金、宿泊施設名の計12項目である。また、計画を立ててもらうに当たって、こちらから幾つかの条件を提示した。それは、例えば、以下のようなものである。

シナリオ

あなたは、今年の研究室旅行の幹事に任命されました。富士山に行く予定です。旅行と同じ日に「精進湖」で開催される研究会に研究室のメンバーの何人かが参加することになっているため、1日目の夜に「精進湖」の「ホテル」で合流することになっています。あなたが計画するのは、おおよそ下の図及び表にある項目です。さあ、富士山観光案内システムを使って、今年の夏の旅行の計画を立てて下さい。健闘を祈ります。

この例では、1日目夜の宿泊先は、「精進湖」の「ホテル」ということがあらかじめ決まっていることになる。このように幾つかの条件を付けたシナリオを以下の6種類用意し、各被験者にランダムに3種類渡して計画を立ててもらった。なお、各被験者には、計画を立てる際に、項目を埋めるためにどうしても必要な情報が対話で得られない場合には、その項目は決定しなくてもいい、という指示をあらかじめ与えた。

- 「自分が所属するテニスサークルの夏旅行の計画」1日目はテニスを、2日目には何か他のスポーツをする」
- 「友達同士3人での富士山への旅行計画」「1日目の昼は湖で遊んで、その日の夜は、本栖湖山荘かニュー山中湖という所に宿泊する」
- 「友達同士3人での富士山への旅行計画」「宿泊は1日目、2日目ともペンションにする」
- 「家族旅行の計画」「1日目の昼は富士急ハイランドで、ゾーラ、ダブルループといったアトラクションを楽しんで、その日の夜はペンションに泊まる」
- 「家族旅行の計画」「1日目は河口湖という湖の周辺の観光地に行き、2日目はサイクリングをする」
- 「研究室旅行の計画」「旅行と同じ日に精進湖で開催される研究会に研究室のメンバーの何人かが参加するので、1日目の夜に精進湖のホテルで合流する」

被験者は、音声対話システムに関する知識を持っていない本学学部生及び大学院生9人である。各被験者には、実験の前に「富士山観光案内日本語音声対話システム」で実際に扱える入力文で構成された適応化文50文によって、認識に用いるHMMの話者適応化を行なってもらった。

4.1.2 対話の形式

システムとの対話は、以下の3種類の形式で行なった。括弧内の記述は、それぞれの対話形式の表における記述を表している。

1. 音声のみの入力、及び、音声のみの出力による対話 （音声入出力）
2. 出力は音声出力に加え、対話の途中経過をディスプレイ上への画像出力で与えるマルチモーダル出力とし、入力は音声のみで行なう対話（音声入力・マルチ出力）
3. 入力は音声のみ、または音声とタッチスクリーンを用いた入力の併用で実現されるマルチモーダル入力、出力は音声及びディスプレイ上への途中経過画像によるマルチモーダル出力で行なわれる対話 （マルチ入出力）

被験者らには、旅行プランのシナリオを3種類渡し、上述の3種類の対話形式によるシステムとの対話を3回行なってもらう。その際、1回目の前に練習セッションを設けてシステムと自由に対話してもらうことにより、「システムにどのようなことが聞けるのか」「旅行計画の各項目を埋めるのに必要な情報を得るにはどのようにシステムに問いかければいいのか」といったことを各自に確認してもらっている。この練習セッションでの対話形式は、各被験者が最後に使用する対話形式と同じである。練習セッションの後、上記の3種類の対話形式でシステムと対話を行なってもらい、それによって旅行計画を3案立ててもらう。今回の実験では、被験者A,B,Cには対話形式3→1→2→3（練習セッションも含む)、被験者E,F,Gには、対話形式1→2→3→1、被験者H,I,Jには、対話形式2→3→1→2の順で対話を行なってもらっている。

4.1.3 評価結果

被験者9人のタスクの達成率（必要な12項目全てを正しい内容で決定できた割合）は74.1%（20タスク/全27タスク）（前回の評価実験[20]では、約30%（9タスク/全30タスク））、平均項目達成率（各項目を正しい内容で決定できた割合）は96.9%（314項目/全324項目）（前回の評価実験[20]では、約79%（190項目/全240項目））であった。その内訳を見ると、項目を埋めるのに必要なデータが対話でどうしても得られなかったケースが4例（全て「音声入出力」の対話）、記入されたデータが間違っているケース（ほとんどが被験者の応答文の聞き間違い）が6例あった。また、これらを対話形式別に見ると、「音声入出力」の対話での項目未達成例が6例、「音声入力、マルチ出力」の対話での項目未達成例が1例、「マルチ入出力」の対話での項目未決定例が3例であった。これを見る限り「マルチモーダルシステム（「音声入力・マルチ出力」&「マルチ入出力」のシステム）」の方がシステムからの応答に含まれる情報をメモとして残す対話履歴やメニューによる表示の機能を持っているため、「音声入出力」のシステムに比べ、正しい検索データを得やすいと言える。記入されたデータが間違っているケースの中には、音声入出力のみの対話で観光地名を聞き間違えた例も1例含まれていた（記入されていた内容：とうよう台、正解：紅葉台（展望台の名称））。

「入力音声認識部」の文認識率（助詞誤りは無視）は、約60%で、そのうちの85%、全体の約51%（前回の評価実験[20]では、約46%）が「対話理解・管理部」における解析で、正しい意味ネットワークに変換できた。ユーザの発話に対する正解応答率は、約47%であった。正しく応答できなかったものは、「もう一度言って下さい」「データベースにないのでお答えできません」といったものや、誤った理解による応答などであった。後者の場合、ユーザはシステムの誤りに気付き、もう一度発声する場合が多かった。また、音声認識用の文法による発話の受理率は、約79%（前回の評価実験[20]では、約72%）であった。受理できなかったもののうち、未登録語を発声した文が43文（約5%）、被験者の発声誤り（鳴沢の氷穴→鳴沢の風穴、など）が10文あった。

表1にシステムの「文認識率」「文理解率」「正解応答率」「対話に費やした時間」及び「発話数」を示す。ここでいう「対話に費やした時間」とは、各被験者が対話を開始してから、旅行プランを完成するまでの時間であり、それぞれの被験者が次の質問を考えたり、システムから提供された情報をメモしたり、旅行プランの記入用紙にシステムから得られた情報を記入する時間も含まれている。

表1において、「対話に費やした時間」と「発話数」について、対話形式間で比較してみると、「発話数」については、「音声入出力」のシステムと「マルチモーダルシステム」の間に優位な差は見られないが、「対話に費やした時間」については、「音声入出力」のシステムとの対話よりも「マルチモーダルシステム」との対話の方に、各被験者は時間を費やしていることが分かる。このような傾向は、これまでに行なってきた予備的な評価実験[19, 20]や他の研究[16]でも見られたことである。このことの理由として考えられるものの一つに、**「マルチモーダルシステム」との対話では、システムから、システムの応答音声に地図やメニューといった画像情報を統合した形で情報提供がなされるため、その豊富な情報によって、次発話を決定するのに戸惑ったり、考え込んだりすることがある**といったことがあげられる。実際の実験でも、「マルチモーダルシステム」との対話では、実際に被験者の対話の様子を観察してみると、画面を見ながら考え込んだり、メニューのどの項目に触れようかと思案しながら指を移動させたりする様子が何人かの被験者から観察できた。「音声入出力」のみのシステムとの対話では、このように次発話について考え込むような様子はほとんど見られなかった。

この結果のみをもって、「マルチモーダルシステム」が「音声入出力」のシステムのどちらが優れているとは言えないが、「マルチモーダルシステム」が持つ特質とその可能性は見い出すことができる。ある被験者は、画面上に表示される地図として、「道路地図が欲しい」と言っていた。これは、観光案内システムに観光地や宿泊施設への行き方の情報も提供して欲しいということを要求していることの現れであろう。このような情報を、「音声入出力」のみのシステムで提供するのは困難であるが、画像情報を提示できる「マルチモーダルシステム」では比較的簡単に実現できる現在のシステムでは、このように情報はサポートしていないため、有効な応答を返すことはできないが、画像による「観光地への行き方」の情報の提示をシステムに実現し、知識データベースの整備をすることによってこれらがサポートできれば、画像情報から対話内容を豊かにできるものと思われる。

このように、マルチモーダルインタフェースには、対話の内容を豊かにすることができる反面、システム側にはより高度な処理技術（データベースの内容の充実を含む）が要求されることになる。

4.1.4 アンケートによる考察

「ディスプレイ上への途中経過表示」について

「地図表示」については、6人の被験者が役に立ったと答えた。役に立たなかった、若しくは、どちらとも言えないと答えた被験者が3人いたが、その理由として、1人の被験者が「地図上の情報が少な過ぎる。どこに触ることができるのか分からない」と指摘した。これは、以前行なった予備評価実験[19, 20]の際に、地図上にたくさん表示されている地名や観光地名などに、システムの扱える語彙が対応しきれていなかった（辞書への未登録など）ために、地図から得られた付加的な情報についてシステムに問いかけてもきちんとした応答が得られず、被験者が不快感を感じていたことから、システムの知識データベースにない情報をできるだけ排除した地図を表示していることによるものである。システムの知識データベースをきちんと整備（データ量を増やす）した上で、システムが扱える知識に対応し

表 1: 評価結果
1段目:「文認識率」 2段目:「文理解率」 3段目:「正解応答率」 4段目:「対話に費やした時間」 5段目:「発話数」

対話形式 使用順	被験者								
	A	B	C	D	E	F	G	H	I
音声入出力	71% 38% 37% 31'05" 73	64% 44% 40% 19'55" 55	50% 44% 38% 10'18" 17						
音声入力・マルチ出力	59% 46% 43% 23'02" 58	77% 81% 73% 11'43" 27	43% 38% 29% 29'05" 47	100% 75% 75% 5'05" 8	59% 55% 55% 9'50" 23	45% 45% 45% 17'55" 34			
マルチ入出力	61% 52% 48% 10'21" 25	93% 74% 78% 11'00" 27	68% 68% 58% 14'41" 21	77% 82% 73% 9'47" 22	49% 44% 44% 16'16" 55	56% 52% 52% 13'29" 24	42% 37% 32% 19'22 41	67% 57% 43% 18'13" 49	80% 64% 56% 31'33" 59
音声入出力				65% 71% 65% 8'45" 17	74% 68% 68% 7'48" 23	62% 69% 62% 7'33" 17	23% 19% 19% 13'35" 32	74% 83% 78% 6'57" 24	46% 39% 29% 14'47" 30
音声入力・マルチ出力							42% 37% 32% 11'32" 29	83% 66% 62% 10'27 33	37% 38% 29% 32'17" 81

た情報が地図上に表示されるのであれば、地図という画像情報の性格上、音声のみの対話で提供できる情報よりはるかに多くを伝えることができ、結果として対話の内容を豊富にすることもできることをこのことは示唆している。

「メニュー表示」については、6 人が役に立ったと答えている。彼らは「音声では聞き逃した部分に対して、表示してくれる部分が役に立った」「これ、とか言うだけで、指したものについて答えてくれるので役に立った」といった理由をあげている。一方で 3 人の被験者が「メニュー表示」は役に立たなかったと答えた。被験者の 1 人が理由として「表示される時は良いが、表示されないことが多い」と指摘した。現在のシステムでは、応答文に「多く (4 個以上) の項目」が含まれていた際にメニューへの表示を行なっているが、応答文に含まれる項目がたとえ 1 個であっても画面上に表示して欲しいとのことであった。実際の対話の際にも、項目一つの応答文を聞き直すために、同じ問いかけをする様子がしばしば見られた (ただし、この被験者は片耳が難聴であった)。

「対話履歴表示」については、5 人が役に立った、4 人がどちらとも言えないと答えている。どちらとも言えないと答えた被験者らは、表示の仕方に問題点があると指摘している。「残っていて欲しいものまで消えてしまい困る。それがなければ役に立つと思う」「残すもの、消してもいいものを選べるといい」といった意見が得られた。この「対話履歴表示」の表示形式、操作については、今後検討、改良を行なっていきたいと考えている。

「タッチ入力」について

以前行なった評価実験 [19, 20] ではタッチ入力の精度が悪かったために、意図したものを指し示すことができないことがあって、「タッチ入力」は使いにくい、役に立たないと評価されていた。現在のシステムではより高性能なタッチスクリーンディスプレイを導入しているため、意図したものが指し示せないといった状況は観察されていない。「タッチ入力」が役に立ったかについては、「大変役に立った」「そこそこ役に立った」と答えた被験者が 5 人、「あまり役に立たなかった」「ない方がいい」と答えた被験者が 4 人という結果になった。「あまり役に立たなかった」「ない方がいい」と答えた被験者らは、「タッチ入力」を使用しなかった、あるいは使用しても「使いにくかった」と答えた被験者である。彼らは、「「タッチ入力」を使用するのが面倒」「「これ」とか「この」の「こ」をシステムが認識してくれなかった」というような理由を答えている。一方で、「現在の「タッチ入力」の機能に便利さを見い出せない」といった理由もあげられた。現在のシステムで実装している「タッチ入力」で行なえるのは、ポインティングによる指示のみである。これに対し、幾つかのマルチモーダルシステムで実現されているグルーピングによる指示 [10] が実現されたり、タッチ動作のみで情報が得られる機能が実現されたり、指示語を用いた言い回しのバリエーションが充実するといった「タッチ入力」の機能の向上がなされれば、「タッチ入力」は「大変役に立つと思う」「そこそこ役に立つと思う」と全ての被験者が答えている。「タッチ入力」のグルーピングによ

る指示機能は今後実現していきたいと考えている。

「どの入出力インタフェースが使いやすいか」

使いやすかった入出力インタフェースを、使いやすかった順に並べると以下のようになる。

- 「マルチ入出力」→「音声入力・マルチ出力」→「音声入出力」：3人
- 「音声入力・マルチ出力」→「マルチ入出力」→「音声入出力」：3人
- 「音声入力・マルチ出力」→「音声入出力」→「マルチ入出力」：2人

この結果は、「タッチ入力」が使いやすかった、使いにくかったかの評価に依存したものになった。「タッチ入力」を「使いにくかった」と評価した被験者らが、「音声入力・マルチ出力」のシステムを最も使いやすいと評価している。

その他の意見として、ほぼ全ての被験者が「データベース（の内容・検索法）が弱い」という意見をあげている。システムのその他の部分の改良と合わせて、知識データベースの充実は、今後取り組んでいきたいと考えているが、もともと整備されている商用データベースを利用する方が実用的であろう[17]。

4.2 エージェントインタフェースの評価

システムのエージェントインタフェースについて、被験者によるアンケート調査で評価した。被験者は、情報関係の大学院生32人である。

実験は、システム応答音声のうち、決まり文句の応答音声を、

1. 実画像 & 実音声
2. 実画像 & 合成音声
3. CG 画像 & 合成音声
4. エージェント無し（システムからの音声応答は、全て合成音声で行なう）

によって出力する4種類のシステムに対し、それぞれ4分半くらいの対話をビデオに収録し、それを被験者らに視聴してもらい、その後、アンケート調査に答えてもらった。それぞれの対話は、1→2→3→4の順に提示した。提示順序による評価結果への影響を避けるために、更にもう一度1→2→3→4の順で各被験者に提示した。

アンケートでは、その対話システム、そのエージェントについて、

- 「使ってみたいか、使いたくないか」「自分にとって好ましいか、好ましくないか」「友好的か、敵対的か」
- 「自然的か、機械的か」「違和感は感じられないか、感じるか」

といった2点について、被験者らに1から7までの7段階尺度で評価を行なうように依頼した。同時に、評価についての理由や、エージェント、システムに関するその他の意見も同時に記入してもらった。

アンケートから得られた評価結果は次のようになった。それぞれのエージェントを実装したシステムは、

1. CG 画像 & 合成音声（20/3：使ってみたい／使いたくない）
2. エージェント無し（17/7）
3. 実画像 & 実音声（12/9）
4. 実画像 & 合成音声（8/16）

の順に、「使ってみたい」「好ましい」「友好的な」システムであり、

1. エージェント無し（16/8：自然的／機械的）
2. CG 画像 & 合成音声（13/10）
3. 実画像 & 実音声（9/17）
4. 実画像 & 合成音声（2/25）

の順に、「自然的」「違和感がない」システムであると、被験者らは評価している。

エージェント無しのシステムに高い評価を与えた被験者らは、次のような理由をあげている。

- （このシステムが）実用的で利用するのに最もシンプルで使いやすそうだと思った。
- 画像がないため、機械らしさを感じることができ、他に比べ自然的であると感じた。

エージェント無しのシステムを高く評価した被験者らは、対話を「計算機との対話」であると割り切って考えており、逆に「機械らしい」対話を「自然な対話」であると考えていることがわかる。これは、これまでの他のシステムでの評価結果と異なるもので[2]、今回は、決まり文句のみエージェントによる応答を試みたことによるものとも考えられるが、もっと掘り下げて検討すべき重要な項目である。

「実画像 & 実音声」のエージェントについては、そのエージェントによる応答自体を見ると、「親しみが持てる」「自然に聞きとれるので違和感が少ない」「注告の時に人間の声で言われるのは悪い感じはしない」と被験者らに評価されているが、応答の主要な部分である情報提供を合成音声で行なっていることによって、

- 音声が2種類あると、2対1で話しているようで嫌だ。
- 機械化されたシステムの中に、実音声が組み込まれているのが、不愉快に感じる。

というような印象を被験者らに与え、低い評価であった。

結論として、情報提供の応答音声出力と一貫性のない「実画像 & 実音声 のエージェント」及び、全ての応答音声出力が一貫しているが実画像と合成音声の質のバランスが取れていない（自然的／機械的）「実画像 & 合成音声 のエージェント」が共に悪い評価（「使ってみたい」システムではない）を得、これらと比べて全てが機械的に一貫している「CG& 合成音声 のエージェント」が違和感が少なく一番高い評価（一番「使ってみたい」システムである）を得た。また、「エージェント無し」が二番

目に「使ってみたい」システムであるという評価を得た。

その他、被験者らから得られた意見として多かったものに、

- 「エージェントは常に表示されていた方が良い」

というものがあった。これは、他の情報提供のための表示スペースを確保するために、決まり文句応答の時のみ表示したことによる。また、応答音声に合成音声と実音声が混在することに違和感、嫌悪感を感じ、音声出力モダリティが一貫していることを好む被験者が多かったことから、「エージェントの常時表示」「実画像 & 実音声 のエージェントで、全ての応答音声をカバーする（評価用に）」ことも、今後考えていく必要があろう。

5 むすび

本研究では、従来音声のみをマンマシンインターフェースとしていた音声対話システムに対し、「対話の途中経過のディスプレイ上への表示」「タッチ入力と、指示語を含んだユーザ発話を組み合わせた入力」及び「エージェントによる決まり文句の応答出力」を実現したマルチモーダル化の改良を行なった。

そして、そのシステムを使用して、被験者によるインタフェース及びシステム全体の評価を行なった。「1対話の総発話数」に関しては、入出力方式間の優位性は見られなかったが、「対話に費やした時間」については、「音声入出力」のみのシステムとの対話よりも「マルチモーダルシステム」との対話の方が、各被験者が時間をかけて行なっているという傾向を見い出すことができた。これは、対話の内容を豊かにできる可能性を含んだ結果である。また、アンケートによる主観的な評価では、「機械らしさ」「首尾一貫性」が好まれる要因の一つであることが分かった。また、マルチモーダルインタフェースの有用性を十分に示すことができた。

今後は、現在のシステムに対し、評価によって明らかになったシステム全体の不備な点や、今回導入したエージェントインタフェースをどのように、マン-マシンインタラクションに生かしていくかを検討していく。

参考文献

[1] D.Teil, Y.Bellik: "Multimodal dialogue interface on a workstation", Venaco Workshop and ETRW on "The structure of multimodal dialogue" (1991.9).

[2] 竹林洋一:「音声自由対話システム TOSBURG II − ユーザ中心のマルチモーダルインタフェースの実現に向けて −」, 電子情報通信学会論文誌 , Vol.J77-D-II, No.8, pp.1417–1428 (1994).

[3] 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」, 電子情報通信学会論文誌, Vol.J77-D-II, No.8, pp.1465–1474 (1994).

[4] 中川聖一, 張建新:「音声と直指操作による入力インターフェイス」, 電気学会論文誌, Vol.114-C, No.10, pp.1009–1017 (1994).

[5] 神尾, 松浦, 正井, 新田:「マルチモーダル対話システム MultiksDial」, 電子情報通信学会論文誌, Vol.J77-D-II, No.8, pp.1429-1437, (1994).

[6] 伊藤, 古川, 中沢, 木山, 張, 岡:「複数ユーザによる音声とジェスチャのマルチモーダルインタフェースシステム:Real-time GSI の一評価実験」, 情報処理学会, 音声言語情報処理研究会報告, 96-SLP-10, pp.3–8 (1996).

[7] 長谷川, 森島, 金子:「「顔」の情報処理」, 電子情報通信学会論文誌, Vol.J80-D-II, No.8, pp.2047–2065 (1997.8).

[8] 長谷川, 坂上, 伊藤, 栗田, 速水, 田中, 大津:「視覚情報を対話的に学習するマルチモーダル擬人化エージェント」, 情報処理研究会報告, CVIM-100-4, pp.33–38 (1996.5).

[9] 土肥 浩, 石塚 満:「WWW/Mosaic と結合した自然感の高い擬人化エージェントインタフェース」, 電子情報通信学会論文誌, Vol.J79-D-II, No.4, pp.585–591 (1996.4).

[10] 知野, 河野, 屋野, 池田, 鈴木, 金沢:「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」, 情報処理学会, 音声言語情報処理研究会報告, 97-SLP-17, pp.115–120 (1997.7).

[11] 長尾 確:「マルチモーダルインタフェースとエージェント」, 人工知能学会誌 Vol.11, N0.1, pp.32–40 (1996.1).

[12] M.Yamamoto, S.Kobayashi,Y.Moriya, S.Nakagawa: "A Spoken dialog system with verfication and clarification queries", IEICE Trans., Vol.E76-D, No.1, pp.84-94 (1993).

[13] 山本, 伊藤, 肥田野, 中川:「人間の理解手法を用いたロバストな音声対話システム」, 情報処理学会論文誌, Vol.37, No.4 (1996.4).

[14] 甲斐, 中川:「冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価」, 電子情報通信学会論文誌, Vol.J80-D-II, No.10 (1997.10).

[15] 甲斐, 伊田, 中川:「大語彙連続音声認識のための音響的先読みによる高速化」, 日本音響学会 平成9年度秋季研究発表会 講演論文集, 3-1-3, pp.91–92 (1997.9).

[16] Abigail J. Sellen: "Remoto Conversations: The Effect of Mediating Talk With Technology", HUMAN-COMPUTER INTERACTION, Volume 10, pp.401–444 (1995).

[17] 酒井, 山田, 伊藤, 小森, 上田, 池田:「CD-ROM を全文検索する音声ガイドシステムとその評価」, 電子情報通信学会論文誌, Vol.J77-A, No.2, pp.232–240 (1994.2).

[18] 傳田 明弘, 中川 聖一:「日本語音声による観光案内システムのマルチモーダルインターフェイス化」, 情報処理学会第 52 回全国大会 (2) , 4D-3, pp.167–168 (1996.3).

[19] 傳田, 伊藤, 中川:「マルチモーダルインタフェースを備えた観光案内対話システムの評価」, 人工知能学会全国大会 (第 10 回) 論文集, pp.431–434 (1996.6).

[20] 傳田, 伊藤, 小暮, 中川:「マルチモーダルインタフェースを備えた観光案内対話システムの評価実験」, 情報処理学会, 音声言語情報処理研究会報告, 97-SLP-15-8, pp.47–52 (1997.2).