マルチモーダル観光案内対話システムの被験者実験による考察

6N－6

傳田 明弘　伊藤 敏彦　中川 聖一
豊橋技術科学大学 情報工学系

1 はじめに

我々の研究室では、「音声入出力」「タッチ入力」「地図やメニュー、対話履歴、エージェントインタフェースからなる出力」を入出力モダリティとするマルチモーダル観光案内対話システム（富士山観光案内対話システム）の開発を行なっており、これまで、このシステムのマルチモーダルインタフェースの評価実験を幾つか行なってきた[1][2]。

今回は、これまでの実験で得られた音声対話データについて、システムの不完全さに起因する、システムの認識誤りによるユーザの言い直しを含む、表層的あるいは意味的に同一な発話を取り除き、再評価を行なった。また、「システムからの全ての応答音声を出力するエージェントインタフェース」を備えた対話システムとの対話の様子を、被験者に視聴してもらう主観的な評価も行なった。本稿では、これらの結果について報告する。

2 音声対話データの再評価

我々の研究室で開発、改良を行なってきたマルチモーダル対話システムは、富士山周辺の観光案内をタスクとしている。システムが備えているマルチモーダルインタフェースの効果やその使い勝手を調べるために、2泊3日の富士山周辺への旅行計画を立ててもらうタスクを被験者に課す形式の評価実験をこれまで行なってきた。これらの評価実験では、マルチモーダルインタフェースを備えた対話システムとの対話の方が、音声入出力のみの対話システムとの対話より、長い時間を各被験者がかけて行なっていて、対話に要する発話の数もマルチモーダル対話の方が多くなっているという結果が得られている[2]。今回は、これまで行なってきた評価実験で得られている音声対話データから、システムの不完全さに起因する発話を評価項目から除外するために、認識誤りによるユーザの言い直しを取り除いて、結果をまとめ直した。評価対象の音声対話データは、9人の被験者（本学学部生及び大学院生）が、対話システムと行なった以下の3種類の対話から得られたものである。

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、あるいは、音声とタッチ入力の組み合わせで実現されるマルチモーダル入力とし、出力も音声と画像の組み合わせ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**）

各被験者の対話から、

- 書き起こしの文字列のレベルで同一な「表層的な同一発話」
- 言い回しが多少異なっているが話している内容は同一な「意味的な同一発話」

を取り除いて、各被験者の各対話における「発話数」及び「対話に費やした時間」をまとめた結果を表1に示す。なお、発話の削除を行なう際には、「発話数」に関しては、「表層的」あるいは「意味的」に同一な発話の固まりの最後の発話を優先して、固まり内のそれ以前の発話を削除した。これとは逆に、「対話時間」に関しては、同一な発話の固まりの最初の発話を優先して、固まり内の以降の発話を削除した。

表を見ると、「音声入出力」対話の発話数より「マルチ入出力」対話の発話数の方が12%多いという関係が、「同一発話を取り除く前」と「取り除いた後」とで変化していない。一方、対話時間の方は、「同一発話を取り除いた後」の方がより差が開いている（「マルチ入出力」対話の方が21%増し→30%増し）。これは、音声入出力のみの対話と、マルチモーダル対話との性質の差によるものであると考えられる。「山中湖には何があるんですか」といった、応答に複数の項目が含まれるような発話をユーザが行なった場合、その次の発話がなされるまでの間隔が「マルチ入出力」対話の方が「音声入出力」対話より開いていた。どこにタッチするか、どのように発話するかを、30秒以上、時には1分近く考え込んでいる様子が、今回再評価した音声対話データ内に見られた。

3 エージェントインタフェースの主観的な評価

我々は、「実画像」「実音声」「CG」「合成音声」といった各モダリティを組み合わせたエージェントインタフェースを対話システムに実装し、システムとの対話の様子を被験者に視聴してもらって、アンケートに答えてもらう形式の、主観的な評価実験を行なってきた[2]。これまでの評価実験では、エージェントインタフェースがカバーする応答が決まり文句のみであったが、「決まり文句の実音声と、観光案内の合成音声の混在に違和感を感じた」というコメントが多く得られたため、今回、システムからの全ての応答を

1. 「実画像エージェント & 実音声応答」
2. 「CG エージェント & 合成音声応答」
3. 「エージェントなし」

でカバーする3種類のインタフェースを実装したシステムを評価用に用意した。被験者は次の2つのグループから成っている。1つ目のグループは、本学情報工学系の大学学部3年生12人と大学院生15人（グループ1：男子学生24人、女子学生3人）、2つ目のグループは、愛知学泉大学家政学部の1年生51人（グループ2：男子学生13人、女子学生38人）である。各被験者には、上記の3種類のシステムの、それぞれ3分ほどの対話を収録したビデオを試聴してもらい、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.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.の順で各被験者に提示した。

アンケートでは、各システムとその出力インタフェースについて、

Consideration of sightseeing guidance spoken dialogue system with multi-modal interface through subjects' evaluation experiments

Akihiro Denda, Toshihiko Itoh and Seiichi Nakagawa
Toyohashi University of Technology

表 1: 全ての被験者の発話数／対話時間 [秒] の合計

	対話の形式		
	音声入出力	音声入力・マルチ出力	マルチ入出力
同一発話を取り除く前	288 / 7193	340 / 8994	323 / 8736
表層的な同一発声を取り除いた後	242 / 5934	275 / 7524	266 / 7473
意味的な同一発話を取り除いた後	194 / 4957	191 / 5250	218 / 6466

- 「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」「友好的に感じるか／友好的には感じられないか」
- 「人間的だと感じるか／機械的だと感じるか」「自然な感じがしたか／不自然だと感じたか」
- 「違和感を感じないか／感じるか

といった3つの尺度に対し、各被験者にそれぞれ5段階で評価するように依頼した。同時に、評価についての意見や、エージェントあるいはシステムに関する意見も同時に記入してもらった。

結果を以下に示す。グループ1の被験者とグループ2の被験者は異なった評価をしていた。グループ1の被験者の評価結果は、次のような順位付けになった（順位付けは得点の平均に基づいて行なっている）。

1. 実画像 & 実音声
2. CG & 合成音声
3. 合成音声のみ

一方、グループ2の被験者の評価結果は、以下のようになった（男女差による違いはなかった）。

1. CG & 合成音声
2. 実画像 & 実音声
3. 合成音声のみ

得られた評価結果は、「実画像エージェント & 実音声応答」に対してどのような印象を持つかによって、グループ1とグループ2の間で評価が異なったということを意味している。グループ1の被験者らは、「実画像エージェント & 実音声応答」に対して、

- 人と会話している感じがして良い
- 人の声で応答を返された方が落ち着く

といった印象を受けており、対話をより自然にするような雰囲気を持ったこのインタフェースを高く評価していた。また、「作成の容易な実画像が表現力の点でも優れていると思う」といった、より表情やしぐさの豊かなインタラクションを実現するという観点に立って評価を行なった被験者もいた。

グループ1の被験者らの評価とは対照的に、グループ2の被験者らは、「実画像エージェント & 実音声応答」に対して、最も自然で（他に比べると）違和感が少ないと判断してはいながらも、このインタフェースが機械的なシステムと組み合わさることにより多少の違和感を感じていた。そのためか、

- 機械に質問しているのに、人間が映って答えてくるから変な感じがした。

という印象を受けた被験者が何人かいた。また、「実画像エージェント & 実音声応答」のインタフェースが、他に比べると自然で違和感が少ないために、

- 人だと遠慮してしまうのが、絵だと話しやすい。
- 人間の顔が出てくるとこわい。

というような、計算機でワープロ、表計算ソフトやCG作成ツールを利用する程度のユーザにとっては、逆にシステムに話し掛けにくい雰囲気を作ってしまう可能性があることが分かった。

「合成音声のみ」のインタフェースについては、グループ1、グループ2の被験者ともに、最も低い評価となった。この「合成音声のみ」のインタフェースは、システム応答の内の決まり文句のみにエージェントインタフェースを適用していた以前の実験 [2] での、大学院生の評価では、

- 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。

という具合にかなり高く評価されていたが、今回、エージェントインタフェースが全ての応答をカバーすることによって、エージェントインタフェースの利点がはっきりと感じられる結果となった。計算機でアプリケーションソフトウエアを利用する程度のグループ2の被験者のみならず、グループ1の被験者も、

- 画像が出ないことで応答のタイミングを逃す可能性があると思う。
- あまりに真面目で見ててつまらない。

という具合にあまり高く評価していない。以前の実験で多かった「機械と割り切って話すことができる」といったコメントもあったが、今回の実験では少数意見であった。

4 むすび

今回、システムの不完全さに起因する認識誤りによるユーザの言い直しを、音声対話データから削除し、再評価を行なった。その結果、前回の評価結果通り、対話形式の相違による「発話数」と「対話時間」の関係を見い出すことができた。また、エージェントインタフェースの評価については、計算機でソフトウエアを利用する程度の一般大学生には、「CG エージェント & 合成音声応答」のインタフェースが、工学系の大学生、大学院生には、「実画像エージェント & 実音声応答」が好まれることが分かった。今後は、今回の評価で得られた知見を元に、本システムのような情報検索の対話システムに合ったエージェントインタフェースの実現を行なっていく。

参考文献

[1] 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).

[2] 中川, 傳田, 伊藤：「マルチモーダル観光案内対話システム」, 人工知能学会誌, Vol.13, No.2 (1998.3).