Chracteristics of Telephone-only and Multimodal Dialogues

3J－3

Kyung-ho Loken-Kim Ryo Furukawa* Fumihiro Yato
Tsuyoshi Morimoto
ATR Interpreting Telecommunications Research Labs.
*Advanced Institute of Science and Technology, Nara

1 INTRODUCTION

The optimal multimedia configuration for an application, such as multimedia interpreting telecommunications, can not be obtained in an ad hoc fashion. It rather requires a series of empirical studies conducted in settings simulating those in which the intended uses are most likely to take place. ATR's Environment for Multimodal Interactions (EMMI) [1] is a simulation tool that supports a variety of realistic environments for interpreting telecommunications. EMMI has been created specifically for collecting speech and language data people might use in multimodal interpreting telecommunications, and it supports three tasks: directions, reservations, and negotiations. This paper reports the results of the preliminary study [2] comparing the characteristics of telephone-only dialogues and multimodal dialogues when carrying out the simulated directions task.

2 EMMI (Environment for Multimodal Interactions)

EMMI is comprised of the following equipment. There are two NeXT computers: one for the agent, and the other for the client. Both computers are equipped with a keyboard and a mouse. One SONY Digital Audio deck, two microphone amplifiers, and two headphones have been installed to obtain high quality speech transmissions and recordings. Two video cameras, connected to the Video Monitor interface of the NeXT Cube, are used to transmit video motion images of the agent and the client. A third video camera has been installed to capture the client's interactions with EMMI. Three telephones have been installed to collect telephone-to-telephone speech-only dialogues. Presently, two telephone lines are used: one for the client, and the other for the agent.

3 METHOD (Scenario, Subjects, and Measurements)

This experiments were carried out in Japanese. A client asks the conference secretariat how to get from Kyoto Station to the Kyoto International Conference Center. The secretariat then gives the directions by displaying maps of the areas surrounding Kyoto Station, the international Conference Center, and Kyoto Park Hotel. The maps are displayed on both of the secretariat's and client's screens, and they can engage in a dialogue while marking and writing relevant information on the maps using a mouse. A total of eight paid Japanese subjects (as clients) participated in the experiment. To minimize learning effect, the subjects were divided into half; the first half ran the telephone-only experiment (Tel) first and multimodal experiment (MM) later, and the second half ran the experiments in reverse order.

Statistics for the following data has been collected to measure the linguistic differences between Tel and MM. (1) Sentence length: average number of words per sentence has been computed. (2) Use of deixis: It was hypothesized that the graphic interface in MM would induce users to use deictic gestures different from Tel dialogues. To investigate this, frequency of different deixis have been counted. (3) Length of dialogue: It was also hypothesized that two-person-communication using a variety of modalities would be more effective in information exchange, thus, the task would be accomplished faster, resulting in shorter dialogue duration.

電話対話とマルチモーダル対話の特徴分析
ローケンキム キュンホ、古川 亮*、谷戸 文廣、森元 逞
ATR音声翻訳通信研究所
*奈良先端科学技術大学院大学

4 RESULTS AND DISCUSSION

In this study, data analysis was limited to two dialogues of the same subjects: one Tel and the other MM. With reference to point 1 above; table 1 shows that the sentence lengths were about the same for Tel and MM. With reference to point 2 above; a major difference between Tel and MM was found in the way deictic markers were used for referent-identifications (Table 2). First, both agent and client were 2 to 3 times more likely to use deictic markers when having a MM dialogue than when having a Tel dialogue. In addition, in the Tel dialogue, third-person pronouns (tpp) starting with そ(so, roughly "that") were most frequently selected whereas in the MM dialogues tpp starting with こ (ko, roughly "this") were favored.

e.g. 1) Tel: バスが出ておりますので、。。。それに乗って頂きまして、

2) MM: バス停がございます。。。こちらのほうの二番乗り場からですね、

In addition, in the MM dialogues, the subjects tended to use こ-tpp for objects which appeared on the map, but tended to use そ-tpp to refer to objects appeared in their dialogues.

3) MM(when referring to an object on the screen): このあたりにバス停がございます。

4) MM (when referring to an object appeared in their dialogues) : そしたらその指定の宿泊先の方を教え願いますか。

Finally in reference to point 3 above; contrary to our expectations; the MM dialogue took longer (585 seconds) and more turn-takings than Tel dialogue (484 seconds). This was partially caused by the time needed for the agent to look for the correct maps and display them.

5 CONCLUSION

In conclusion, although only three areas were studied; sentence length, deictic markers, and overall dialogue length and form, several differences in the linguistic characteristics of Tel and MM were observed. The increased dialogue time for MM dialogues may be the result of the agent's unfamiliarity with the maps and equipment, and may be lessened through practice or by simply replacing the current machine with a faster one. However, the differences in use of deictic markers appears more fundamental, and may also be affecting the pattern of turn-taking and dialogue length. The presence of a visual image (i.e. map) seems to increase the tendancy of the client to confirm information supplied by the agent by indicating specific points and asking for confirmation that "this" is the position to which the agent referred. This tendency toward redundant confirmation may make the information exchange more effective, but may act to lengthen rather than shorten the dialogue. We are currently developing a third simulator for a translator which will allow us to examine further the linguistic phenomena and user behavior in multimodal interpreting telecommunications.

Table 1 Dialogue Length

	agent		client	
	Tel	MM	Tel	MM
No. of Words	782	898	636	726
No. of Sentences	107	116	114	123
Words/Sentence	7.3	7.7	5.6	5.9
No. of Turns	77	94	76	94

Table 2 Deictic Markers

	agent		client	
Deixis	Tel	MM	Tel	MM
ここ	0	7	0	2
こちら (tpp)	0	11	1	3
こっち	0	0	0	1
この	0	4	1	3
これ	1	0	5	9
そこ	0	2	2	0
そちら (tpp)	1	1	0	1
その	1	0	0	1
そのような	0	0	1	0
それ	3	0	5	1
そちら (fpp)	4	3	0	0
そちら (spp)	0	0	7	3

REFERENCES

[1] Loken-Kim, Yato, Kurihara, Fais, Furukawa (1993): "EMMI-ATR Environment for Multimodal Interactions," ATR TR, unpublished

[2] Furukawa, Yato, Loken-Kim (1993): "Analysis of Telephone and Multimedia Dialogues," ATR TR, unpublished