

LINGUISTIC AND PARALINGUISTIC DIFFERENCES BETWEEN MULTIMODAL AND TELEPHONE-ONLY DIALOGUES

*Kyung-ho Loken-Kim, Fumihiko Yato, Laurel Fais,
Tsuyoshi Morimoto, Akira Kurematsu**

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
+Department of Electronics Engineering, University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182 Japan

ABSTRACT

The results of a pilot study comparing linguistic behavior of speakers in telephone-only and multimodal communicative environments, are reported. The subjects, native speakers of Japanese, conducted goal-oriented conversations in each environment. Linguistic differences (disfluency rates, and use of intention types, syntactic structures, and deictic expressions) and paralinguistic differences (ease of use, and utilization of media) are discussed. Suggestions are made for incorporating these findings into the design of multimodal spoken language interpretation systems.

1. INTRODUCTION

Multimedia systems seem to have great potential for facilitating human communication, yet the issues concerning multimedia system configurations that induce optimum performance in human information conveyance are rarely addressed and barely understood. A multimedia system configuration should be based on our understanding of human communicative behavior in that environment. Oviatt et. al. [1] report that linguistic variability in people's speech and writing can be reduced with the careful selection of input modality and presentation format. Their findings are very intriguing, but do not address the issue of a task that is a bit more open-ended, such as, a map-based direction-finding task in which two people communicate to each other using a variety of modalities (typing, marking, speaking..). How do the linguistic and paralinguistic characteristics of such multimodal dialogues differ from those of unimodal dialogues such as the telephone dialogues?

This paper is an attempt to answer some of these questions. It summarizes the results of our pilot study aimed at finding out the linguistic and communicative characteristic differences between telephone-only and multimodal dialogues. As a first step, our experiment was conducted in a monolingual (Japanese-to-Japanese) setting, but issues involving multimodal, multilingual interpretation are also discussed.

2. MULTIMODAL SIMULATOR - EMMI -

EMMI (ATR Environment for MultiModal Interaction) is a simulation environment designed to realistically simulate the setting of a monolingual (bilingual in the later version), telephone-only communications, as well as multimodal communications. EMMI is neutral in the sense that it does not contain any sort of intelligence; it is simply a man-machine-man interface. EMMI has been created specifically for collecting data on the speech and language people might use in multimodal telecommunications. EMMI is designed to accommodate

three main tasks: giving directions, making reservations, and negotiating. The interlocutors are allowed to communicate using a variety of modalities, such as speaking, typing, writing, marking, and tracking, and looking.

For example, the agent, who is acting as the conference secretariat, can give directions verbally to the client over the headphone while showing and writing on the map displayed on both the agent's and the client's screens. On the other hand, a subject acting as a client may ask the agent for directions from the Kyoto station to the International Conference Center (for detailed hardware and software configurations of EMMI, see [2]). EMMI, like any other experimental system, is undergoing constant modification, and the most recent version enables us to simulate a three-person bi-lingual multimodal interpreting telecommunication task.

EMMI is equipped with an array of multi-media data collection equipment. Currently, three video cameras, two used to transmit the full-motion video images of the participants, and the third used to record the client's interaction with the system, are available. Three telephones and a Digital Audio Tape deck have been installed to collect speech-only dialogues and ensure the high quality recording of all verbal transactions.

Participants interact through the multi-media window illustrated in Figure 1. This window is divided into four sub-windows: information, video, input/output, and logo. The INFORMATION window is used for displaying maps, reservation forms, and calendars, as well as for marking and writing. For example, participants, while engaged in a dialogue, can mark and write necessary information on the map by pressing the left button of the mouse and dragging the cursor across the screen (touch panels are installed in the later version). In order to distinguish the client's and the agent's marks and writings, different colors are

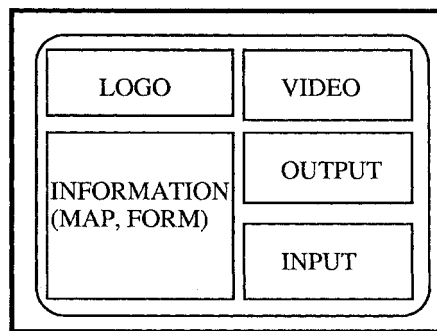


Figure 1. User Interface

used. When a reservation form appears on the screen, both the agent and the client can fill out the form simply by typing.

The VIDEO window is used for displaying full-motion video images of the client and the agent. Participants of course can turn off the video camera if they choose. The limitations of the video camera positions do not allow direct eye contact between the participants.

The INPUT and OUTPUT windows are provided to aid verbal communication by allowing participants to exchange information in text. Japanese proper nouns, for example, are more easily described in text than in verbal descriptions.

3. METHOD

The linguistic differences between multimodal dialogues and telephone-only dialogues were measured by collecting statistics for the 1) disfluency including interjections (e.g., "ah", "eh", etc.), false starts and hesitations¹, 2) intention types and syntactic classifications [3], and 3) sentence lengths, and 4) deictic expressions. Paralinguistic differences were measured by observing the subjects' behavior during the experiments and through post-experiment interviews.

A total of eight paid Japanese subjects (clients) and one non-paid female Japanese subject (agent) participated in the experiment. Each subject acting as a client was given detailed instructions on how to use EMMI; e.g., how to use the headphone-microphone while using the telephone, how to mark and write on the map using the mouse, etc. The subject acting as the agent was trained in all aspects of operating EMMI, as well as the geography of Kyoto. After the experiment, a post interview was conducted to elicit the subjects' responses.

The main goal to be achieved was attending an international conference which was held at the Kyoto International Conference Center. The two subgoals were 1) finding one's way from Kyoto Station to the International Conference Center, and 2) reserving a hotel room.

To minimize learning effect, the subjects were divided in half: the first half ran the telephone-only experiment (Tel) first and multimodal experiment (MM) second, and the second half ran the experiments in reverse order.

All the subjects were videotaped from the front to record facial and head movements and from the side to record manual movements. The agent was also videotaped from the front (this image appeared on the clients' monitor).

4. RESULTS

4.1 Linguistic Differences

4.1.1 Disfluency Rate

The agent's disfluency rates in both Tel and MM environments were lower than those of the clients' despite uttering longer sentences than the clients (Table 1, Table 2). The agent, however, showed higher disfluency rates in Tel than MM, in sharp contrast to the clients' rates. This suggests that the agent with training and

¹ Disfluency rate is computed by dividing the number of interjections by the length of the dialogue in seconds. This approach was taken instead of the standard measures for disfluency commonly used for English, which represents the number of interjections per 100 words, because the concept of word in Japanese does not exactly correspond with the concept of word in English.

repeated experiments with the multimodal system, actually reduced her disfluency rate to a level lower than that when using the telephone in spite of the complexity of multimodal interactions: frequent modality shifts and parallel use of multiple modalities.

The clients, on the other hand, had higher disfluency rates in MM than in Tel in spite of the pretraining. Especially those clients who went through the experiment in Tel->MM order showed a higher disfluency rate than those who went through MM->Tel, suggesting that experimental order was another factor influencing the rates. The agent's disfluency rate, however, remained fairly constant in both experimental orders in MM, but showed the higher disfluency rate in Tel in the Tel->MM order.

Speaker	Agent		Client	
	MM	Tel	MM	Tel
Tel-> MM	0.084	0.110	0.238	0.135
MM-> Tel	0.089	0.096	0.209	0.121
Mixed	0.086	0.102	0.223	0.128

(interjections/second)

Table 1. Disfluency Rate

Environment	Agent	Client
Tel	9.29	6.18
MM	10.97	5.89

(words/sentence)

Table 2. Average Sentence Lengths

4.1.2 Intention Types and Syntactic Classifications

In MM, the agent often had to explain the maps and reservation format in detail, making it apparent that the agent acted more as an information source in MM than in Tel (*inform* 35.8% in MM vs. 29% in Tel, see Table 3). The detailed explanations resulted in longer sentences (Table 2) and longer time for the multimodal dialogues than the telephone-only dialogues (on average, 400 seconds in MM vs. 256 sec in Tel).

Operating EMMI also caused further distinctions in the speaker's role. For example, the delay between finding the appropriate map and displaying it not only lengthened the dialogue, but also induced the agent to generate more *request* utterances, such as "Wait a minute" (5.3% in MM vs. 1.9% in Tel) in MM than Tel.

The clients, on the other hand, responded to or acknowledged the agent more in MM (*response* 54.6% in MM vs. 41.5% in TEL, and *acknowledge* 7.9% times in MM vs. 5.8% in Tel) while uttering shorter sentences (Table 2), clearly acting more as information receivers in MM than in Tel.

Another interesting point to note is that the clients, most of whom were not as familiar with the multimodal system as the agent, ended up generating more meaningless utterances than the agent (utterances with no meaning 2.7% for the client vs. 1.2% for the agent in MM).

Syntactic classifications show (Table 4) that the agent acting as an information provider used declarative sentences more than half of the time in both Tel and MM, while the client responding to the agent uttered short idiomatic sentences most of the time. The agent used a slightly higher proportion of declarative sentences in MM (56.9%) than in Tel (53.6%) even though the clients asked slightly fewer questions (19.1% in MM than in Tel (20.7%). Both clients and the agent used a slightly higher percentage of the interrogative sentences in Tel than MM, reflecting the necessity for clarification in telephone dialogues.

Speaker	Client				Agent			
	Tel		MM		Tel		MM	
Environment	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.
Intention Type	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.
promise	0	0.0	2	0.4	2	0.4	1	0.2
offer	0	0.0	1	0.2	6	1.3	6	1.0
suggest	0	0.0	0	0.0	2	0.4	0	0.0
invite	0	0.0	0	0.0	0	0.0	1	0.2
permit	2	0.5	0	0.0	0	0.0	0	0.0
phatic	11	2.6	10	1.9	24	5.1	29	4.9
expressive	9	2.1	13	2.5	8	1.7	18	3.1
response	178	41.5	236	45.6	144	30.5	154	26.1
acknowledge	25	5.8	41	7.9	58	12.3	51	8.7
confirmation	73	17.0	80	15.5	53	11.2	49	8.3
request	26	6.1	30	5.8	9	1.9	31	5.3
questionref	19	4.4	21	4.1	16	3.4	14	2.4
questionif	26	6.1	23	4.4	12	2.5	19	3.2
inform	55	12.8	46	8.9	137	29.0	211	35.8
utterances with no meaning	5	1.2	14	2.7	1	0.2	5	0.8
total	429		517		472		589	

Table 3. Intention Types

Speaker	Client				Agent			
	Tel		MM		Tel		MM	
Environment	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.
Sentence Type	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Freq.	Rel. Freq.
interrogative	89	20.7	99	19.1	66	14.0	81	13.8
imperative	0	0.0	0	0.0	0	0.0	0	0.0
idiom	183	42.7	230	44.5	153	32.4	173	29.4
declarative	157	36.6	188	36.4	253	53.6	335	56.9
Total	429		517		472		589	

Table 4. Syntactic Structure

4.1.4. Deictic Expressions

A major difference between Tel and MM was found in the way deictic markers were used for referent-identifications. First, both agent and client were 1.5 times more likely to use deictic markers when having an MM dialogue than when having a Tel dialogue (0.0815 deictic markers/sec in MM, 0.056 deictic markers/sec in Tel, see Table 5). In the Tel dialogues, mid-range deictic expressions starting with *so*, roughly "that" were most frequently selected whereas in the MM dialogues, near-oriented deictic expressions starting with *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す (There is a bus... take that bus)

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、(There is a bus stop... from the number two platform on this side)

In addition, in the MM dialogues, the subjects tended to use the *ko*-marker for objects which appeared on the map, but the *so*-marker to refer to objects referred to 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。(There is a bus stop around here)

4) MM (when referring to an object referred to in their dialogues): そしたらその指定の宿泊先の方を教えてください。(Then please tell me about that hotel)

Speaker	Client		Agent	
	Tel	MM	Tel	MM
Deixis	Tel	MM	Tel	MM
koko (here)	1	5	0	26
kochira (I, here)	2	7	10	83
kore (this)	9	16	3	7
kono (this)	2	14	0	37
kotchi (this side)	0	0	0	1
kokora (around here)	0	0	0	1
soko (there)	6	2	1	2
sochira (you, there)	20	16	36	15
sore (that)	12	9	8	10
sono (that)	6	7	3	3

Table 5. Deictic Marker

4.2 Paralinguistic Differences

4.2.1 Ease of Use

Subjects' impressions of EMMI varied from mostly *simple* to a few *some efforts*: *simple* may have been caused by the fact that all the subjects were familiar with computers, and *some effort* may have resulted from the complicated system configuration. For example, several subjects were asked to use the keyboard to type in their last names in Chinese characters which involves moving the cursor to the Input window, using the mouse and typing, and some subjects felt that this process was a bit confusing and complicated.

4.2.2 Video Image

Surprisingly, only two subjects made any comments about the video image: one subject felt that the map was enough, and the other subject thought the video image was very useful.

The problem with the video image is that the present video camera positions do not allow direct eye contact between the participants. One would have to look straight into the camera lens in order to make direct eye contact. In other words, the screen and the camera would have to be one and the same. With the current setting, it seems that one is talking to another who is looking away. However, this is not a major problem because participants look at the information window most frequently and seem to keep the video image in their peripheral vision.

4.2.3 Hardcopy

In both Tel and MM, all the subjects took some notes on scratch paper while they were engaged in the dialogue. Note taking was especially common in the Tel dialogues. In MM, on the other hand, the abundance of available information seemed to give the subjects the impression that they understood the agent's message, thus, taking fewer notes; later some subjects remarked that they wished they had taken more notes or had a hardcopy of the map.

4.2.4 Cursor Position, Mouse, and Keyboard

Several subjects were having difficulty finding their own as well as the other party's cursors, especially when the cursors were inactive on the map. These static

cursors were later changed to round rotating cursors. The initial cursor position on the map can be used as "You are here" sign, since many clients were having a hard time figuring out their position on the map. Another problem the subjects were having was that when they drew a line on the map using the mouse, they could not confirm whether the line appeared on the other party's map, causing confirmation messages such as "Do you see this line?" Marking and drawing with the mouse, if one is not used to it, can be very awkward. This may be another reason most clients did not use the mouse very much. In the more recent version, the mouse has been replaced by a touch panel.

Most of the subjects (clients) felt that the keyboard was not very useful because of the aforementioned reason. Pen writing over the touch panel may eventually replace the keyboard.

5. DISCUSSION.

Our assumption, prior to the experiment, was that the multimodal system would greatly facilitate human communication. While it seems to be true, the results of our experiment suggest that training and familiarity with such a system are important factors affecting the effectiveness of the communication. For example, the agent (a trained user), regardless of the complexity of handling the multimodal system, consistently showed smaller disfluency rates in the MM environment than the clients - even less than when using a telephone, and remained very active in voluntarily providing information throughout the experiments.

The clients, on the other hand, showed greater disfluency rates in MM in spite of uttering rather simple and short sentences, and remained passive, resulting in their not utilizing the full capacity of the multimodal system. An interesting point is that this, however, does not necessarily mean that the clients disliked the multimodal system. The results of a second study [4] show that the subjects enjoyed using the multimodal system more than the telephone, and they felt that the multimodal interactions facilitated their understanding of the agent's instructions [5].

From these findings, we can conclude that the human interface of the multimodal systems for a naive user should be simple (for example, only video and map without the keyboard and the mouse), and allow the user to initiate only the initial contact, while a multimodal system for an experienced user could be fairly complex giving the user the freedom to select whichever modality he/she chooses to convey the information.

How can we incorporate these findings into the designing of a spoken language interpretation system [6]? The following results open up the potential for introducing multimodality to a spoken language interpretation system: 1) The clear distinction of the speakers' roles in the multimodal dialogues may help us to define speaker models. 2) The agent's consistently low disfluency rates can help us obtain higher automatic speech recognition and machine translation rates. The agent can be trained to talk to a machine, which is, of course, a much easier task than training the general public (naive users). 3) The parallel use of modalities (mainly marking and speaking) increased the number of referent expressions, opening up the possibility of using the information provided by one modality to resolve the ambiguity occurring in the information provided by another modality.

There are areas, nevertheless, which could be problematic. For example, synchronizing multiple sources of information in a bilingual context is tricky and requires a deep understanding of the information in one language in order to translate it to another language in a well coordinated manner. Without such information understanding and synchronization, the bilingual multimodal interactions will be incomprehensible.

6. Conclusion

The results can be summarized as follows: 1) there was a clearer distinction in the speaker's roles in multimodal dialogues than in telephone-only dialogues, 2) the frequency of near-oriented deictic expressions (*koko* "here", *kore* "this") was greater in multimodal dialogues, and mid-range deictic expressions (*soko* "there", *sore* "that"), which are common to telephone exchanges, were rare in multimodal dialogues, 3) the agent often gave verbal directions while using the mouse to mark on the screen (use of multiple modalities in parallel), 4) both the clients and the agents shifted to the keyboard modality to express some portion of the dialogue, name, and telephone numbers in text.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to Dr. Yamazaki, the president of ATR Interpreting Telecommunications Research Laboratories, for the support, and cheerful encouragement which made this study possible. We also would like to thank Mr. Kurihara for his efforts in realizing EMMI.

REFERENCES

- [1] S. Oviatt, P. Cohen, M. Wang and J. Gaston, "A Simulation-Based Research Strategy for Designing Complex NL Systems", ARPA Workshop on Human Language Technology, Mar. 1993
- [2] Kyung-ho Loken-Kim, Fumihiko Yato, Tsuyoshi Morimoto, A Simulation Environment for Multi-modal Interpreting Telecommunications, IPSJ, AV Multiple Information Processing Workshop 4-1, 3/18, 1994 in Japanese
- [3] Masaaki Nagata, Masami Suzuki, Sachiyo Tsukawaki, First Steps Toward Annotating Illocutionary Force Types to a Bilingual Dialogue Corpus, ATR Technical Report TR-I-0298, Mar. 1993
- [4] Laurel Fais, Kyung-ho Loken-Kim, Effects of Mode on Spontaneous English Speech in EMMI, ATR Technical Report TR-I-0059, July, 1994.
- [5] Ryo Furukawa, Fumihiko Yato, Kyung-ho Loken-Kim, Analysis of Telephone and Multimedia Dialogues, ATR Technical Report TR-IT-0020, Sept. 1993
- [6] F. Yato, T. Morimoto, Y. Yamazaki and A. Kurematsu, "Important Issues for Automatic Interpreting Telephone Technologies", Proc. ISSD-93, pp. 235-238, Nov. 1993