TR - IT- 0087

An Experiment for Telephone versus Multimedia Multimodal Interpretation: Methods and Subject's Behavior

Young-Duk Park, Kyung-Ho Loken-Kim, Larel Fais.

December, 1994

Abstract

We performed an experiment to investigate the user's conversational behavior when communicating with a multimedia and multimodal interpreting service. The experiment was carried out for two tasks (asking direction and making hotel reservations) and participants used two kinds of experimental terminals: a telephone and a multimedia multimodal terminal. The collected experimental data were the speech of clients, agents, and interpreters, and the interpreter's computer screen. We analyzed the collected data from the viewpoint of linguistic and paralinguistic behavior. From the experiment results, it is clear that the multimedia and multimodal communication is necessary and very useful to the users in the interpreting telecommunication. However, the multimedia and multimodal affects the dialogue pattern, and conventional media processing schemes should be changed in the multimedia interpreting service.

1. General

Future telecommunication systems will be multimedia and multimodal systems. In such an environment, users may prefer the multimedia terminal to the conventional telephone, because the former provides a more convenient and easily understandable communication environment. In addition, the user interface of the future system will be a multimodal one, which eliminates the need for a lengthy definite description, and affects the dialogue pattern. [1]. Thus, research on interpreting telecommunication systems must consider the dialogue patterns and the system architectures affected by multimedia and multimodal system, because future interpreting telecommunication services can not avoid being in a multimedia and multimodal environment. [2]

Therefore, we performed an experiment to investigate the user's conversational behavior when dealing with a multimedia and multimodal system in an interpreting service. We also tried to identify the system requirements of future multimedia multimodal interpreting systems. The first experiment was carried out based on a direct connection between clients and agents, and we got some useful results. [3]-[7]

Following the previous experiment, we performed an interpretation experiment on the telephone and multimedia multimodal system. The main purpose of this experiment was to collect data for an analysis of telephone and multimedia multimodal dialogues through an interpreter. To investigate the difference between telephone and multimedia multimodal conversations in interpreting service, we decided to collect the linguistic and paralinguistic behaviors of the subjects. On the basis of collected data, we measured the number of turns, the number of words, the number of words per turn, the disfluency rates and the vocabulary. Also from the participants' interview after the experiment and observation of their behavior during the experiment, we found the paralinguistic behavior of the subjects. We used EMMI (ATR Environment for Multi-Modal Interactions) to collect the data. EMMI is a an environment designed to simulate the setting of a bilingual, telephone-only communications and multimedia multimodal communications session through interpreter. [8] Through the experiment, we recorded the speech of clients, agents, and interpreters onto DAT (Digital Audio Tape). The interpreter's computer screen with audio was also videotaped.

In this report, we describe the method of the experiment, the participants' gesture behaviors through the experiment, and the participants' linguistic differences between the telephone and the multimedia/multimodal interpreting experiment. At the end of this report, we also mention the subjects' opinions of the interpreting service and our future directions.

2. Objective

The goal of this experiment was to :

1) collect the user's conversational behaviors of telephone and multimedia multimodal communication through an interpreter,

2) investigate the influence of interpretation intervention in the telephone and multimedia multimodal dialogues, and

3) analyze the effect resulting from mixing of media and modalities on the interpretation dialogues,

4) collect the elementary data and necessary conditions for the next experiment based on the Wizard of Oz (WOZ) simulation method.

3. Method

3.1 Tasks

o Asking directions:

North American native speakers of English were asked to imagine that they were arriving for the first time in Kyoto Station and that they wanted to know the way to the International Conference Center. They called on the "Interpreting Service Center" and talked to an "agent" who gave directions to the foreign visitors through an interpreter.

o Making hotel reservation:

If the foreign visitor wanted to reserve a hotel room, the agent gave information about hotel reservations through the interpreter.

3.2 Subjects and equipment

3.2.1 Subjects

o Two Japanese interpreters

o Five Japanese agents

o Ten north American native speakers of English clients

3.2.2 Equipment

1) Three telephone for voice communication

2) Two NeXT Workstations for multimedia multimodal communication (client and agent)

3) A SS-10 Workstation for multimedia and multimodal communication (interpreter)

4) Three Headset for recording the subjects' speech

5) Two Digital Audio Tape Recoders for recording the subjects' speech

6) Three Amplifier for the speech amplification

7) A Down converter for converting the video signal of the interpreter's screen

8) One VCR for recording the interpreter's screen

3.3 Experimental configuration

Fig. 1 illustrates the layout of the multimodal research lab. for the data collection. The interpreter was physically separated from the agent and the client, and sound-absorbing partitions were placed between the agent and the client rooms to prevent transmission of the participants' voices. We selected a telephone and a multimedia multimodal terminal as the experiment terminals.

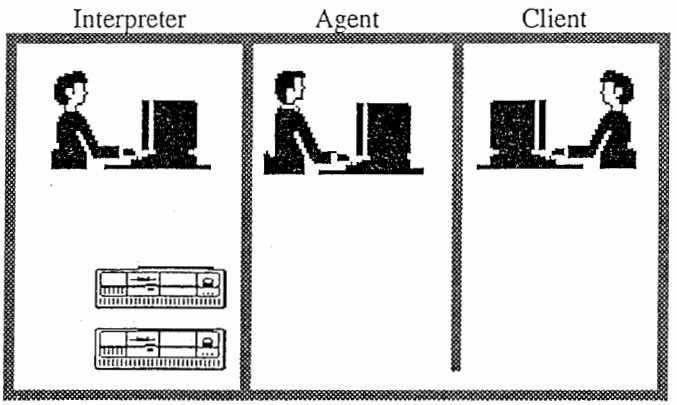

Fig.1 Experimental room configuration

For the telephone experiment, three telephones were installed to collect dialogues. In this experiment, two telephone numbers were used; one for the client, and the other for the agent and interpreter. The line to the agent was interconnected with interpreter's telephone line. Fig. 2 shows the physical connection scheme of the telephone experiment.

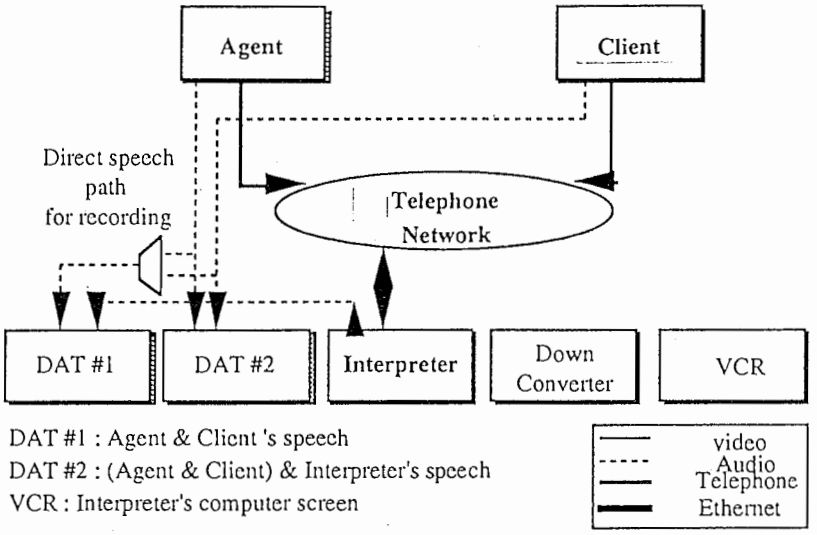

Fig.2 The configuration of the telephone experiment

In the multimedia multimodal experiment, participants could use various kinds of modalities as the style of interaction to access the system. To support these modalities, we used two NeXT computers and a SUN sparc station for the multimedia multimodal experiment. The NeXT computers that were used by the agent and client were equipped with a keyboard, mouse, touch screen, and video camera. The modalities which were used were looking, hearing, speaking, typing, marking, and touching. We prepared three kinds of maps (Kyoto Station area map, conference building area map, and subway line map) and hotel reservation form for the agent. The user interface windows of the agent and client consist of four sub-windows: an information window for map and hotel reservation form displaying, a video window, a text input and output window, and a logo window. For details of the user interface see [3]. A SUN sparc station was equipped with a keyboard and a mouse for the interpreter. The agent's (client's) video image was transmitted to the client (agent) and interpreter through a direct connection. Therefore, the interpreter could see the agent or the client by clicking the mouse. However, the agent (client) only saw the client's (agent's) face. The participants' speech was also directly transmitted in the multimedia multimodal experiment. If the client or agent dragged some point, that position transmitted to the interpreter only. After that procedure, the interpreter dragged his/her mouse along the client's or agent's dragging position with the interpreted speech. However, the interpreter's dragging was only displayed on the agent's or client's computer screen with the interpreted speech. Also the interpreter could transmit his/her dragging information to the agent and the client at the same time by clicking the option button. Information for the map pointing and the hotel reservation was transmitted through an Ethernet network and displayed on each subject's computer screen. Two each of DAT decks, microphone amplifiers, main amplifiers, audio distributors, and headphones were installed to collect high-quality speech. A downconverter and Hi-Fi stereo VCR were used to record the interpreter's computer screen. Fig. 3 shows the physical connection scheme of the multimedia multimodal experiment.

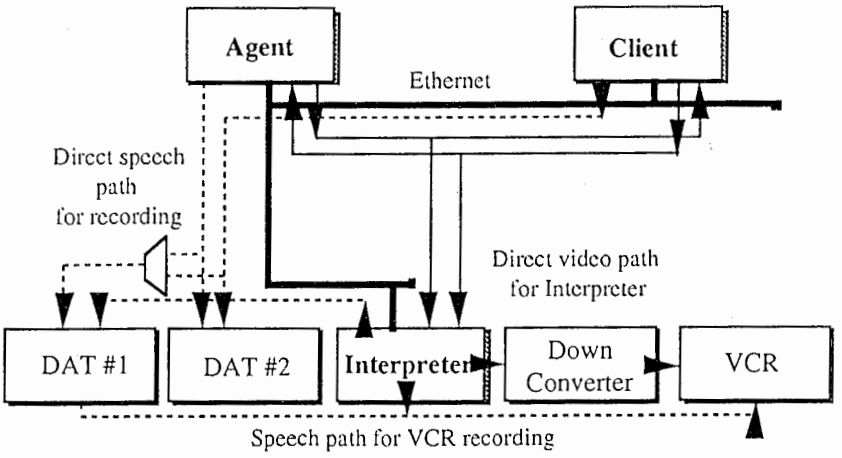

Fig.2 The configuration of the multimedia multimodal experiment

3.4 Collected data

1) Client's speech

2) Agent's speech

3) Interpreter's speech

4) Computer screen of the interpreter

5) Subject's suggestions in the pre and post experiment interview

6) Subject's conversational behavior during experiment

3.5 Experiment sequences

Clients were divided into two groups. "A" group began with the telephone and "B" group began in the reverse sequence to decrease the client's learning effect on the experiment. Each group carried out the experiment in a day. The participants were given instructions about the background of the experiment, the tasks of the experiment, and the operation of the system. After the introduction, participants practiced operating the system. Then the participants performed the actual experiment with a one-page brochure about the conference, and clients asked the two tasks in random order. We did not give any prepared scenarios to help the subjects' conversation. Also, we asked them to talk naturally, and not to be constrained when they had a problem during the actual experiment. After the experiment, we asked some questions to learn the subjects' impressions and opinions about the experiment.

4. Experimental Results

4.1 Linguistic behaviors

To analyze linguistic differences, the collected speeches of subjects was transcribed, including all the speech problems like false starts, interjection words, simultaneous speech, etc. Then we measured the communication duration, the number of turns, the number of words, disfluency rates (including interjections, false starts), and the vocabulary that each participant used in the call. On the basis of the measured data, we analyzed the linguistic characteristics and differences between the telephone and the multimedia interpreting conversations.

4.1.1 Communication duration and turns per call

The average communication duration and turns per call for the telephone (TEL) and multimedia multimodal (MM) sessions are shown in Table 1. The communication duration of the MM calls was greater than that of the TEL calls except for one call. In the exceptional call case, the interjection words and false starts that will be described below, and the average number of words per turn used by the client were fewer than the other MM cases.

Table 1. Average communication time per call

Time Parameters	TEL	MM
Avg. duration (sec)	775	1033
Avg. turns (turns)	254	277
Avg. duration per turn (sec)	3.05	3.73

The average number of turns for MM was also greater than the TEL case even though the subject used fewer turns in the MM dialogue to explain the same kinds of questions. For example, when the agent asked the location of the client, dialogues between the subjects were as follows.

{ MM dialogue example;

A: はい分かりました今京都駅のどちらにいらっしゃるか分かりますか

I: All right. [ah] Do you know which part of Kyoto station you are in right now?

C: Let's see. I'm in front of the travel information center across from the taxi stop.

I:[あのー]タクシー乗り場のちょうど向かいっかわの旅行情報センターの方にいますが

A: はい[えー]地図で言いますと このあたりですね

I: All right. Let me mark that on the map. [are] !(where) Where is that [oh] OK OK (let).! So you are standing right here.

C: That's right.

}

{ TEL dialogue example;

A: はい[えー](きょ)京都駅のどちらにいらっしゃいますか

I: Yes. Which part of Kyoto Station you are in right now.

C: I'm in front of the [ah] taxi [um] taxi stop.

I:今[あのー]タクシー乗り場の前にいるんです

A:[えー]タクシー乗り場は[あー]北と南にあるんですけれども[えー][そうですね]

タクシー乗り場近くに京都タワーが見えますか!大きな[えー]タワーなんですけれども

I: OK. There are two taxi stops. There are one in north, one in south but.

C: [uhuh]

I: Do you see a Kyoto tower in front of you?

C: Yes, I do.

I: 見えます

A: はい 分かりましたそちらの方が北の乗り場になります

I: All right. That means you're at thi north taxi stop.

C: OK

}

This dialogue shows a simple example. The difference was quite significant in the direction explaining case. Therefore, this suggested that subjects asked different questions in MM much more, because the communication environment of MM provides more information to the subjects than the TEL one. The average duration per turn for MM was greater than the TEL case because of the system operation time, such as shifting back & forth among modalities, file opening, typing and so on.

4.1.2 Number of turns per participant

The average number of turns per participant for the TEL and MM are shown in Table 2. The agent had more turns than client, the InsJc (Interpreter speaking Japanese case) and the InsEc (Interpreter speaking English case) in both TEL and MM.

Table. 2 Avg. No. of turns per participant

(turns)

Mode	Agent	InsJc	InsEc	Client	Total
TEL	68.44	60.67	61.56	63.11	253.78
	27%	24%	24%	25%	
MM	73.11	61.56	70.67	71.89	277.22
	26%	22%	26%	26%	

InsJc; Interpreter speaking Japanese case

InsEc; Interpreter speaking English case

The turn distributions among the participants were different when compared with direct monolingual communication, in which both agent and the client had the same number of turns. The reasons for this difference are as follows.

- There was some dialogue between the agent and the InsJc, and between the client and the InsEc merely to solve the ambiguities of the question and answer.
- When the client or the agent spoke a long sentence that was too hard to translate in one turn, the interpreter translated that sentence in two or more turns.
- There were some direct interpretations. For example, when the agent spoke short sentences continuously, the interpreter translated the agent's speaking whether the client responded to it or not.
- There were some direct responses from the agent to the client. For example, the Japanese agent responded immediately after the client spoke a well-known English word like "O.K.", "Yes", "Good", "All right", etc.
- When participants performed the hotel reservation task, the agent asked the telephone number and name of the client. In that case, there was local dialogue (dialogue between the client and the

InsEc or the agent and the InsJc only) to speak and confirm the above information. For example, when the client said his telephone number, the InsEc confirmed the number one by one, and the InsEc began to interpret after receiving all the digits.

4.1.3 Linguistic measurements

Because the language characteristics of Japanese and English are quite different, it is very difficult to find a base reference to compare both languages simultaneously. Therefore, we analyzed the linguistic differences between TEL and MM from each language's view point.

Tables. 3 and 4 show the average number of words per call and average number of words per turn, the average number of interjection words per call, and the average number of false starts per call in both languages.

Table. 3 Linguistic measurements in Japanese dialogues

Subject	Agent		InsJc	
Mode	TEL	MM	TEL	MM
Words/call	874	1023	694	813
Words/turn	12.77	14.00	11.45	12.91
Interjections /call	46.22	52.44	31.56	32.78
Interjection rate	5.288	5.126	4.546	4.032
False starts/call	4.89	4.22	1.78	3.56
False starts rate	0.559	0.412	0.256	0.438
No. of vocabulary	721	733	679	695

Table. 4 Linguistic measurements in English dialogues

Subject	InsEc		Client	
Mode	TEL	MM	TEL	MM
Words/call	575	640	459	561
Words/turn	9.34	9.06	7.28	7.81
No.Interjections/call	28.11	31.67	35.67	38.56
Interjection rate	4.889	4.948	7.771	6.873
No. False starts/call	3.78	4.78	7.67	8.56
False starts rate	0.657	0.747	1.671	1.526
No. of vocabulary	732	741	732	795

The average numbers of words per call and the average number of words per turn in MM were greater than in the TEL case. The InsEc and the agent spoke more words than the client and the InsJc. However, if we consider InsEc to be the agent in the direct English connection and the InsJc to be the client in the direct Japanese connection, the difference of the interpreting call was smaller than the direct monolingual connection case. [7] The reason for this is that when the client asked long questions in the interpreting conversation, the interpreter translated that the sentence in two or more sentences. For the client, in most cases, the words per turn (91% in telephone and 92% in MM) were less than 20. Especially the cases in which the client spoke only one word per turn were about 24% in TEL and 25% in MM. However, that was 11% in the TEL and 14% in the MM of the InsEc case.

Tables. 3 and 4 also show the average number of interjection words and false starts and their percentages to the total number of words in both languages. The disfluency rates in TEL was greater than that in the MM case. The agent spoke more interjection words than the InsJc in Japanese, and the client spoke more interjection words than the InsEc in the English dialogue. These were the same results as the false start's case.

The total number of English vocabulary words used in the TEL and MM dialogues were 1073 and 1068 words respectively. In the Japanese case, the total number of vocabulary words that were used in the TEL and MM dialogues were 952 and 955 words each. "the", "you", "ah", "O.K.", "I", "to", "a", "right", "is" and "and" were the top ten vocabulary words in English. "de", "su", "i", "ma", "-", "no", "e", "shi", "ni" and "ka" were the top ten vocabulary words in Japanese. The client and the interpreter used a similar number of vocabulary words in both the TEL and MM dialogues. However, the deictic expressions that were used for referent-identifications were quite different. For example, the client and the InsEc in MM dialogues, used "here" (118 times) and "this" (101 times) much more for map pointing in contrast to TEL dialogues, in which the expression was seldom used ("here": 42 times, "this": 68 times).

4.2 Gestures behaviors

To investigate the effect of the gestures, we measured the number of turns that included gesture, the shapes of gesture, the usage patterns and intentions of gesture, and the timing relation between speech and gesture. From the transcription dialogue and the videotape that recorded the experiment, we classified subject's gestures into five classes. These were circling (draw a circle-like shape in the area of map), dragging (draw a line-like shape in the map), pointing (click a point in the map), spiraling (draw a spiral-like shape in the area of map), and checking (check a point in the map).

4.2.1 Number of turns with gestures

The gestures occurred mainly in direction finding task, and total 50 turns that included gesture were collected in total 556 turns of the direction finding task. The agents used more gestures than client, because the agent had to explain direction or location on the map in detail with gestures. The gesture turn ratio of the agent and the client was 14% and 4% respectively.

Table.5 Number of turns with gestures

turns	Agent	Client	Total
No. of total turns	272	284	556
No. of gesture turns	38	12	50

The client usually spoke only one sentence with gesture, for example, "I am here (click a point in the map)" when they were asked their location by the agent or did not use any gesture in a dialogue session (four clients did not use any pointing gesture through the experiment). The reason for this is the system unfamiliarity of the subject. In addition, they think that their major consideration point (the reason that the user used the interpreting service) was the translated speech in the interpreting service. Especially, clients preferred the speech to the gesture, because the former provides a more convenient access.

4.2.2 Shapes of collected gestures

Table.6 shows the shapes of collected gestures. The agent's main gestures were the circling and the dragging gesture (95% of agent's gesture) in direction finding task. The clients ,however, used the circling, the dragging, and pointing gesture to express their location (92% of client's gesture).

Table.6 Shapes of collected gestures

Subject	Circle	Drag	Point	Spiral	Check	Total
Agent	13	23	0	1	1	38
Client	4	4	3	1	0	12
Total	17	27	3	2	1	50

4.2.3 Usage pattern of gestures

The subjects used circling gestures mainly for location-indications as shown in Table.7. Therefore, there were many deictic words (82% of circling gesture included deictic word, examples of deictic word; kochira, kono, koko, etc.) and proper nouns (18% of circling gesture included proper noun, examples of proper noun; name of hotel and subway station) in the duration of the circling gesture. However, there were many kinds of parts in a sentence with dragging gesture, because the main usage of the dragging gesture was direction expression (93% in dragging gesture). Also, referent-identification and direction presentation were included in the dragging gestures, and subjects dragged under the object when it used for referent-identifications. Lastly, the pointing gestures, the spiraling gestures and checking gestures were mainly used by the clients for referent-identification and location answer as shown in Table 7.

Table.7 Usage pattern of gestures

Meaning	Circle	Drag	Point	Spiral	Check	Total
Location indication	17	2	3	2	1	25
Show the way	0	25	0	0	0	25
Total	17	27	3	2	1	50

4.2.4 Timing between speech and gestures

As shown in the below e.g. 1 and e.g. 2 sentences, the circling gestures occurred in the middle of the sentence where the deictic words and proper nouns were located. The brackets in the example sentences denote the start and end time of gesture. Also the circling gestures started with deictic words or proper noun, and included the back words of the deictic words because of the drawing time of circling gestures (65% of the circling gestures started with deictic word or proper noun).

e.g. 1; [このあたり] ですね。
e.g. 2; 新幹線は[ここです]。

The dragging gesture and the scrambling gesture included many words in contrast to the circling gestures (e.g. 3 and e.g. 4), and their start time was random. Also the drawing time of dradding gesture was longer than circling gesture. For example, 26% of dragging gestures occurred in a whole sentence.

e.g. 3; [まっすぐ歩いてすぐ右のところ]です。
e.g. 4; [この地下鉄乗り場てなっていますよね、こちら]。

4.3 The subject's impressions and suggestions of the experiment

4.3.1 The subject's impression of the experiment

Because most of the participants were not familiar with the MM experiment system, they said that the TEL experiment was much easier to use than the MM experiment even though 7 clients had some experience dealing with computers among the 10 clients. They also thought that the telephone dialogues provided a more natural dialogue environment, good flow (not interrupted by waiting, typing, shifting back & forth among modalities) for getting the necessary information.

However, after the MM experiment, every subject felt they got more information about the given tasks than from the TEL experiment, and the MM experiment provided a better communication environment for giving the information to the other subjects. The first reason was that MM could give

the information more conveniently and precisely than telephone. They also mentioned that dragging on the map, writing the name of a place and hotel reservation form, the touchscreen, and looking at the eyes of the communication partner were very helpful. However, many participants said that there were some difficulties getting the information in the multimedia multimodal environment because of the following problems:

- long system waiting time,
- being slow- a lag between asking a question and getting the answer,
- nice agent picture; but she was not looking at me so I didn't feel like I was talking to her,
- screen was too busy ,
- typing information; not comfortable,
- inexactness of the touchscreen,
- the abruptness and casual manner of the interpreter, and
- system unfamiliarity.

The interpreters said that they didn't feel any difference between TEL and MM interpreting. If they had enough system experience, MM interpreting was easier than TEL interpreting, because it provided a map, video and other kinds of media simultaneously. They said some problems occurred in interpreting the agent and client's speech, because:

- In overlapping speech periods, interpreting was very difficult.
- Is it necessary to interpret the "OK", "OK" when spoken to speak each other at the same time?, and
- It's very difficult to decide the start and ending time of the interpreting during conversation.

4.3.2 The subject's suggestions for the interpreting service

After the experiment, participants suggested many opinions that would have made the system easier, clearer, or more comfortable. The suggestions are as follows:

- It would be better to see the interpreter than to see the agent.
- There's no need to hear the agent; hearing only the interpreter was enough.
- Video information was not so important.
- It's better to show information about the hotel on the screen. E.g., a picture of the hotel.
- Agent needs automatic directions paint function.
- The agent and client want some cues (e.g. beeps) when interpreter starts typing, , and
- The interpreter should introduce his/herself.

The first suggestion was an opinion about the media communication path among the participants. In this experiment, the agent's (client's) video was transmitted to the client (agent) and the interpreter. Therefore, the agent (client) only saw the client's (agent's) face, and the interpreter could see the agent or the client by clicking the mouse. However, most of the clients thought that it would be better to see the interpreter than to see the agent only. The conventional videoconferencing system controls the video switching function based on two principles; the current speaker sees the previous speaker and

the other sees the current speaker. However, first the principle must be changed in the interpreting service case as follows.

- When the agent or the client speaks, they always see not the previous speaker but the interpreter's scene.
- When the interpreter speaks, it depends on the interpreted language. For example, when the interpreter translates from English to Japanese, the interpreter sees the client.

Also in the interpreting service, there are many short dialogues. This is another characteristic that influences the video switching function. For example, the percentage in which the interpreter and client speak sentences within 5 words is 40% or 50% respectively. In that case, if there were no optimal control scheme for the video switching, there would be too many scene changes.

The second suggestion was an opinion for the speech path. Though the conventional telephone interpreting service provides the mixed speech of the agent, interpreter and client, many clients want to hear the interpreter's speech only. We think that if the client does not hear the speech of the agent or the InsJc, they will probably lose the conversation flow whether they understand that language or not. We will verify that relationship in next our experiment.

The third suggestion is about the information presentation scheme of the interpreter. They wanted some cues when the interpreter starts typing, because the multimedia terminal presents many information windows simultaneously. When a client concentrated on a specific window, he or she would not know what was happening in other windows. Therefore, it is necessary to produce sound or some other kind of cue that presents the interpreter's behavior in the multimedia multimodal interpreting case. Also, many participants said that the interpreter should introduce herself. This is the another important design factor for the future machine interpreting system. The other is a waiting signal for the silent period during which the agent or the interpreter is preparing some information for the client's question.

5. Conclusion

We performed an interpretation experiment to investigate human communicative behaviors in the telephone-only and multimedia multimodal environment. From the experiment results, it is clear that multimedia and multimodal communication is necessary and very useful in interpreting telecommunication, and that the multimedia and multimodal system will provide a more convenient service to users. However, the multimedia and multimodal affects the dialogue pattern, and media processing schemes for conventional multimedia systems should be changed. The major results of this experiment are as follows:

1) Effects of the multimedia and multimodal

- The average communication duration, turns, and duration per turn for the MM call were greater than those of the TEL call. Also, the average number of words per call and the average number of words per turn in MM were greater than those of the TEL case.

- The disfluency rate in MM was lower than that of the TEL case. The agent and the client spoke more interjection words than the InsJc and the InsEc.

- There are many sentences which were affected not only by the multimodal and multimedia, but also by the multimedia only in the MM dialogues. Especially, the deictic expressions that were used for referent-identifications are quite different in TEL and MM.

- The subject's major gestures are the circling and the dragging gesture (88% of total gestures) to express their intention in direction finding task. The circling gestures were mainly used to express the location-indication, however, the main usage of the dragging gesture was direction expression.

- The timing between referent-identification words and circling gestures was clear. The 65% of the circling gestures started with deictic word or proper noun. But the timing between speech and dragging gesture was very random. Therefore, synchronization presentation between translated speech and dragging gesture require further examination.

2) Effects of the interpreting

- The turn distributions among the participants were different in comparison with the direct monolingual communication case in which both the agent and the client had the same number of turns.

- The InsEc and the agent spoke more words than the client and the InsJc. However, the difference in the words per turn between them in the interpreting call is smaller than in the direct monolingual communication case.

- The video switching function of the conventional videoconferencing system must be changed in the interpreting service case.

- It's better to reexamine the mixed speech scheme of the speech processing, because many clients want to hear the interpreter's speech only.

- When the interpreter is managing information, that was sent by the client or the agent, the store and forward method by machine is better than the interpreter's reaction.

These experimental results are based on a human interpreter. However, to build an automatic interpreting system, it is also necessary to know the characteristics of interaction between the user and the machine interpreter. However, it is impossible to know what such interaction would be like until such a system has been built. One of the solutions for these problems is the simulation based on the WOZ (Wizard of Oz) method. [8] A human plays the role of the machine in a WOZ experiment. We will collect free dialogues in a WOZ based situation to extract the requirements of the future multimedia machine interpreting system.

References

1. Meera M. Blattner, Roger B. Dannenberg, "Multimedia Interface Design", ACM Press, 1992.

2. Fumihiro Yato, Tsuyoshi Morimoto, Yasuhiro Yamazaki, and Akira Kurematsu, "Important Issues for Automatic Interpreting Telephone Technologies", Proc. of EuroSpeech-93, Berlin, Sep., 1993.

3. Kyung-ho Loken-Kim, Fumihiro Yato, Kazuhiko Kurihara, Laurel Fais, and Ryo Furukawa, "EMMI-ATR Environment for Multi-Modal Interactions", ATR Technical Report TR-IT-0018, Sep.,1993.

4. Ryo Furukawa, Fumihiro Yato, and Kyung-ho Loken-Kim, "Analysis of Telephone and Multimedia Dialogues", ATR Technical Report TR-IT-0020, Sep.,1993.

5. Kyung-ho Loken-Kim, Fumihiro Yato, Laurel Fais, Kazuhiko Kurihara, Ryo Furukawa, and Yoshihiro Kitagawa, "Transcription of Spontaneous Speech Collected Using a Multimodal Simulator--EMMI, in a direction-finding task (Japanese-Japanese; English-English)", ATR Technical Report TR-IT-0029, Dec.,1993.

6. Laurel Fais, "Structure in Spontaneous English Conversation", ATR Technical Report TR-IT-0040, Feb.,1994.

7. Christian Boitet and Kyung-ho Loken-Kim, "Human-Machine-Human Interactions in Interpreting Telecommunications", Proc. of EuroSpeech-93, Berlin, Sep., 1993.

8. Kyoung-ho Loken-Kim, Fumihiro Yato, Laurel Fais, Tsuyoshi Morimoto, Akira Kurematsu, "Linguistic and Paralinguistic Difference Between Multimodal and Telephone-only Dialogues", Proc. of ICSLP, Aug., 1994.

9. Norman M. Fraser and G. Nigel Gilbert, "Simulating Speech Systems", Computer Speech and Language, Academic Press Limited, 1991.