A SIMULATION ENVIRONMENT FOR MULTI-MODAL INTERPRETING TELECOMMUNICATIONS

Kyung-ho Loken-Kim, Fumihiro Yato, Tsuyoshi Morimoto

ATR Interpreting Telecommunications Research Laboratories

This report describes the hardware and software construction of the ATR Environment for Multi-Modal Interactions (EMMI) developed to collect spontaneous speech and language data used in multi-modal, mono-lingual, two-person telecommunication settings. The primary task that EMMI supports is travel informtion service and its subtasks; directions, reservations, and scheduling. Since ATR aims at the eventual replacement of human telephone interpreters with spoken language interpreting systems, we are currently upgrading the existing EMMI to accomodate bi-lingual, three-person, multi-modal communications. Problems involved with simulating multi-modal, bi-lingual situations and related issues are further discussed here.

マルチモーダル音声翻訳通信のためのシミュレータ

ローケン・キム　キュンホ、谷戸　文広、森元　逞

エイ・ティ・アール　音声翻訳通信研究所

本稿では、マルチモーダル音声翻訳通信における音声、および言語データを収集するため構築したシミュレータであるEMMI (ATR Environment for Multi-Modal Interactions)のシステム構成に関して述べる。EMMIのマルチモーダルな対話環境は音声、動画像、キーボード、マウスなどの入出力手段が利用可能なもので、タスクとしては旅行代理店とそのサーブタスクである道案内、要約、スケジュリングなどの模擬が可能である。現在は、異言語間、通訳者を介する３者対話の模擬会話の収集が可能になるようにシステムの改良が行っており、このようなマルチモーダル対話環境を用いた３者対話の収集を行う際に生じる問題点などを報告する。

1.INTRODUCTION

Recent developments of application specific ICs (ASICS) [1] open up the possibilities of realizing personal communicators that integrate voice, data, handwriting, fax, electronic-mail, still images, and full-motion video within a decade. No doubt such multi-media systems will have a profound impact on the world of communications, and they will change the form of human communications forever. It is not well understood, however, how these technologies should be optimally amalgamated to induce maximum efficiency in human-machine-human communications, especially in multi-media, multi-lingual, multi-party interpreting telecommunication settings. The optimal multi-media configuration for an application, such as multi-media 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 Multi-Modal Interactions (EMMI) [2] is a simulation tool that supports a variety of realistic environments for multi-media, mono-lingual (Japanese-Japanese, English-English), two-person telecommunications. EMMI is neutral in the sense that it does not contain any sort of intelligence; that is, it is simply a man-machine-man interface.

EMMI has been created specifically for collecting data about the speech and language people might use in multi-modal telecommunications. We have selected a travel information service task, and it is divided into three sub-tasks; directions, reservations, and scheduling.

Collecting mono-lingual multi-media speech and language data requires a minimum of two participants: one acting as an agent, and the other a client. In EMMI, participants can communicate with each other using a variety of input/output modalities: speech, text, graphics, and video image. 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.

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.

In this report, the authors: 1) describe the hardware and the software configurations of EMMI, 2) discuss the current limitations of EMMI, and 3) address issues related current limitations of EMMI, and 3) address issues related to the introduction of multi-modality to bi-lingual, three-person interpreting telecommunications.

2. USER INTERFACE

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. In order to distinguish the client's and the agent's marks and writings, different colors are 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 the 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, nevertheless, do not allow direct eye contact between the participants.

The INPUT and the 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.

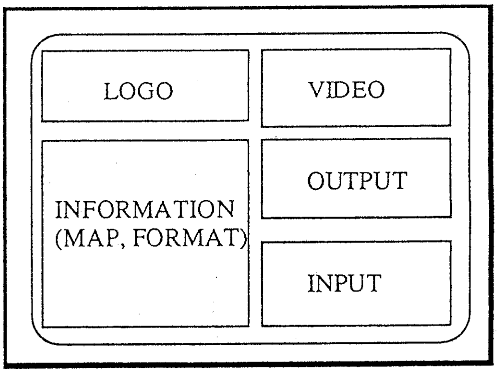

Figure 1. User Interface

3. TASK

The primary task that EMMI supports is international conference registration and its sub-tasks. Specifically, the following tasks are supported.

1) DIRECTIONS TASK

For this task, a client asks the conference secretariat for directions, for example, from Kyoto Station to the Kyoto International Conference Center. The agent gives the directions by displaying one of three maps of the areas surrounding Kyoto Station, the International Conference Center, and Kyoto Park Hotel. As previously explained, the maps are displayed on both of the agent's and client's screens, and the agent and the client can engage in a dialogue while marking and writing relevant information on the maps using a mouse. Figure 2 illustrates a mid-session screen of the client.

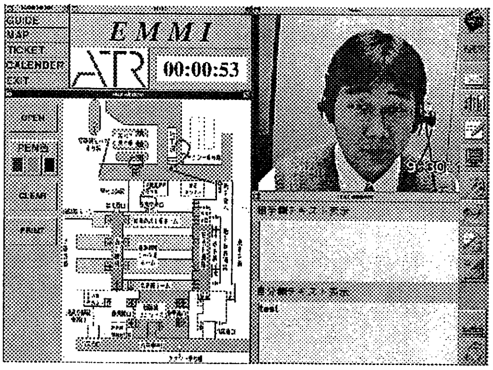

Figure 2. Client's Midsession Screen

2) RESERVATIONS TASK

In this task, a client needs to make a reservation, such as a hotel reservation. The agent displays a reservation form on the screen and makes the reservation by filling out the form using the keyboard and the mouse. The client may also fill out the form. Currently four different reservation forms are available: train, airline, hotel, and package tour.

3) SCHEDULING TASK

In this task, the agent negotiates with the client on the client's possible paper presentation date and time using the calendar displayed on the screen. It is assumed that the conference schedule has not been determined yet.

4. HARDWARE CONFIGURATIONS AND SCHEMATICS

Table 1 is a list of equipment used for EMMI. There are two NeXT computers: one for the agent, and the other for the client. A SUN Sparc Station has been allocated for the interpreter. All three computers are equipped with a keyboard and a mouse.

One Digital Audio Tape deck, microphone amplifiers, and 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. The line to the agent is interconnected with another telephone for the interpreter.

A scanner is attached to the SUN workstation to scan the maps and other graphic images.

Table 1. Hardware

Computers	Two NeXT Cube Workstations, SUN Sparc Station-10
Audio Equipment	Two SONY Digital Audio Tape Decks DTC-77ES
Amplifiers	Two ONKYO Integrated Stereo Amplifier A-812 EX, Two Technics Mic Mixing Amplifiers SH-3026
Headphones	Four SONY MDR CD850 Headphones, Three Sennheiser HMD 410 Headphones with microphones
Video Equipment	SONY Video Hi-8 Handycam PRO 3dd, Two SONY Video Hi-8 Handycams
Video Recorder	SONY Video Television Recorder
Telephone	Three telephones
Scanner	SONY Color Video Scanner UY-T55V

Figure 3 is the hardware schematic of EMMI. The section within the circle has not been completely installed..

5. SOFTWARE CONFIGURATIONS

EMMI has been developed on a NeXT computer using Interface Builder and Objective C [3]. The software construction of the client's side is almost the mirror image of that on the agent's side. As illustrated in Figure 4, the two top processes, i.e., start.new.csh on the agent's and the client's sides activate EMMI by sending messages to corresponding TEL processes. Upon receiving the messages, the agent TEL process displays

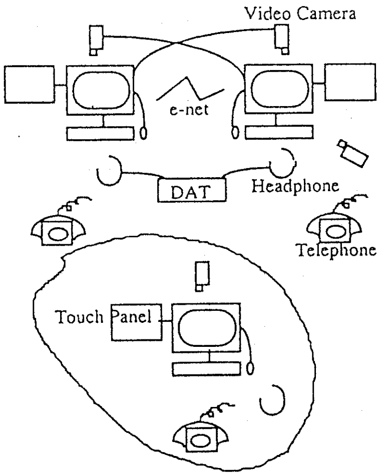

Figure 3. Hardware Schematic

the telephone window while the client TEL process displays a slightly different one. This is the initial idling state of EMMI waiting for a client.

When a client types in the telephone number of the conference center, the number is transmitted to the agent TEL process. This process, then, generates the sound of a telephone ring, and switches the telephone window to notify the agent the client's call.

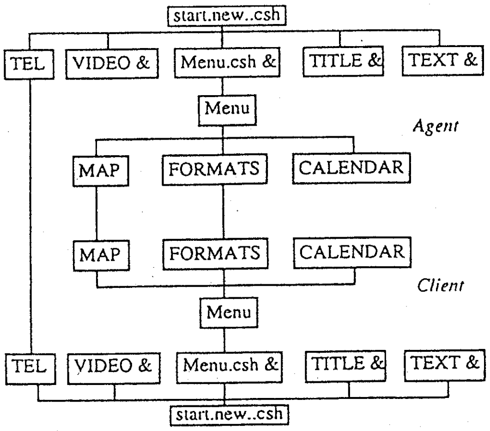

Figure 4. Software

The agent can answer the call by pushing the return key which automatically executes four background processes running in parallel: TITLE, TEXT, VIDEO, and Menu.csh. The TITLE process displays the EMMI logo window and the clock using the control object. The TEXT process displays input and output windows through which participants communicate using keyboards. The VIDEO process displays a video window on which full-motion video facial images of the participants are projected. The Menu.sch displays a small menu window at the top left corner of the agent's screen, so the agent can select items on the menu. This menu window, however, is not displayed on the client's screen.

For clients who read only English, there is an English version start.new.csh.eng process.

6. DISCUSSIONS

6.1. Future Improvement

As with any other experimental system, EMMI is going through a period of evolution and fine-tuning. Some of the current limitations can be eliminated simply by replacing the hardware settings. Marking and writing, for example, on the CRT screen with a mouse is awkward for people who are not accustomed to doing it. Our pilot study [4] indicates that many subjects prefer taking pen and paper notes instead of writing on the screen. Awkwardness in using a mouse may prevent them from using the mouse more actively. To improve the situation, we have added a touch-panel to each display enabling users to mark and write by simply writing on the screen with their fingers.

Another problem is that the present video camera positions do not allow direct eye contact between 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 which, of course, is impossible. Placing the combination of a mirror and half silvered mirror in front of the screen [6], although very cleaver, does not allow the users to touch the screen. With the current setting, it seems that one is talking to another who is looking away. However, this does not seem to cause a major problem because participants look at the information window most frequently and seem to keep the video image in their peripheral vision.

Record keeping is another area that needs to be improved. The client's interactions with EMMI, for example, were initially video-taped by placing a video camera behind the right-hand of the client. It turned out to be almost useless, because the client's writings and markings on the screen were too faraway and too small to be visually recognized. The situation was somewhat improved when the video camera was moved to behind the left shoulder of the client, and closer to the screen. In this position, although the camera had a better view of the screen, it was often obstructed by the instruction sheet the client was waving. To alleviate this problem, we have added a video down converter (Chromatek 9125) to EMMI, and now we are able to directly video tape any information appeared on the CRT screens.

6.2. Multi-modal Communications, and Task

Currently, EMMI supports only mono-lingual, two-person communications, and we are currently improving the system to accommodate bi-lingual, three-person, multi-modal dialogues. We have recently completed a newer version EMMI featuring an interpreter's station. The screens for the interpreter and the agent are more-or-less the same; they differ only in the features for displaying maps and other objects, i.e., the agent has a window that allows to display maps and formats.

In addition,.the interpreter can see both the agent and the client through the VIDEO window (Figure 1), but both participants can not see the interpreter. This imitation is rather intentional because we want to collect speech and language multi-modal data generated under the simulated situation where the users think they are interacting with a machine, not a human interpreter. Wizard of Oz (WZ)[8] technique is commonly used for experiments where a subject is told she/he is interacting with a computer, but in fact a human operator mimics the behavior of the computer. Using the technique, when the communication is uni-modal, e.g., typing, is fairly straightforward. However, designing a multi-modal WZ simulator, especially for bi-lingual, three-party interpretation setting is tricky even for a fairly restricted task due to the complexity of synchronizing multimodal information [9].

Selecting proper tasks that can take advantage off multi-modal communication is also an important issue. Our pilot studies [4, 5] suggest that tasks, such as, scheduling in which two speakers attempts to setup a date for a meeting by looking at the calendar is not very well suited for a multi-modal communications simply because, in reality, people do not want to share their private schedules with other people. On the other hand, we found that maps and video windows are extremely useful for directions tasks.

7. CONCLUSIONS

In this paper, the authors 1) introduced the system architecture of ATR's Environment for Multi-Modal Interactions - EMMI, and 2) discussed some issues introducing multi-modality to interpreting telecommunications: We are currently upgrading EMMI so bi-lingual, three-party, multi-modal spoken language data can be collected in the future.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to Dr. Yamazaki, the president of ATR Interpreting Telecommunications Research Laboratories, for all his 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] Lauren Burst, *Mean-SEA Tsay. "Mixing signals & voltages on chip,"* IEEE SPECTRUM (8/1993).

[2] K.H. Loken-Kim, Fumihiro Yato, Kazuhiko Kurihara, Laurel Fais, Ryo Furukawa, "*EMMI - ATR Environment for Multi-Modal Interactions,*" ATR Technical Report TR-IT-0018 (1993).

[3] Bruce F. WEBSTER. "*The NeXT Book,*" Eddison Wesley Publishing Company (1989).

[4] Ryo Furukawa, Fumihiro Yato, K.H. Loken-Kim, "*Analysis of Telephone and Multimedia Dialogues,*" ATR Technical Report TR-IR-0020 (1993).

[5] Yamamoto, Fumihiro Yato, K.H. Loken-Kim, Kazuhiko Kurihara, Kitagawa, Akira Kurematsu, "*Analysis of Telephone-only and Multimodal Scheduling Dialogues,* " ATR Technical Report TR-IT-0027 (1993).

[6] Bill Buxton, "*Telepresence: Integrating Shared Task and Personal Spaces,*" in the Proceedings of Groupward'91, (1991).

[7] Sharon Oviatt, Philip Cohen, Michelle Wang, and Jeremy Gaston, "*A Simulation-Based Research Strategy for Designing Complex NL Systems,*" ARPA Workshop on Human Language Technology (1993).

[8] P. Green and L. Wei-Hass in Human-Computer Interface Design Guidelines (C. Brown), Ablex Publishing Corporation (1988), "*The rapid development of user interfaces: experience with the Wizard of Ox method,*" Proceedings of the Human Factors Society - 29th Annual Meeting (1985).

[9] Christian Boitet and K.H. Loken-Kim, "*Human-Machine-Human Interactions in Interpreting Telecommunications,*" in the Proceedings of International Symposium on Spoken Dialogue (1993).