RWC MULTIMODAL DATABASE FOR INTERACTIONS BY INTEGRATION OF SPOKEN LANGUAGE AND VISUAL INFORMATION

S. Hayamizu, O. Hasegawa, K. Itou, K. Sakaue, K. Tanaka (ETL),
S. Nagaya, M. Nakazawa, T. Endoh (RWCP-TRC),
F. Togawa, K. Sakamoto (RWCP-Sharp), K. Yamamoto (Gifu-Univ)

Electrotechnical Laboratory,
1-1-4 Umezono, Tsukuba 305 Japan

ABSTRACT

This paper describes our design policy and prototype data collection of RWC (Real World Computing Program) multimodal database. The database is intended for research and development on the integration of spoken language and visual information for human computer interactions. The interactions are supposed to use image recognition, image synthesis, speech recognition, and speech synthesis. Visual information also includes non-verbal communication such as interactions using hand gestures and facial expressions between human and a human-like CG (Computer Graphics) agent with a face and hands. Based on the experiments of interactions with these modes, specifications of the database are discussed from the viewpoint of controlling the variability and cost for the collection.

1. INTRODUCTION

Recently multimodal interaction between human and computer has attracted keen interest in the research field of speech and image processing[1-10]. But little effort has been focused on the construction of multimodal database for the integration of spoken language and visual information.

Especially, a multimodal database is necessary for the development of a system which interacts with a human by recognition and synthesis of spoken language and visual information in a way that multiple modes are integrated for the enhancement of interactions.

ETL collected speech and images of humans using a Wizard of OZ simulation to analyze non-verbal elements for maintenance of dialogue[1]. RWCP-Sharp collected and analyzed multimodal interactions between humans[2]. Based on these experiences, we started a working group to construct a multimodal database under the framework of Real World Computing Program in November 1994. The working group has had discussions about the database since then.

The database is intended for research and development of integration of multiple modes combining image recognition, image synthesis, speech recognition and speech synthesis.

Many points in the multimodal database should be clarified before specifications of the database are decided. What interactions should be collected and the way in which they could be collected should be clarified. Since these points have not yet fully been understood, we had to conduct experiments to collect pilot data.

The design issues (what and how) and prototype data collections are described in the following sections.

2. OBJECTIVES

Spoken language has been used as the media to exchange information between humans. Thus it is interactive by nature. Various types of speech database (or corpus) have been constructed and used in the research field of speech.

On the contrary, database of visual information in the interactions are very few compared with speech database. Sometimes humans use visual information to exchange information in the interactions with others. One way is to use their body (hands, face, etc.) movements to express some information. The other is to give some information to others by showing some objects. Also, humans use visual information with spoken language simultaneously or sequentially.

As far as the authors know, these kinds of multimodal interactions have not yet been organized into database. Here, construction of database for multimodal interactions has two objectives.

The first objective of the database is to share the knowledge about what kinds of information are used in the interactions between humans. Hopefully, these types of interactions are useful and seem to be natural in the interactions between human and computer, too.

The second one is to collect some considerable size of multimodal data to train and to test some modules of the system, such as speech recognition, image recognition, and/or integration module. The collected data can be used for the train-

[1] ETL : Electrotechnical Laboratory
[2] RWCP-TRC : Tsukuba Research Center, Real World Computing Partnership
[3] RWCP-Sharp : Novel Functions Sharp Laboratory, Real World Computing Partnership
[4] Gifu-Univ : Gifu University

ing of statistical methods in the modules or can be used for evaluation purposes. Common task and common database will be very effective for the progress of the research field.

3. CONTENTS OF DATABASE

Multimodal interactions have two aspects. One is integration of multiple modes and the other is the non-verbal aspect of communication.

Typical integration is a complimentary one, such as "put that there" paradigm. One channel (for example speech) becomes complimentary to the other channel (image) and the integration enhance the interactions.

Body movements play the role of exchanging non-verbal information (prosody has a similar role in speech). Body movements can be classified as follows (modified after [3]);

- sign,
- exemplar,
- continuous control,
- emotion,
- turn-taking, and
- concern.

Human can understand and express all of these movements. But the current systems have limited abilities in these points. That is, systems with current technology of automatic recognition and synthesis can understand and express only some parts and to some limited extent. So, we have to select some cases from the viewpoint of easiness in processing by the systems.

We raised the following categories as candidates of the data to be collected.

(a) Movements of head (nodding and shaking head horizontally) and speech (yes/no).

These are used to show affirmative or negative responses to others. Three aspects, number of times, amplitude, and speed of movements, are specified. For strong negative gestures, a hand is also shaken.

(b) Speech and hand gestures to show directions (upwards, downwards, right, left, here, there) and others (size, emphasis, etc.).

To show directions, one way is to use a palm of a hand and the other is to point by using an index finger.

These categories are selected for task-oriented interactions between human and computer. They are intended to represent basic movements which could be used as input to the system by integration of image recognition and speech recognition. And they are expected to serve as test samples for speech and/or hand gesture recognition.

Hand gestures represented by movements of hands are selected for the first step of data collection. Gestures represented by the shape of hand (movements of fingers) are to be collected in the second step.

4. CONDITIONS AND METHODS

In order to clarify the problems for collection of multimodal data, preliminary experiments were conducted under the following conditions and methods.

4.1. Types of Interactions

The interactions tested in the experiments were as follows.

(i) Between humans in two separated rooms with transmitted speech and images,

(ii) between a human and re-produced video images and speech which had been recorded in advance, and

(iii) between a human and a CG agent [Figure 1] with a face, hands, and synthesized speech [IMAGE A404G01.GIF].

These interactions were basically different in several points. The first one was easier to collect and would be more natural but it was more difficult to control the conditions to be usable for system development. It is because humans tend to behave more freely and they go beyond the scope which systems can deal with. The other reason is that preparation of interactions needs much effort and situations of multimodal interactions do not happen so frequently.

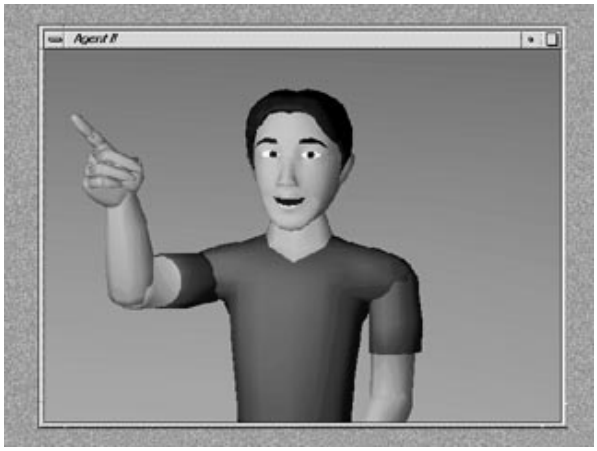

Figure 1: Interactions with a CG Agent.

4.2. Recording Environment

The face and upper body of humans were recorded. Figure 2 is an example of the image [IMAGE A404G02.GIF]. The lower left is the face of the subject. The upper left is the

upper body of the subject. The upper right is the upper body of another person. The lower right is the display which the subject sees. The four images were composed into one image as each image has a quarter size.

Black curtains were used as the background. Humans wore blue shirts with green, red and purple markers, for grips, elbows and shoulders, respectively. If necessary, these markers could be used to make the first step of image recognition easier.

Speech of two humans (or a human and synthesized speech) were recorded in separate channels (left and right channels of stereo sound input). The images of human's face and body, and that of a CG agent, the speech of two talkers were recorded simultaneously in the beta-cam tapes.

Images were recorded using digital cameras (SONY DCR VX-1000). Speech were recorded using a close-talking head-mounted microphone (SENHEISER HMD 410).

Figure 2: An example of an image of the interaction between humans. The upper-right person tells and shows by pointing the block to be selected.

4.3. Methods

Three cases were tested for recording.

(1) Instructions given in the form of written texts.
(2) Imitation of pre-recorded hand gestures and speech.
(3) Interactions in task of object construction with colored blocks.

(1) Instructions given in the form of written texts.

The intention for each gesture and utterance was written on the instruction sheet. The subject was asked to read it and behave accordingly. It was found that variations between humans were considerably large. It was difficult to specify timing and movements of the hand gesture in details. But this method seemed good to find out what gestures and utterances would be used to represent some intentions.

(2) Imitation of pre-recorded hand gestures and speech.

One person's hand gestures and utterances were recorded in advance. Then these recorded video and speech were shown to the subject. The subject was asked to imitate each hand gesture and utterance as it was. This method was more efficient and easier to collect many data than other ones. The variations between humans were reasonably small. They would fit for the first step of research and development as training and testing materials, for example, image recognition of hand gestures, and integration of the image recognition and speech recognition.

(3) Interactions in task of object construction with colored blocks.

This case was intended to collect natural hand gestures and utterances in the actual interactions of goal-directed task. But the interactions were constrained considering limitations in the current recognition technology.

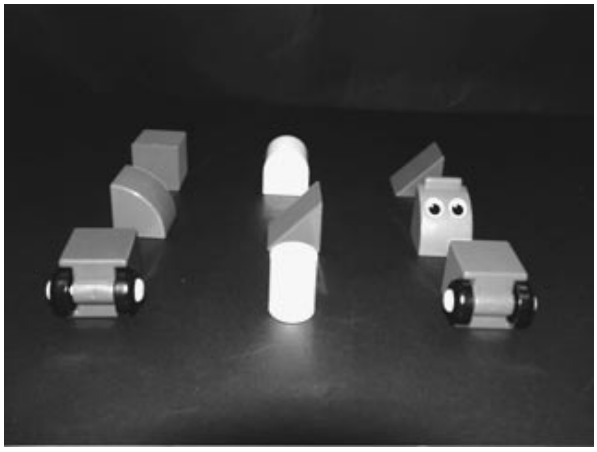

Figure 3: Colored blocks to be selected in the object construction task.

Colored blocks were used in the task. Each block has one of four colors, red, blue, yellow, green and has a simple shape such as a cube, a cylinder, and a triangular prism. Some objects such as a car and a ship can be constructed by combining these blocks. They were selected in order to make image recognition an easier task. Figure 3 shows the colored blocks to be selected in the task [IMAGE A404G03.GIF].

The subject saw the upper body of the other person and the upper body of himself/herself on the monitor (display). The situations were that one person knew the way to construct and the final shape to be constructed and the other person did not know them. The person was asked to inform the other by hand gestures and speech.

Two persons were in the separate rooms. Images and speech were submitted to see and to hear each other. Images of the other person's upper body were displayed on the monitor (see the lower-right of Figure 2). The person's own speech and the other person's speech were given in the separate channels of head phones.

If this task was done without any constraints, people had the tendency to inform the other only using speech. The other kept silent and saw only the blocks. Under this situation, the communication was almost one way.

Considering this experience, usages of information in the communication were constrained so that multimodal interactions occurred more frequently. Usage of the vocabulary was restricted to the words for colors and directions. By this constraint, it was expected that hand gestures for directions would occur. For example, when there were two red blocks and ambiguity about the selection, it could be decided by hand gestures.

This task consisted of three stages;

- selection of blocks (pointing by finger and speech),
- changing the position of blocks (pointing, direction of movement and speech), and
- construction of target objects (rotation, direction, emphasis, etc. and speech).

In all stages, subjects were asked to use both of speech and hand gestures. Interactions with a CG agent were tested only for the first and second stages. The interactions were controlled by the WOZ (Wizard of OZ) simulation. Another person saw the subject and heard his voice. He controlled body movements of the CG agent by keyboard. He also controlled speech synthesis of sentence templates by keyboard.

The order of selection, that of changing the position were decided precisely in advance. Under these constraints, hand gestures and speech had relatively small variations. They could be used for the analysis and system design of the interactions with spoken language and visual information.

But the cost of collection in this situation was large compared with the collection by imitation. At present, we have a plan to put the first priority on the collection by imitation.

5. CONCLUSION

Design policy and prototype data collection of RWC multimodal database were described.

In order to realize interactions by integration of spoken language and visual information, it is necessary to know the way of such interactions which humans do and to know the limited capability of the interaction system with current technology.

As strategy to promote the progress of research and development, we plan to construct multimodal database with speech and images for such interactions.

Through the prototype data collections, specifications of the database were decided. Two types of data are on the way of collections. One is the database of short hand gestures with speech which are collected by imitation. This is intended mainly for the training and testing materials of vision and/or speech recognition modules. The other is the compilation of interactions under the limited usage of information for the task of block constructions. This is intended for the analysis and system design of such interactions.

The database is planned to be available outside of RWCP (the release date has not yet been decided). The sample data of the database will be available on CD-ROMs.

ACKNOWLEDGMENT

This research was done under the Real World Computing Program. The authors would like to thank Mr. Jiro Kiyama and Mr. Susumu Seki for their contribution to the working group, and Miss Naoko Ueda and Miss Yukiyo Uehori for their technical assistance in the experiments.

REFERENCES

[1] K. Itou, T. Akiba, O. Hasegawa, S. Hayamizu, K. Tanaka : "Collecting and analyzing nonverbal elements for maintenance of dialog using a wizard of OZ simulation", ICSLP-94, S17-10.1, pp. 907 - 910 (1994).

[2] K. Watanuki, K. Sakamoto, F. Togawa : "Analysis of multimodal interaction data in human communication", ICSLP-94, S17-8.2, pp. 899 - 902 (1994).

[3] T. Kurokawa : "Nonverbal interface", Ohm-sha (1994).

[4] R. A. Bolt : "Put that there : Voice and gesture at the graphics interface", Computer Graphics, Vol.4, No.3, pp.262-270 (1980).

[5] Y. Suenaga, K. Mase, M. Fukumoto, Y. Watanabe : "Human reader : an advanced human machine interface based on human images and speech", Trans. IEICE, Vol. J75-D-II, No.2, pp. 190- 202 (1992).

[6] M. T. Vo, A. Waibel : "Multimodal human-computer interaction", Proc. Int. Symp. on Spoken Dialogue, pp. 95-101 (1993).

[7] K. Nagao, A. Takeuchi : "Speech dialogue with facial displays : Multi-modal human-computer conversation", Proc. ACL, pp. 102-109 (1994).

[8] O. Hasegawa, K. Itou, T. Kurita, S. Hayamizu, K. Tanaka, K. Yamamoto, N. Otsu : "Active agent oriented multimodal interface system", Proc. IJCAI-95, pp. 82-87 (1995).

[9] R. Oka, J. Kiyama, H. Kojima, Y. Itoh, S. Seki, and S. Nagaya : "Real-time integration of speech, gesture, graphics, and database", 95 RWC Symposium (1995/06).

[10] O. Hasegawa, K. Itou, H. Asoh, S. Akaho, T. Akiba, T. Kurita, S. Hayamizu, K. Sakaue, K. Tanaka, N. Otsu : "Human factor analysis in human computer interaction", Technical Report of IEICE, PRU95-57 (1995).