

# Multimodal Interaction System that Integrates Speech and Visual Information

Satoru HAYAMIZU, Osamu HASEGAWA, Katunobu ITOU  
Takashi YOSHIMURA, Tomoyoshi AKIBA, Hideki ASOH  
Shotaro AKAHO, Takio KURITA, Katsuhiko SAKAUE

This paper presents the studies related to multimodal interaction systems. It also describes our new direction in the research, 'Intermodal Learning'. The prototype system has four modes: vision, graphical display, speech recognition, and speech synthesis sub-systems, and an interaction manager. We demonstrated that it can learn user's face and name and the appearance and names of objects. A speech recognition technique to estimate phonetic transcriptions from multiple speech samples was used to learn new words. This is similar to a baby learning about the real world by communicating with its parents.

## § 1 Introduction

To achieve communication between humans and machines through the integration of spoken language and visual information, it is important to learn how machines use media to interact with humans. This is similar to a baby learning about the real world by communicating with its parents.

The work presented here aims at a system which will be able to see, hear, talk, and act. This system will observe human behavior and the environment with a vision module. It hears spoken language with a speech-recognition module, and responds by using a speech-synthesis module. The system generates facial expressions and hand gestures (postures) with an image-synthesis (computer-graphic) module. The integration of these modules will enable the system to interact with humans smoothly and naturally.

Many works on multimodal interaction have been reported<sup>4,6,18-23</sup>. Research on multimodal interaction systems can be classified into three categories:

- information integration,
- non-verbal communication,

- interactivity.

These three categories are related, to each other but most work done has treated them individually, not as an integrated unit.

There are several advantages to using the multimodal interaction system which these works have tried to bring about. They are:

- a) Information integration will enhance the existing functions of communication. For example, the integrating speech recognition with sensors and pointing devices will ease the task of speech recognition by limiting the vocabulary required for cases specified by pointing. And the integration of visual recognition of lip movements with speech recognition will improve the recognition accuracy in a noisy environment<sup>22</sup>.
- b) Recognition and synthesis of gestures (hand postures) and facial expressions will provide a natural and familiar interface for humans. Information that is difficult to convey by using the written word can be more easily communicated using other modes. For example, non-verbal information can be

expressed by facial expressions generated by computer graphics.

- c) Smooth interactions can be realized by a new way of communicating, that controls temporal (time domain) aspects in a sophisticated way. Maintenance of the interactions will be more natural if the control of speaking in turns, such as ‘chiming in’, interruption, and natural recovery, can be done automatically.
- d) The integration of multiple sources of information will create a new function that cannot be realized by any other method. For example, integration of facial identification and speech synthesis enables a state of interactivity where the system initiates the interaction by speaking talking first. Another example is that the integration of the facial identification capability of the vision module and the spoken dialogue (speech recognition and speech synthesis) system enables the transfer of messages from one person to the system and then from the system to the other person<sup>8,14)</sup>.

These multiple modes can be integrated either complementarily or independently, and the integration of inputs and outputs is important.

Although the advantages cited above are widely accepted, their reality is still far beyond the state of current technology. For non-verbal information, it remains unknown even what type can be used to create an interaction between humans and machines in a real-world application.

This paper summarizes our efforts to develop a multimodal interaction system, explain our new research, and describes our prototype system.

## § 2 Studies on Multimodal Interaction System

Research on multimodal interaction systems is interdisciplinary in nature. It is related to various fields of study, such as speech recognition, speech synthesis, compute vision, computer graphics, natural language, mathematical theory, and cognitive science.

In our research we have attempted to clarify the

problems in a multimodal interaction system by collecting data on speech and images during the interaction of humans and machines by performing statistical analysis of facial images, and by developing a prototype multimodal interaction system<sup>1,8,14)</sup>.

First, we collected the non-verbal elements of interaction (facial expressions, prosody of speech, etc.) using Wizard-of-Oz (WOZ) simulation. The subjects were asked to speak into a spoken-dialogue system to collect data on the speech, facial expressions and gestures used during interactions between the subjects and the system<sup>13)</sup>.

The WOZ technique is one in which a person (called the wizard) plays the role of the system in a simulated subject/agent interaction<sup>5)</sup>. Our WOZ experiment was designed to collect data to allow an analysis of the elements that enable interaction between humans and machines (**Fig.1**). In particular, we designed the experiment to collect the data that identify the non-verbal aspects of an interaction. Analyzing the collected data, we identified various phenomena related to human behavior and the system’s responses.

Figure1 shows the experimental setup. The subject’s room is separated from the wizard’s (person playing role of system), to hide the wizard from the subject. The wizard listens to the subject with the speaker (a) and observe them on the display (c) through the microphone (g) and the video camera (d). The wizard runs the program of spoken dialogue through a

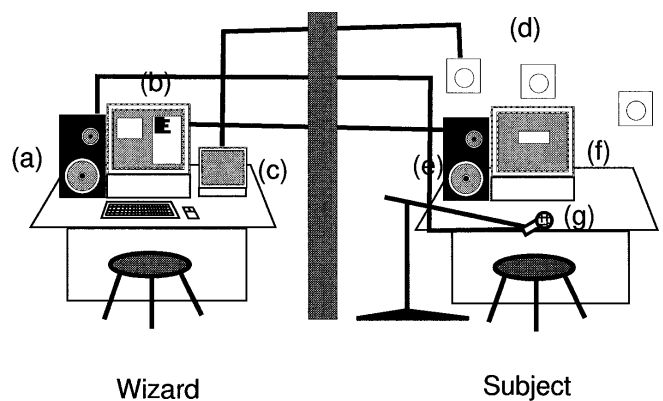


Fig.1 Experimental setup of WOZ system

workstation (b), and the system's responses are output through speaker (e) in the subject's room. The workstation (f) is used only to display the task being performed.

After performing the WOZ simulation, we did a statistical analysis of the facial images collected. By clustering the data on the locations of positions marked for the different facial expressions, some interesting features were found to correspond to characteristic expressions and facial movements during the simulation<sup>7</sup>.

Next, we developed a prototype multimodal interaction system incorporating four modes<sup>8,14</sup>. The purpose of the prototype development was to identify the actual problems, and advantages and disadvantages of a multimodal interaction system.

The four modes of our prototype system are speech recognition, speech synthesis, computer vision, and a visual agent generated by computer graphics. These modes are integrated into one system that nominally replicates the functions of the human eyes, ears, mouth, face, and upper body.

This prototype shows how an 'active' agent can be created by the integration of multiple modes. Our prototype displays a moving human-like agent with realistic facial expressions to promote smooth interaction with users. The agent can identify the user on sight and interact in spoken language. The agent can initiate spoken interaction by talking to the user first. The agent can respond differently to individual users. These different responses are made through the integration of image and speech recognition/synthesis technologies.

Next, we developed a language called MILES (Multimodal Interaction LEading Script) which can express time relations between events and communicative elements in dialogues<sup>3</sup>. We also developed a human-like software robot which interactively learns and manages visual information in real world<sup>10</sup>. And a gesture recognition was developed which uses HLAC (Higher order Local Auto-Correlation) features of PARCOR (PARTial auto-CORrelation coefficient) images and

HMM (hidden Markov Model) based recognizer<sup>17</sup>.

### § 3 Intermodal Learning

From our experience in pilot studies on multimodal interaction systems, we are convinced that 'learning by integration' will become the key focus of the new direction research will take.

Problems to be solved can be summarized as follows:

- How can we reduce cost in designing interactions that are necessary to accomplish a given task ?
- How can we define non-verbal elements for smooth interactions and specify them in detail ?
- How can we describe the relationship between situations and their contents in the flow of interactions ?

'Learning' will be a general solution to solve these problems. Interactive learning about superficial pattern processing will be the main target. In this learning, multiple sources of information (multiple media and modes) are given to the system and the patterns are reciprocally used for supervision.

We call this 'Intermodal Learning', in the sense that the learning about the usage of media is done through interaction with humans by multiple-mode integration. Under this framework, we have studied the learning of a user's name and face, and the appearance and names of objects<sup>9</sup>. We have also studied the learning of the attributes of colored objects using the Expectation Maximization (EM) algorithm<sup>2,11</sup>.

The prototype system presented here learns the relationships between the user's name and face, and between the appearance and name of objects. First, a user shows a sequence of images to the system and tells the system his or her name. The system learns the relationship. Then, when a sequence of images is shown to the system and the name is asked, the system replies the name by using a speech synthesizer.

This type of learning is useful for the cases where it is difficult to define categories in advance, or those where variations in the linguistic expression are large and unpredictable. By utilizing the actual utterances,

## § 4 Overview

explicit definition of categories and linguistic expression are not needed. This process is also simple and easy to apply.

The speech recognition method used here is to acquire phonetic transcription automatically from speech samples<sup>12,15)</sup> to learn new words through interactions with images and spoken language. The target of the learning is the vocabulary and its usage about images. The method uses information from multiple speech samples that have the same transcription in common<sup>15)</sup>. It first estimates multiple (N-best) candidates of phonetic transcriptions for each speech sample using a phonetic typewriter that can recognize any combination of phonetic sequences in Japanese. Then, it re-estimates a transcription of the word from those candidates.

The estimated phonetic transcription is used for speech synthesis and for further interactions. An image frame is cut from a time sequence of the images during the interactions. A standard pattern for the object (or face) is defined using these images.

This paper presents a prototype multimodal interaction system developed at the Electrotechnical Laboratory. The prototype system has four sub-systems (modules): speech recognition, speech synthesis, computer vision, and a visual agent generated by computer graphics. These are integrated into one system which nominally replicates the functions of the human eyes, ears, mouth, face and upper body.

This system consists of four sub-systems (modules) and an interaction manager (**Fig.2**).

### (1) Vision sub-system

The vision sub-system uses the higher order local auto-correlation features (**Fig.3**) of the input variable density image (90 × 60 pixels). A learning mechanism based on linear discriminant analysis or multiple regression analysis is used to identify the user<sup>16)</sup>. By using this method, invariant recognition is possible with respect to the position of the facial image within the

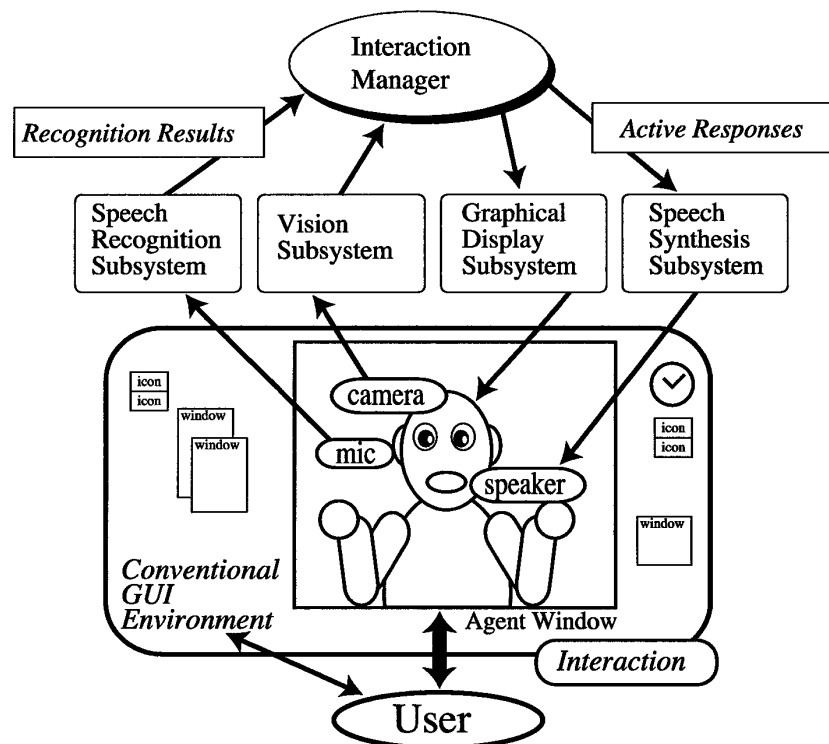


Fig.2 Configuration of the prototype multimodal system

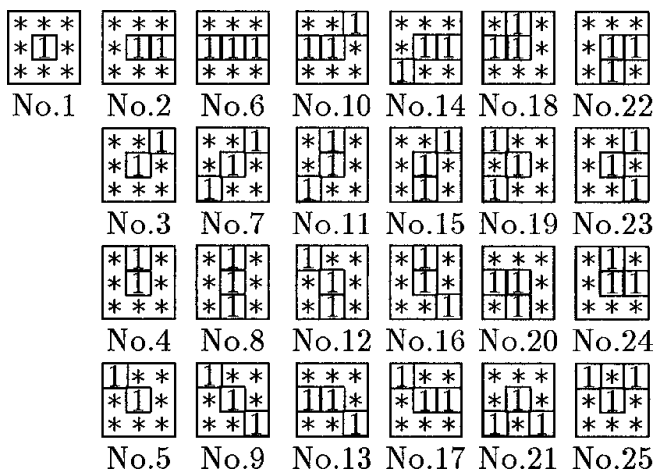


Fig.3 Higher-order autocorrelation features

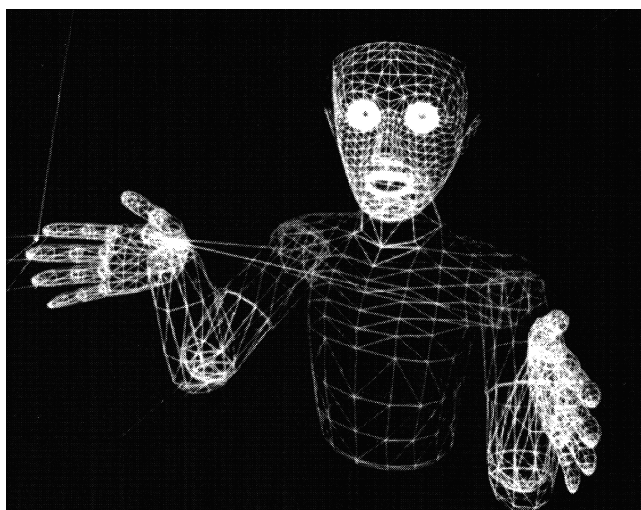


Fig.4 Wireframe models of the CG agent

screen. A time sequence of images is cut during the interactions. A standard pattern for the object (or face) is defined using these images.

(2) Graphical display sub-system

In the first version of our facial display sub-system, the face of the agent was generated by computer graphics (CG). The face was composed of approximately 500 polygons and was modeled three-dimensionally. The appearance of the face was rendered by using the texture mapping technique. The eyebrows, eyeballs, eyelids, mouth and head orientation of the facial model were movable. The system was designed to control the agent's eyes so as to maintain contact with the user during the interaction.

Our latest version of the sub-system shows a waist-up view (a head, two hands and an upper body. The face is composed of 1700 polygons, and the other body parts are composed of 3000 polygons. The agent now has hand gestures (postures) in addition to its facial expressions. **Figure 4** shows the wireframe models of the agent.

(3) Speech recognition sub-system

In the speech recognition sub-system, a block consisting of a collection of 40 frames (frame period is 10 ms) is used to cut out the speech. Each sequence of continuous blocks is assumed to be one utterance and is

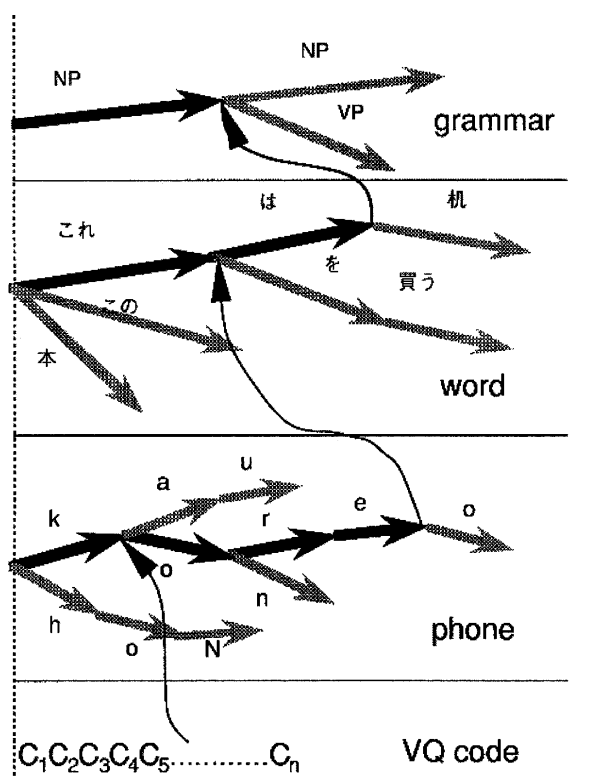


Fig.5 Search space of speech recognition

recognized as continuous speech. **Figure 5** shows the search space of speech recognition in which hypotheses at different levels are linked. The N-best hypotheses are maintained in this tree-structured search method. The result of continuous speech recognition is converted into a specified task command and sent to an interaction

manager. The speech recognition method acquires phonetic transcription automatically from speech samples. It first estimates multiple (N-best) candidates of phonetic transcriptions for each speech sample using a phonetic typewriter.

#### (4) Speech synthesis sub-system

In the speech synthesis sub-system, a commercial synthesizer is used. The speech synthesizer uses software to convert a Kana/Kanji text sentence produced by the interaction manager into a readable format to which accent information and so on is added. Next, the readable format is transmitted to the synthesizer.

#### (5) Interaction manager

The whole system is controlled by an interaction manager. The interaction manager receives messages (recognition results) from both the vision and speech recognition sub-systems. To obtain a high level of accuracy, it examines the order of the received messages and discards inadequate ones by following the state of the dialogue. It then analyzes the messages received and generates commands to control the display and speech synthesis sub-systems.

## § 5 Sample Scenario

**Figure 6** shows our prototype multimodal interaction system. Sample scenarios to learn the name of a person and an object is as follows. The original dialogue is in Japanese.

(User-1 sits in front of camera and looks at monitor. The visual agent initiates a conversation while it learns user's face.)

System : Hello.

System : May I ask your name ?

System : Please repeat it three times.

User-1 : Hasegawa, Hasegawa, Hasegawa.

System : Thank you, 'Hasegawa'-san.

(User-1 shows a bear doll to the system. The system looks at the doll and learns a sequence of its images.)

System : 'Hasegawa'-san, what's that ?

System : Please repeat it three times.

User-1 : Bear, bear, bear.

System : I understand.



**Fig.6** Interactions with the prototype multimodal system

(User-2 shows the doll to the system and asks its name.)

User-2 : What's this ?

System : It's 'Hasegawa'-san's 'Bear'.

.....

Note that during a period of spoken language interactions the system learns about images and associates acquired phonetic transcriptions with the images. Then the system responds using the acquired transcription ('Hasegawa') in speech synthesis when it sees his face. The system also responds using the acquired transcription when it sees images of an object and is asked its name.

This framework is different from the usual one where the vocabulary is defined a priori. The system will learn any new words as long as the user shows the images and tells their names to the system.

This type of learning is flexible in selection of vocabulary for spoken language interactions. For example, a user may call the object a 'doll', a 'bear', 'a bear doll', etc., but the system can use the actual utterances to determine the words (phonetic transcriptions) through interactions.

## § 6 Conclusion

This paper presents the studies related to multimodal interaction systems. It also describes our new direction in the research, 'Intermodal Learning'. The prototype system has four modes: vision, graphical display, speech recognition, and speech synthesis sub-systems, and an interaction manager.

We demonstrated that it can learn user's face and name and the appearance and names of objects. A speech recognition technique to estimate phonetic transcriptions from multiple speech samples was used to learn new words.

## Acknowledgement

These works have done under the Real World

Computing Program. The authors would like to thank those concerned.

## References

- 1) Akaho, S., Hayamizu, S., Hasegawa, O., Itou, K., Akiba, T., Asoh, H., Kurita, T., Sakaue, K., Tanaka, K., Otsu, N.: Recent Developments for Multimodal Interaction by Visual Agent with Spoken Language, Proc. ICMI-96 (1996).
- 2) Akaho, S., Hayamizu, S., Hasegawa, O., Yoshimura, T., Asoh, H. : Concept acquisition from multiple information sources by the EM algorithm, IEICE Trans. A, Vol. J80-A, No.9, pp. 1546-1553, (1997) (in Japanese).
- 3) Akiba, T., Kamishima, T., Itou, K. : MILES: multimodal interaction leading script, which can express time relations between events and communicative elements in dialogues, Technical Report of IEICE, NLC97-53, SP97-86 (1997) (in Japanese)
- 4) Bolt, R. A.: Put that there : Voice and gesture at the graphics interface, Computer Graphics, Vol.4, No.3, pp.262-270 (1980).
- 5) Fraser, N.M., Gilbert, G.N.: Simulating speech systems, Computer Speech and Language, Vol 5, No.1, pp. 81-99 (1991).
- 6) Hasegawa, O., Yokosawa, K., Ishizuka, M.: Real-time parallel and cooperative recognition of facial images for an interactive visual human interface, Proc. of 12th ICPR, Vol. 3, pp. 384-387 (1994).
- 7) Hasegawa, O., Itou, K., Asoh, H., Akaho, S., Akiba, T., Kurita, T., Hayamizu, S., Tanaka, K., Sakaue, K., Otsu, N.: Human factor analysis in human computer interaction, Technical Report of IEICE, PRU 95-57 (1995) (in Japanese).
- 8) Hasegawa, O., Itou, K., Kurita, T., Hayamizu, S., Tanaka, K., Yamamoto, K., Otsu, N.: Active agent oriented multimodal interface system, IJCAI-95, pp. 82 - 87 (1995).
- 9) Hasegawa, O., Sakaue, K., Itou, K., Kurita, T., Hayamizu, S., Tanaka, K, Otsu, N.: Agent oriented interactive image learning system, Proc. of Symposium on Sensing via Image Information, SII-96, pp. 145 - 150 (1996) (in Japanese).
- 10) Hasegawa, O., Sakaue, K., Hayamizu, S. : A human-like software robot which interactively learns and manages visual information in real world, IEICE Trans. D-II, Vol. J82-D-II, No. 10, pp. 1666-1674 (1999) (in Japanese).

- 11) Hayamizu, S., Akaho, S., Hasegawa, O., Yoshimura, T, Asoh, H.: Intermodal learning between speech and visual information, Information Intelligence Media Symposium 96, pp. 61-68 (1996) (in Japanese).
- 12) Itou, K., Hayamizu, S., Tanaka, H.: Detection of unknown words and automatic estimation of their transcriptions in continuous speech recognitions, Proc. ICSLP-92, pp. 799 - 802 (1992).
- 13) Itou, K., Akiba, T., Hasegawa, O., Hayamizu, S., Tanaka, K.: Collecting and analyzing nonverbal elements for maintenance of dialog using a wizard of OZ simulation, ICSLP-94, pp. 907 - 910 (1994).
- 14) Itou, K., Hasegawa, O., Kurita, T., Hayamizu, S., Tanaka, K., Yamamoto, K., Otsu, N.: Active multimodal interaction system, ESCA Workshop on Spoken Dialogue Systems, pp. 169 - 172 (1995).
- 15) Itou, K., Hayamizu, S., Tanaka, K.: Recognition of transcriptions from speech samples, Technical Report of IEICE, SP96-104 (1997-01) (in Japanese).
- 16) Kurita, T., Otsu, N., Sato, T.: A face recognition method using higher order local autocorrelation and multivariate analysis, Proc. of 11th ICPR, Vol. 2, pp. 213-216 (1992).
- 17) Kurita, T., Hayamizu, S. : Gesture recognition using HLAC features of PARCOR images and HMM based recognizer, Proc. of International Conference on Automatic Face and Gesture Recognition, pp. 422-427 (1998).
- 18) Kurokawa, T.: Nonverbal interface, Ohm-sha (1994) (in Japanese).
- 19) Nagao, K., Takeuchi, A.: Speech dialogue with facial displays : Multi-modal human-computer conversation, Proc. ACL, pp. 102-109 (1994).
- 20) Oka, R., Kiyama, J., Kojima, H., Itoh, Y., Seki, S., and Nagaya S.: Real-time integration of speech, gesture, graphics, and database, 95 RWC Symposium (1995/06).
- 21) Suenaga, Y., Mase, K., Fukumoto, M., Watanabe, Y.: 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) (in Japanese).
- 22) Vo, M. T., Waibel, A.: Multimodal human-computer interaction, Proc. Int. Symp. on Spoken Dialogue, pp. 95-101 (1993).
- 23) Watanuki, K. , Sakamoto, K. , Togawa, F.: Analysis of multimodal interaction data in human communication, ICSLP-94, S17-8.2, pp. 899 - 902 (1994).

(Accepted May 19, 2000)