A MULTIMODAL-INPUT MULTIMEDIA-OUTPUT GUIDANCE SYSTEM: MMGS

Toshiyuki Takezawa and Tsuyoshi Morimoto†

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
E-mail: takezawa@itl.atr.co.jp morimoto@tlsun.tl.fukuoka-u.ac.jp
†currently with Department of Electronics Engineering and Computer Science, Fukuoka University, Fukuoka, Japan

ABSTRACT

We have built a multimodal-input multimedia-output guidance system called MMGS. The input of a user can be a combination of speech and hand-written gestures. The system, on the other hand, outputs a response that combines speech, three-dimensional graphics, and/or other information. This system can interact cooperatively with the user by resolving ellipses/anaphora and various ambiguities such as those caused by speech recognition errors. It is currently implemented on a SGI workstation and achieves nearly real-time processing.

1. INTRODUCTION

We have built a multimodal-input multimedia-output guidance system called MMGS. The input of a user can be a combination of speech and hand-written gestures. The system, on the other hand, outputs a response that combines speech, three-dimensional graphics, and/or other information. This system can interact cooperatively with the user by resolving ellipses/anaphora and various ambiguities such as those caused by speech recognition errors. It is currently implemented on a SGI workstation and achieves nearly real-time processing.

Recently, there have been many projects aimed toward intelligent multimodal systems [1], and great progress has been made in multimedia presentation systems [2]. However, much more effort is required to develop useful systems that exploit speech recognizers. Our research concentrates on speech recognition techniques for interactive systems.

From the viewpoint of accurate speech recognition in a multimodal environment, studies have already been made on the integration of speech recognition and speech reading [3, 4]. Our research focuses on the integration of speech recognition and language processing.

From the viewpoint of multimodal interaction systems using speech recognition, there have also been studies on dialogue systems with facial displays [5, 6]. We prefer three-dimensional graphic outputs because we assume that users would feel familiar with three-dimensional graphics. Such graphic information may lead to interesting situations.

Section 2 gives an overview of the system. Section 3 describes two key features. One is a technique to integrate speech recognition, hand-written gesture recognition and language processing. The other is for resolving ellipses/anaphora and ambiguities such as those caused by speech recognition errors. Section 4 gives preliminary system evaluation. In Section 5 we offer discussion and describe future works. Section 6 gives our conclusion.

2. SYSTEM OVERVIEW

Figure 1 shows a sample display of our system. Its current implementation guides visitors around facilities in the Hikaridai area of Kansai Science City and offers sight-seeing guidance around Nara City.

Figure 2 shows the system configuration. The system consists of a speech recognition sub-system, a hand-written gesture recognition sub-system, a semantic analysis sub-system, a problem solving sub-system, an object database, a graphics output sub-system, and a speech synthesis sub-system. The problem solving sub-system has four major functions: ellipses/anaphora resolution, object search, graphics operation and response generation.

3. KEY FEATURES

3.1. Integrated processing of speech and language

We have prepared a Japanese grammar for both speech recognition and semantic analysis. This grammar consists of CFG rules and annotations based on feature structures.

The speech recognition sub-system adapts an HMM-LR speech recognition method and uses

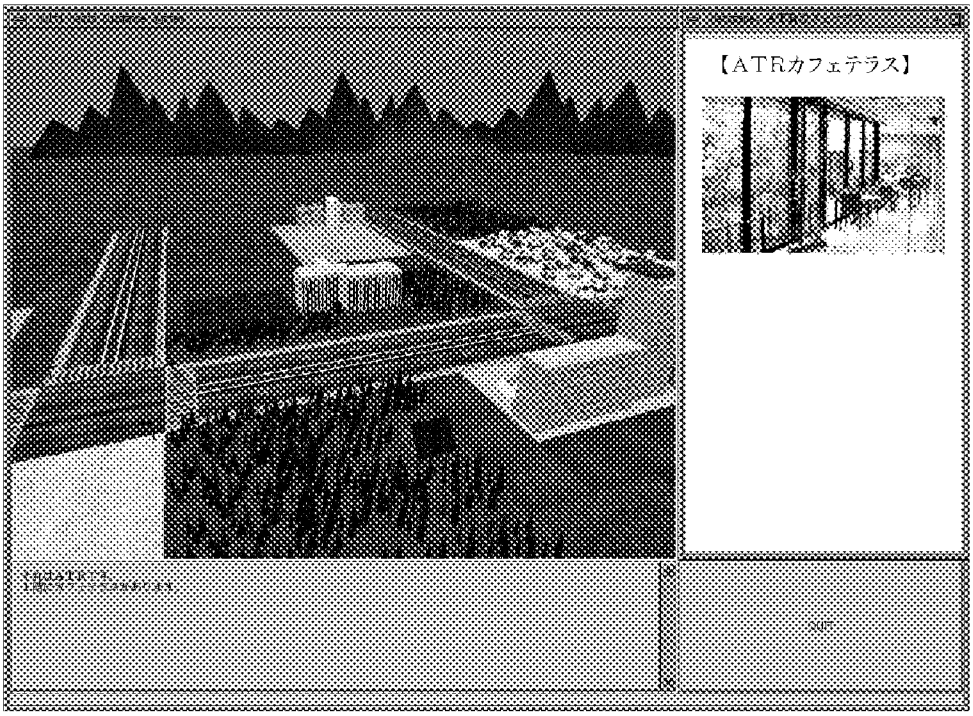

Figure 1. Sample display of system

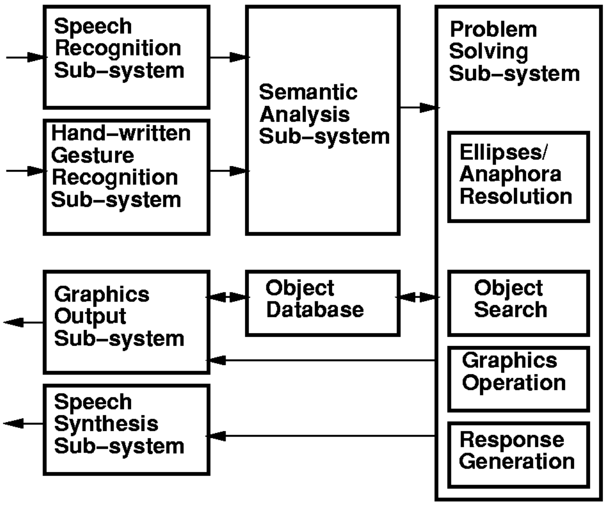
Speech Recognition Sub–system
Hand–written Gesture Recognition Sub–system
Semantic Analysis Sub–system
Problem Solving Sub–system
Ellipses/ Anaphora Resolution
Graphics Output Sub–system
Object Database
Object Search
Graphics Operation
Speech Synthesis Sub–system
Response Generation

Figure 2. System configuration

only CFG rules and the bigrams of preterminal symbols as a language model [7, 8]. Accordingly, this sub-system outputs sequences of CFG rules. The semantic analysis sub-system can generate semantic structures from the sequences of CFG rules and corresponding annotations based on feature structures by unification processing.

Users can point to facilities (objects) in three-dimensional graphic responses through hand-written gestures for questions like *"Kore wa nan desu ka?"* which means *"What is this?"* The hand-written gesture recognition sub-system supports pointing and circling gestures [9] and can recognize such gestures and corresponding facilities (objects). The semantic analysis sub-system obtains the information and is able to generate semantic structures. This semantic analysis sub-system is based on the semantic analysis module in our previous speech-to-speech translation system ASURA [10].

Figure 3 shows a part of our object database. It is necessary for such question and answer systems to provide information on objects. In our database, objects are defined in a hierarchical structure of classes and instances. White ovals in Fig. 3 indicate class definitions. These have properties of names, addresses, accesses and so

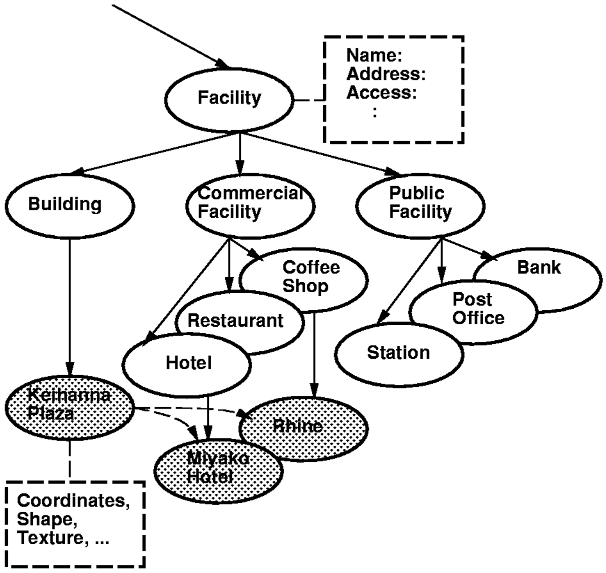

Figure 3. Part of object database

on. Gray ovals in the figure indicate instance definitions. These can be shown in the three-dimensional graphic display and have properties of coordinates, shapes, texture, and so on.

3.2. Resolving ellipses/anaphora and ambiguities

The problem solving sub-system refers to the object database and outputs responses with three-dimensional graphics, synthesized speech, texts, and/or pictures. The three-dimensional graphics and synthesized speech are synchronized as follows:

> **User input:** *"Kore wa nan desu ka?"* (What is this?)
> **System output:** *"Kore wa ATR desu."* (This is ATR.)
> [The ATR building is blinking in the display during this output.]

In addition, ellipses/anaphora resolution is achieved as follows:

> **User input:** *"ATR ni kissaten wa ari masu ka?"* (Is there a coffee shop in the ATR building?)
> **System output:** *"Hai, ATR niwa Cafeteria ga ari masu."* (Yes, there is a Cafeteria in the ATR building.)
> **User input:** *"Keihanna Plaza niwa nai n desu ka?"* (How about at Keihannna Plaza?)
> **System output:** *"Keihanna Plaza niwa Rhine to iu kissaten ga ari masu."* (There is a coffee shop Rhine in the Keihanna Plaza.)

The user's second input does not have any words to directly indicate a coffee shop. A problem-solving sub-system can resolve such ellipses by referring to a user's previous input.

Table 1. Results of preliminary system evaluation experiments

	Number	Average Response Time [s]
Total Utterances	447	3.7
Speech Detection Error due to Noise	44	4.1
Target Utterances	403	3.6
Succeed Utterances	273	3.6
Fail Utterances	130	3.7

We assume that this function can deal with not only the above examples but also with ambiguities such as those caused by speech recognition and/or hand-written gesture recognition errors.

> **User input:** *"Koko ni ..."* (This/Here ...) [The "..." means some kind of speech detection error.]
> **System output:** *"Sore wa Keihanna Plaza desu ga, dono you na go youken de shou ka?"* (This is the Keihanna Plaza building. How may I help you?)

4. PRELIMINARY SYSTEM EVALUATION

We have conducted a preliminary system evaluation with the task of sight-seeing guidance around Nara City. Sample conversations and instructions for microphone and mouse operations were given to the subjects before the preliminary evaluation experiments. Table 1 summarizes the results of these experiments. The number of subjects was eight. The accuracy rate was 67.7% (273/403 × 100), and average response time was 3.6 seconds. Speech detection error due to noise indicates that the system made some responses to some noise because the microphone was always active. The number of words in the dictionaries of speech recognition and language analysis modules is about 300.

A summary of users' impressions is as follows.

- Three-dimensional graphics are interesting.
- Not only three-dimensional graphics but also two-dimensional maps or name plates for landmarks may be necessary to understand the relationship between a user's current position and a landmark.
- Confirmation questions were frequently asked.

5. DISCUSSION AND FUTURE WORKS

Three-dimensional graphics seem to be impressive to users. However, users sometimes lose their current position and/or relationship between the current position and landmarks. We plan to introduce two-dimensional maps and/or name plates for landmarks in the three-dimensional graphics.

Confirmation questions might frequently occur in the current implementation. The last system output in the following example is a kind of confirmation question.

User input: *"Nara-ken Shin Koukaidou wa dore desu ka?"* (Which is the Nara-ken New Public Hall?)
System output: *"Nara-ken Shin Koukaidou wa kochira desu."* (This is Nara-ken New Public Hall.) [The Nara-ken New Public Hall is blinking in the display during this output.]
User input: *"Kintetsu Nara Eki kara aruite ike masu ka?"* (Can I go from Kintetsu Nara Station by walking?)
System output: *"Kintetsu Nara Eki kara Nara-ken Shin Koukaidou made desu ka?"* (You mean from Kintetsu Nara Station to Nara-ken New Public Hall, don't you?)

We are investigating several strategies of such kinds of confirmation questions. For example, the system would avoid the need to output a confirmation question when ellipses/anaphora resolution is completed by simply referring to the previous users utterance. We plan to implement several strategies in our system and examine users' level of comfort according to the situations.

In the current implementation, a Japanese grammar for both speech recognition and semantic analysis must be prepared in the first step. The speech recognition sub-system uses the CFG rules. This integration technique is quite convenient for designing and developing such kinds of systems. However, the number of words and the type of wording are limited. To achieve much more robust speech recognition, we plan to introduce a word-spotting mechanism as well as the current CFG-based approach.

6. CONCLUSIONS

This paper reported a multimodal-input multimedia-output guidance system called MMGS. The input of a user can be a combination of speech and hand-written gestures. The system, on the other hand, outputs a response that combines speech, three-dimensional graphics, and/or other information. This system can interact cooperatively with the user by resolving ellipses/anaphora and various ambiguities such as those caused by speech recognition errors. We plan to conduct further system evaluation after making modifications and revisions such as introducing word-spotting techniques.

ACKNOWLEDGMENTS

The authors wish to thank Mr. Goro Tanaka, Mr. Jun Shiozaki and Mr. Takeo Baba for their contributions to building our system.

REFERENCES

[1] *Proceedings of the Workshop on Intelligent Multimodal Systems*, IJCAI-97 (1997).

[2] Bordegoni, M., Faconti, G., Maybury, M. T., Rist, Th., Ruggieri, S., Trahanias, P. and Wilson, M.: "A Standard Reference Model for Intelligent Multimedia Presentation Systems," *Proceedings of the Workshop on Intelligent Multimodal Systems*, IJCAI-97, pp. 85–99 (1997).

[3] Vo, M. T. and Waibel, A.: "Multimodal Human-Computer Interaction," *Proceedings of International Symposium on Spoken Dialogue*, pp. 95–102 (1993).

[4] Stork, D. G. and Hennecke, M. E. (ed): *Speechreading by Humans and Machines*, NATO ASI Series, Springer (1996).

[5] Nagao, K. and Takeuchi, A.: "Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation," *Proc. of ACL-94*, pp. 102–109 (1994).

[6] Hasegawa, O., Itou, K., Kurita, T., Hayamizu, S., Tanaka, K., Yamamoto, K. and Otsu, N.: "Active Agent Oriented Multimodal Interface System," *Proc. of IJCAI-95*, pp. 82–87 (1995).

[7] Takezawa, T. and Morimoto, T.: "Dialogue Speech Recognition Using Syntactic Rules Based on Subtrees and Preterminal Bigrams," *Systems and Computers in Japan*, Vol. **28**, No. 5, pp. 22–32 (May 1997).

[8] Takezawa, T. and Morimoto, T.: "Conversational Speech Recognition Using Subtree-Based CFG Rules and Class-Based Bigram Models," *Proceedings of the International Conference on Speech Processing (ICSP '97)*, Seoul, Korea, pp. 355–360 (August 1997).

[9] Loken-Kim, Kyung-ho: "Prototyping of a Multimodal Spoken Language Translation System," *Technical Report of IEICE*, MVE96-49 (November 1996).

[10] Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M. and Kurematsu A.: "ATR's Speech Translation System: ASURA," *Proc. of EuroSpeech '93*, pp. 1291–1294 (1993).