Vol. 40 No. 4 Transactions of Information Processing Society of Japan Apr. 1999

Regular Paper

A Prototype of an Automatic Japanese-Korean Interpretation System

MASAMI SUZUKI,† NAOMI INOUE† and KAZUO HASHIMOTO†

In this article, we report the design principle and the outline of our experimental interpretation system from Japanese into Korean. Our objective is to provide a prototype of a semi-realtime spoken language translation system with a higher performance for a limited task domain, in order to actually investigate underlying factors towards realizing applicable interpretation systems. We also describe our original technique for speech recognition incorporating with dialogue state prediction, as well as human interface issues for achieving the above subgoal. A comprehensive evaluation result in a dialogue environment showed the current feasibility of the system and some significant remarks towards our future directions.

1. Introduction

This project was motivated with a concept of realizing a prototype of a semi-realtime spoken language interpreting system for a limited task domain (hotel reservation). The response time is very crucial for effective communication through interpreted dialogue between people in different languages. Of course, the performance of the system should be of high-level both in speech recognition accuracy and quality of translation. We also intended to implement such a system using minimum computing resources.

The next step we need is to investigate whether a limited capacity of automatic interpreting system could be applicable for certain actual task, or what kind of factors might be considered for interpreting communications. For this purpose, a high-speed, compact and real-working speech translation system should be realized as a test bed for evaluating usability of the system and improving the human interface. Based on the above consideration, our Japanese-Korean speech translation system has been developed in a joint work with Korea Telecom (KT) and Electronics and Telecommunications Research Institute (ETRI) in Korea[5].

In the following sections, our basic strategy of system design, the synopsis of the components and the experimental results are described. Afterwards, we discuss related issues and our future directions.

2. Task and System Design

2.1 Basic Concepts

Our current task domain is "Hotel Reservation" (hereafter, HR task). The dialogue participants are a client and a hotel clerk. The task is cooperative and basically asymmetrical between the two speakers. It also means that the hotel clerk is leading the dialogue session to achieve the goal (= room reservation). First we interviewed a professional hotel front operator about the typical procedure of room reservation, and then collected sample dialogues with role playing by human subjects. The 80 dialogues (4 trials for 20 different plots respectively) were transcripted and only essential patterns were extracted with omitting ungrammatical or unusual ones. Within these extracted expressions, we judged that we could write CFG-style recognition grammar, with following insights☆.

In order to incorporate dialogue constraints into the stage of speech recognition, grammar-based approaches are very convenient, compared with n-gram based approaches that would include only local constraints.

(1) In the grammar rules, it is possible to introduce discourse-level nonterminal symbols according to the types of utterance in a quite natural way.

(2) It will enable to drive speech recognition in accordance with dialogue states based on utterance types and their transition (see Sections 2.2 and 3.1).

The grammar consists of about 800 CFG-rules (static branching factor is about 17) and 1,000 words, that cover the minimum HR task with various expressions of requesting and confirming dates for stay, room types and numbers, 200 popular Japanese family names and typical

† KDD Research and Development Laboratories Inc.

☆ Moreover, the recognition result can be directly processed as translation input.

Korean names spelling, etc.

Furthermore, it covers grammatical patterns with changing words order and frequently used sentence-final auxiliary expressions.

2.2 Speech Recognition with Next Dialogue State Prediction

An important point is how to implement an effective speech recognition for spoken dialogue translation systems. Based on an existing HMM-LR based continuous recognition platform (specifications are described later), we integrated a light-weight discourse monitor which detects the current dialogue state and predicts the next states. Considering the features of cooperative goal-oriented dialogue, the next dialogue states are affected by previous states[1]. We adopted a rather simple prediction model to limit the number of applicable recognition rules, in order to achieve a higher performance and reducing the computation cost simultaneously. The mechanism is described later.

2.3 Dialogue Translation

Another important issue is to provide a robust and high speed dialogue translation module. For this purpose, we developed a variable word unit transfer program. As is well known, Japanese and Korean languages are similar in grammatical features. Kim, et al.[2] proposed a morphology based transfer scheme for text translation. We enhanced this to achieve more natural dialogue translation, with using flexible translation units. This also reduces the computation cost.

2.4 Systems Connection and Human-System Interaction

We assumed that another Korean-Japanese speech translation system for the reverse direction would be connected for bi-directional interpretation, in our co-operation with Korean research partners. The common pre-requisite is to provide an interface for speech input, confirming recognition result and displaying translation, etc.

The human interface of a dialogue system is very important as well as fundamental features. We have tried to lighten the burden on a user who actually speaks at the system in the above environment.

In our system, the user interface was designed in order that the series of actions was to be smoothly carried out from accepting the speech input to sending the processing result with a turn transfer signal. Our strategy was to provide a user with effective visual aids and to make his/her interaction with the system simple, for fluent communication with the other dialogue partner.

3. System Configuration and Implementation

The outline of the system configuration is shown in **Fig. 1**. The hardware resource is a single SPARC-20 workstation with 4 DSP chips for acoustic analysis as recognition preprocessing.

3.1 Speech Recognition Module with a Discourse Monitor

4,000 Phoneme-balanced sentences uttered by 80 speakers were used for learning our speaker-independent discrete type HMMs. The read corpus is general and completely independent from our task domain. The recognition method is based on one-pass N-best search algorithm with frame-synchronous HMM verification[3]. A CFG-based recognition grammar (about 800 rules and 1,000 word units) are precompiled into word predicting LR-tables.

3.1.1 Discourse-sensitive Speech Recognition

For the purpose of providing discourse constraints for speech recognition (also for dialogue translation), we implemented a light-weight discourse monitor which is easy to deal with. The monitor observes the dialogue status by detecting a combination of parameters: Speaker's Role (Hotel/Client), Utterance Type (15 sorts of general communicative act types in **Table 1**).

The information for the latter two parameters is known from the used recognition grammar

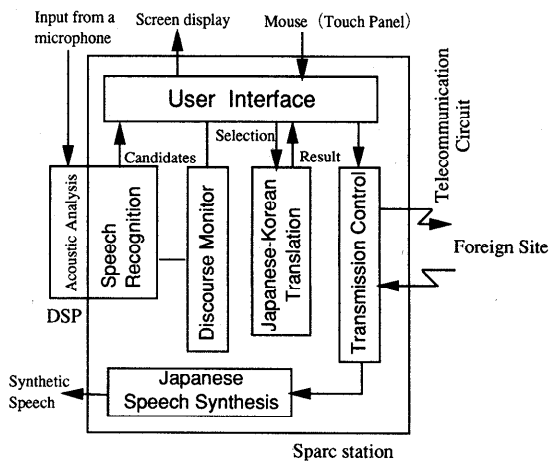


Fig. 1 System configuration.

Table 1 Utterance types.

Type label	(abbr.)	Meaning
Acknowledge	(AC)	acknowledgment
Confirmation	(CF)	confirmation
Demand Attribute	(DA)	request value attribute's value
Demand Behavior	(DB)	request certain action
Express Feeling	(EX)	express feeling
Greet Open	(GO)	greeting in beginning dialogue
Greet Close	(GC)	greeting in closing dialogue
Offer and Suggest	(OS)	offer or suggest
Question Polarity	(QP)	yes/no question
Question WH	(QW)	WH question
Response Positive	(RP)	positive response
Response Negative	(RN)	negative response
State Positive	(SP)	positive statement
State Negative	(SN)	negative statement
Wish	(WS)	Desire

rules☆.

This means that the CFG style grammar itself is organized to reflect a classification from the above viewpoint of dialogue parameters. In order to realize this with a simple technique, we labeled each rule with a header indicating its discourse related features (if it has such features). Also, a matrix table for transition of dialogue states was prepared. It provides knowledge about the possibility of the next dialogue status. This enables us to reduce the number of applicable rules for speech recognition in given dialogue states and to achieve a higher accuracy.

The detailed mechanism of the above mentioned dialogue state-sensitive recognition framework is as follows[4].

3.1.2 Dialogue States Transition Table

As mentioned before, each dialogue state is defined as a combination of Speaker's role and Utterance type. Then, possibilities of transition from a dialogue state to the next can be described as a matrix-like transition table as shown in the following example (**Table 2**). This table was made from labeled sample dialogue corpus within our task domain and supplementary consideration on possible next dialogue states. In this table, 1 stands for transition possibility and 0 for no transition. The reason why we adopted the 0/1 discrete value for each transition is as follows:

- Our labeled corpus was rather small to calculate meaningful stochastic values for dialogue states transition.
- Thus, we carefully checked the table so that any possible transition (though it rarely happens) would have a value "1", considering not a few typical example dialogues but various kinds of dialogue situation. Therefore, the table represents a rather general flows of cooperative dialogues☆☆.

Table 2 A part of a dialogue states transition table.

Dialogue state	H/GO	C/GO	H/QW	C/WS	···
H/GO	0	1	0	1	···
C/GO	1	0	1	0	···
H/QW	0	0	0	1	···
C/WS	1	0	1	0	···
···					

3.1.3 Labeled Recognition Grammar and Selective Rule Application

The recognition grammar, described in CFG format, includes top-level nodes corresponding respective utterance types, and each rule is annotated with a header indicating relevant dialogue parameters (e.g., Speaker's role, Utterance type) as far as they can be specified, as described in **Table 3**.

If ignoring these headers, a whole grammar with all rules is used. On the other hand, when referring the above transition table, dialogue state-sensitive word predicting LR tables can be complied for each dialogue state, from the extracted subset of the grammar. This effects selecting the recognition candidates into the predicted scope of a given dialogue state, with reducing computation costs. Additionally, in the case of a retry caused with mis-recognition occurs, the new utterance is treated in the same dialogue state as the previous one. Accordingly

☆ It can be also detected from the translated Japanese text with surface pattern matching.

☆☆ Those utterance types themselves in Table 1 are considerably general for the task domain.

Table 3 A part of an annotated CFG.

```
R1) -GO-<Sentence> = <Sentence_GO>
R2) HGO-<Sentence_GO> = <hotel_name> <v_copula>
R3) ----<v_copula> = <desu>
R4) H---<desu> = <でございます>      (degozaimasu)
R5) ----<desu> = <です>               (desu)
```

(Note: R1 ~ R5 for rule numbers, <Sentence_GO> is one of Sentence top-level nodes corresponding the utterance type *Greet Open*. The first of the 4 column-header shows the constraints with the speaker's role, *Hotel* or *Client*. The 2nd and 3rd columns indicates the relevance with utterance types. The 4th column is used for topic label assignment.)

Table 4 Speech recognition with/without dialogue monitor.

Discourse monitoring	OFF	ON
Sucess rate for top candidate	68%	86%
Sucess rate within 5th rank	75%	90%

this framework would be robust, as far as the utterance is within constraints.

A preliminary evaluation result (using prerecorded speech) showed that this method considerably improved the recognition rate compared with the non-prediction mode as shown in **Table 4**.

3.2 Japanese-Korean Translation Module

The basic method is a word level transfer using a Japanese-Korean translation dictionary, which is described as a regulated table containing a source (Japanese) expression, syntactic/semantic category, selective condition, target (Korean) expression and so on. A selective condition consists of adjacent words' syntactic/semantic and functional features.

The dictionary entries correspond to Japanese morphological units (flexible in case of collocations) and the target Korean expressions also contain morphological information. Thus, the final translated text is generated without any additional phase for morphological synthesis. This robust and powerful method can be easily extended to the reverse K-J translation. Furthermore, the processing units of translation are synchronous with those of speech recognition, and it guarantees easier maintenance of the linguistic resources of the system.

A subjective evaluation was performed on the translation result of 45 sample independent sentences as follows. Three measure were considered: intelligibility, grammaticality and naturalness, which were scored by 3 grades (1=perfect, 2=slightly defective, 3=poor), based on an idea originally suggested by Nagao, et al.[6]. The result of **Table 5** suggests that most translations are acceptable, but it might be difficult to pursue naturalness[5].

Table 5 Subjective evaluation rersult for Japanese-Korean translation.

Measure	grade 1	grade 2	grade 3	Total
Intelligibility	45	0	0	
Grammaticality	45	0	0	45
Naturalness	39	6	0	

Table 6 Inter-system connection protocols.

Packet form	{marker}*return* {Recognition result text}*return* {Translated result text}*return*
Marker	0 Transmission of results 1 Exchange of turn 2 Closing of dialogue
Recognition result	<Japanese strings> S-JIS
Translated result	<Hangul strings> KSC

A few examples of input Japanese and output Korean sentences are shown as follows.

J-input: 部屋を予約したいのですが。
(I would like to reserve a room.)
K-output: 방을 예약하고 싶은데요.

J-input: いつお泊まりのご予定でしょうか。
(When are you going to stay?)
K-output: 언제 숙박하실 예정입니까?

3.3 Inter-system Connection and User Interface

The peripheral conditions of our system were presupposed as follows, cooperating with our research partners in Korea.

- Speech Input from a microphone (headset)
- Displaying plural recognition candidates on the monitor screen
- Selecting an appropriate candidate with mouse clicking
- Invoking translation with the selection
- Transmitting recognition and translation result together with turn transferring information (see **Table 6**)

Though a user was requested a few more mouse clicks for starting utterance and sending a turn signal in the above environment, we reduced the click times into only once (selecting a candidate) in a turn, giving an input-ready

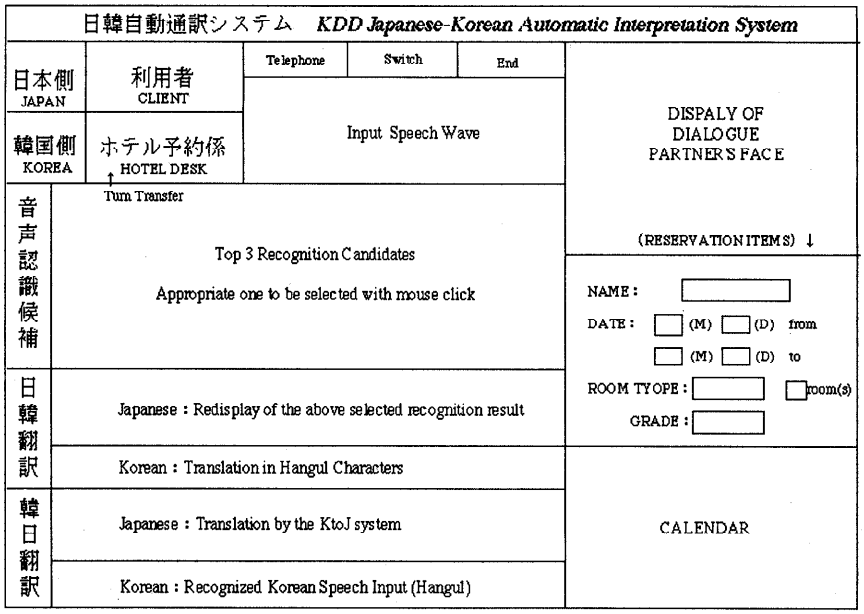


Fig. 2 System monitor screen.

Table 7 Performance in the joint experiments.

Japanese speakers' role	First utterance's success rate	Mean duration of one turn
Client	98%	6.0 sec
Hotel	95%	8.1 sec

sign and sending a turn automatically. This treatment was considered with human factors of dialogue systems[7].

Additionally, we have introduced a display of the reservation items (dates, room type/grade and name) and a calendar which is switched according to the mentioned (recognized) date, as an attempt of task-dependent visual interface. Using the above reservation items, the confirmation procedure can be smoothly performed. Furthermore, from the viewpoint of multimodal dialogue interface, we equipped a video display sub-window on the monitor screen to see the partner's face. See **Fig. 2**.

Previous to the system evaluation described in the next section, we actually connected our system with our partner's systems in Korea[8),9)]. For the reference, we note here the success rate and the time duration in a turn☆ at our site in the (prearranged) demo-dialogues of this inter-system connection, as shown in **Table 7**. It is regarded as the system's performance when a few expert users speak as their own roles.

4. System Evaluation

4.1 Comprehensive System Evaluation Experiment

For the purpose of evaluating our system's performance in an interpreted dialogue environment, we implemented a simple quasi Korean-Japanese speech translation system which emulates the reverse direction processing to our J-K system. Using those two (real and simulated) systems, we evaluated the comprehensive performance of our Japanese-Korean speech translation. In other words, ths following experiments will show the usability and defects of the system in a semi-real situation.

The simulated K-J system was operated by a human experimenter so that he could select a pair of Korean and Japanese sentences. Under such an arrangement, a quasi input sentence was synthesized by a Korean text-to-speech software☆☆, and a quasi output (translated Japanese) was transmitted with its input to the real J-K system and was also synthesized to be heard by the subject. This setting is almost similar to the actual interpreted communication experiments performed between Korea and Japan.

16 novice subjects were asked to play a role

☆ It means the duration from the input-ready indication till the sending the translated text to the other site.

☆☆ This was developed by one of our research partners in Korea.

Table 8 Comprehensive evaluation result (average of 16 subjects).

Success rate measure	1C	2A	2B
First utterance's sucess rate[1]	95.1%	94.2%	96.5%
Actual/minimum number of utterances[2]	1.05	1.13	1.07
Top candidate ratio in successful trials[3]	88.7%	86.3%	87.2%
Mean duration of one turn	6.3 sec	6.3 sec	6.0 sec

Note:
1) First utternace's success rate shows how a user's first speech action for a given sutuation was acceptable to proceed the dialogue, with selecting an appropriate recognition candidate from three.
2) Actual/minimum number of utterances suggests how much a user had to actually speak, compared with when he/she could perform without fail.
3) Top candidate ratio indicates the percentage of recognition candidates scored as the top within successful speech action.

of clients who are going to make a reservation. The procedure of the experiment was as follows.

(1) Explanation about the system was given by an edited video tape containing the above Korea-Japan (actual) bidirectional connection.

(2) Instructions through a demonstration (one reservation dialogue). These two kinds of explanation was completed within 10 minutes.

(3) Experiment 1C. This is a role play by the subject, reading a prepared whole scenario, as a client role speaker (9 sentences for reserving a room).

(4) Experiment 2A and 2B. In this case, a subject was requested to make an own reservation memo (containing dates, room type and grade, ...) in advance (2A), and then performed a role play as a client without looking any other text. Next, he/she was given another memo including a different setting (except name) to perform one more dialogue (2B).

Currently, we are estimating the system's performance by the success rate of a user's action as one of speaker roles and the time duration spent for a dialogue. **Table 8** shows several success rate measures for overall experimental dialogues.

4.2 Experimental Result and Remarks

Table 8 suggests that the subjects performed well at the automatic interpretation system, though they were at the first time to speak in such a novice environment, yet the content of the performed dialogues was basic and simple (no negotiation between clients and the hotel)☆.

Next, we will examine the result and investigate additional features observed in the experiment.

Regarding the observed recognition errors in the experiment is as follows.

- 93% of the mis-recognized utterances were acceptable with the grammar.
- The above misrecognition was limited in the utterance type SP, where either dates, room types or person names were mentioned.
- The utterances that were out of grammar contained undefined sentence -final expressions.
- However, most of illegal expressions including hesitation were accepted as suitable recognition candidates.

Concerning the response time, we observed an interesting result as follows. Roughly speaking, the mean duration in a turn (about 6 sec) consists of 2 sec speech input, 2 sec processing delay and 2 sec user's response time. The deviation of a turn duration by person was greater in 2A and 2B (0.50) than in 1C (0.22). This shows that the subjects' behavior interacting with the system varied person by person in a more spontaneous situation compared with a controlled one. Additionally, a slight learning effect was observed in the reduced response time between 2A and 2B. Actually, 12 of 16 subjects shortened their response time in 2B.

Such an interpreted dialogue performed in our current systems connection takes a certain long time duration, even if the processing delay would be reduced with realtime techniques. The major factors are the duration of synthetic speech for the translation and the user's response time. In our framework of system design, it would be possible to cut the speech synthesis, with giving only visual text information on the monitor screen. On the other hand, it might be difficult to reduce the response time, considering the current accuracy of speech recognition. However, some categories of short utterances (corresponding the certain utterance types, e.g., AC, EX, GO, GC, RP in Table 1) seem to have been recognized 100% in our experiment. It will offer a key of skipping confirmation of speech input under

☆ The utterances in the conditions 2A and 2B may be regarded as semi-spontaneous, where the dialogue initiative was taken by the hotel side.

certain conditions*.

5. Discussion

So far, several major research activities are known in this field of spoken language translation. ATR, CMU and Siemens, et al. demonstrated their speech translation systems connected through international circuits[10),11)]. Generally, most of those papers report implementation issue and describe occasionally the evaluation of system components or overall performance of static speech translation processing. Though such evaluation is also needed, it is now more important to investigate how the system works in an interaction with a user. This viewpoint is that we intended to examine the efficiency of interpreted dialogues through experiments on a semi-realtime interpretation system.

Fortunately, we can refer some important papers on interactive spoken dialogue systems (e.g., Refs. 12), 13)), where advanced design of multimodal human interfaces between systems and users are proposed and their evaluation result are reported. On the other hand, the most significant features of interpretation systems is that they have to intermediate two users one at each end. Although there seem to be few investigations about human factors in speech translation systems, we have to incorporate these considerations by designing effective interfaces for automatic interpretation.

We also have to improve the basic performance level of our speech recognition and translation modules. Nevertheless, we believe that various discourse knowledge (like the goal of a task or range of values, etc.) should be integrated with these components from the viewpoints of providing intelligent interpretation. In our current task domain, we need little discourse knowledge for resolving contextual ambiguities. However, we know that certain task-dependent knowledge is necessary for achieving more natural interpretation[14)], as well as improving next utterance's prediction. We suppose that a multimodal dialogue interface would be effective, when its design is closely related with discourse and embedded domain knowledge. This issue should be further investigated for feasible speech translation systems.

* It is suggestive that the recognition rate of sentences of utterance types SP and SN was relatively lower, where various kinds of utterance are included. In other words, such types of sentences can contain a wide range of content words in the HR task domain, and are difficult to recognize correctly.

6. Conclusion

We reported a prototype of an automatic Japanese-Korean interpretation system with semi-realtime processing, using minimum computing resources. We recognized the effects of a light-weight discourse monitor with next utterance type prediction for the purpose of improving speech recognition accuracy and reducing computation cost. Our translation module based on flexible transfer units also showed its robustness.

Furthermore, we showed the current feasibility of our automatic J-K interpretation system through an evaluation experiment. Though, the result depends on our arrangement of experimental conditions, it suggested that a well designed system would yield a relatively high performances by novice users, even when comparing them with those of expert users.

Our system can be used for further experimental study on important factors of interpreted telecommunication, and such investigation should be continued and enhanced in order to achieve an advanced system design and user interface, towards realizing a practical interpretation system.

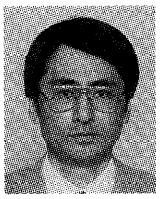
Acknowledgments We would like to thank Dr. Seiichi Yamamoto and Dr. Fumihiro Yato for motivating and executing our research project. We are also grateful to Dr. Takuro Muratani, Dr. Hitomi Murakami and Dr. Kenji Suzuki for their advices and encouragement. Finally we acknowledge all the efforts done by many people who contributed from corpus analysis, software implementation to system evaluation.

References

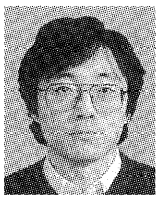
1) Nagata, M. and Morimoto, T.: An Experimental Statistical Dialogue Model to Predict the Speech Act Type of the Next Utterance, *Proc. ISSD-93*, pp.83–86 (1993).
2) Kim, T.S., et al.: A Japanese-Korean Machine Translation Based on Conjugated Word Analysis, *ICEIC '91*, P.R.China (1991).
3) Kuroiwa, S., Takeda, K., et al.: A Voice-Activated Extension Telephone Exchange System, *Proc. EUROSPEECH '93*, Vol.3, pp.1793–1796 (1993).
4) Suzuki, M., Inoue, N. and Yato, F.: An Efficient Dialogue Speech Recognition Method

based on Utterance Type Prediction, *ASJ Fall meeting*, pp.87–88 (1995).
5) Suzuki, M., Inoue, N., Yato, F., Takeda, K. and Yamamoto, S.: A Prototype of a Japanese-Korean Realtime Speech Translation System, *Proc. EUROSPEECH '95*, Vol.3, pp.1951–1954 (1995).
6) Nagao, M. and Tsujii, J.: Evaluation of Japanese-English Translation Result in Mu-Project, IPSJ SIG Notes on NL, Vo.85, No.47-11 (1985).
7) Suzuki, M., Inoue, N. and Yato, F.: A Design of User Interface for an Automatic Interpretation System considering the Task Environment, Technical Report of IEICE NLC95-29, pp.75–80 (1995).
8) Koo, M.-W., et al.: KT-TS: A Speech Translation System for Hotel Reservation and Continuous Speech Recognition System for Speech Translation, *Proc. EUROSPEECH '95*, Vol.2, pp.1227–1230 (1995).
9) Lee, Y.-G., et al.: Korean-Japanese Speech Translation System for Hotel Reservation – Korean front desk side, *Proc. EUROSPEECH '95*, Vol.2, pp.1197–1200 (1995).
10) Morimoto, T., et al.: ATR's Speech Translation System: ASURA, *Proc. EUROSPEECH '93*, Vol.2, pp.1291–1294 (1993).
11) Waibel, A.: Interactive Translation of Conversational Speech, *Proc. ATR International Workshop on Speech Translation* (1996).
12) Takebayashi, Y., et al.: Spontaneous Speech Dialogue System TOSBURG II – Towards the User-Centered Multimodal Interface, *IEICE*, Vol.J77-D II, No.8, pp.1417–1428 (1994).
13) Zue, V.: Research and Development of Multilingual GALAXY: A Status Report, *Proc. ATR International Workshop on Speech Translation* (1996).
14) Suzuki, M.: A Method of Utilizing Domain and Language-specific Constraints in Dialogue Translation, *Proc. COLING-92*, Vol.2, pp.756–762 (1992).

(Received September 30, 1998)
(Accepted February 8, 1999)



Masami Suzuki was born in 1955. He received his Master degree from Keio Univ. in 1980. He has been working in KDD since 1980 and now is a senior research engineer of KDD R&D Laboratories Inc. Since 1989 through 1993 his research activity was in ATR Interpreting Telephony Labs. His current research interests are cross-language information access and communication support. He is a member of IPSJ, IEICE, JSAI, NLP and ACL.



Naomi Inoue was born in 1959. He received his M.S. and Ph.D. degrees from Univ. of Kyoto in 1984, 1998 respectively. He joined KDD in 1984 and now is a senior research engineer of KDD R&D Laboratories. Since 1987 until 1991 he researched in ATR. He has been engaging in the research areas of spoken dialogue processing, especially on speech recognition systems for telephony applications. He is a member of IPSJ, IEICE, JSAI and ASJ.

Kazuo Hashimoto was born in 1953. He received his Master degree from Tohoku Univ. in 1979. Since then, he has been working for KDD R&D Laboratories, and currently he is a senior manager of Knowledge-based Information Processing Laboratory. His basic background is artificial intelligence and his research scope is from realtime expert systems to information retrieval, data mining, etc. He is a member of IPSJ, IEICE and JSAI.