PAPER Special Issue on Speech and Discourse Processing in Dialogue Systems

System Design, Data Collection and Evaluation of a Speech Dialogue System

Katunobu ITOU†, *Nonmember*, Satoru HAYAMIZU††, Kazuyo TANAKA†† *and* Hozumi TANAKA†, *Members*

SUMMARY This paper describes design issues of a speech dialogue system, the evaluation of the system, and the data collection of spontaneous speech in a transportation guidance domain. As it is difficult to collect spontaneous speech and to use a real system for the collection and evaluation, the phenomena related with dialogues have not been quantitatively clarified yet. The authors constructed a speech dialogue system which operates in almost real time, with acceptable recognition accuracy and flexible dialogue control. The system was used for spontaneous speech collection in a transportation guidance domain. The system performance evaluated in the domain is the understanding rate of 84.2% for the utterances within the predefined grammar and the lexicon. Also some statistics of the spontaneous speech collected are given.

key words: speech dialogue system, spontaneous speech, continuous speech recognition, speech understanding

1. Introduction

This paper describes design issues of a speech dialogue system, the evaluation of the system, and the data collection of spontaneous speech in a transportation guidance domain.

Natural conversation is the ultimate interface between humans and computers. Humans use speech as an interactive interface with each other. In this sense, "dialogue" is essentially important for a speech interface.

Several phenomena are characteristic of dialogues. Firstly, spoken language is different in many ways from written language. It has specific phenomena such as interjections, restarts, filled pauses, and hesitations. Speech which includes these phenomena is called "spontaneous speech". Secondly, humans change their speaking styles and attitudes according to the other person's behavior. It is not clear how humans change them when they talk to computers. Thirdly, the information of an utterance in a dialogue depends on the situation in which it is produced.

As it is still difficult to collect spontaneous speech and more difficult to use a real system for collection and evaluation, these phenomena have not been quantitatively clarified yet. The construction of an actual operating speech dialogue system, which operates in real time, with acceptable recognition accuracy and flexible dialogue control, is indispensable in order to study these phenomena.

Several systems have been constracted with the aim of evaluating speech dialogues between humans and computers.[(1)-(6)] However, many of the problems of realizing speech dialogue systems are still unsolved. Most systems have been neither implemented nor evaluated in real-time operation. An experimental system is necessary to clarify the system design, data collection and evaluation of speech dialogue systems.

System design will have a great influence on the speech corpora to be collected and the results of evaluation should be fed back to the system design. In addition, the evaluation depends on the system design. Therefore, collection of spontaneous speech corpora and construction and evaluation of the system should be performed simultaneously.

The "Wizard of OZ" paradigm has often been adopted for the collection of speech dialogue corpora which include spontaneous speech.[(5)] Only recently, some efforts have started in the collection of spontaneous speech using real systems.[(6)-(8)]

The authors have constructed a speech dialogue system.[(7)-(8)] The system consists of a speaker-independent continuous speech recognizer, niNja,[(9)-(10)] and a dialogue manager. It runs in almost real time on conventional work-stations without any special hardware. This paper is focused on system design, data collection and evaluation of the speech dialogue system constructed. Experimental results in a transportation guidance domain are shown.

2. System Design

2. 1 Requirements for Speech Dialogue Systems

In order to realize natural conversations between humans and computers, there are three requirements for speech dialogue systems.

(1) real-time response

It is important not to interrupt the user's train of

Manuscript received June 22, 1992.
Manuscript revised August 27, 1992.

† The authors are with the Faculty of Engineering, Tokyo Institute of Technology, Tokyo, 152 Japan.
†† The authors are with Electrotechnical Laboratory, Tsukuba-shi, 305 Japan.

thought. In the least, some sort of response is needed within a few seconds after the user has stopped speaking.
(2) acceptable recognition accuracy
The user's utterance must be correctly recognized to some degree, even if complete speech recognition is almost impossible. A framework is thus needed so as not to destroy the flow of dialogues even when there are recognition errors.
(3) flexible dialogue control
The response must be appropriate to the user's remarks. In some dialogues, it is not necessary to completely understand and respond to everything the user says. When the utterance includes ambiguities and errors, it is possible to resolve these through several interactions. In some applications, if some ambiguities and errors cannot be resolved, it is enough to just indicate this and it is not necessary to answer the utterance itself.

2.2 Specifications of Dialogue

In view of current continuous speech recognition technology and computer performance, only limited dialogues can be handled in accordance with the above three requirements.

Also, when a speech dialogue system is planned for a practical application, specifications of dialogue become important. It is necessary to determine the following specifications for designing a dialogue system.
(1) Goal of dialogue—What is the goal of the dialogue that the user can accomplish ?
(2) Form of dialogue—The way of utterance, initiative, turntaking, media of system response (text, table, figure, speech output etc.).
(3) Answerable questions—What can be answered by the system ?
(4) Assumed situation—What can be assumed about the user's prior knowledge of the domain ? What information should be provided to the user ?

Changes of these specifications will influence the user's utterances. A clear definition of them is important for the evaluation of the system.

2.3 Usage of Real System for Data Collection

The "Wizard of OZ" paradigm has often been adopted for the collection of speech dialogue corpora which include spontaneous speech.[5] Typically, there is a person behind a curtain who listens to the user's utterances and types in its content with a keyboard.

The authors used a real speech dialogue system for the collection for the following reasons.
(1) Rapid system response
Real-time or near real-time response is the first requirement as described in Sect. 2.1. Input with a keyboard usually takes more than several seconds and will result in slow responses of the system. Usage of a real-time system will make the response time shorter.
(2) Improvement of the dialogue system step by step
There are some limits in the dialogue system in the fulfillment of the real-time response. These limits can gradually be loosened and a larger domain can be implemented according to the research progress.
(3) Influence by the user's system image can be studied
The user's utterances are influenced by the system image which is conveyed through unexpected response, recognition errors, etc. To study relations between the system image and the utterances, flexibility is necessary.

3. Implementation of a Dialogue System

The authors have constructed an experimental speech dialogue system which is intended to satisfy the three requirements in Sect. 2.1. Figure 1 shows the overall structure of the system constructed. The system consists of a speaker-independent continuous speech recognizer, niNja, and a dialogue manager.

The flow of the dialogue is modeled by the state transitions. Dialogue control includes resolving ambiguities, ellipses, rejections and escape from the flow. It runs in almost real time on conventional work-stations without any special hardware. Collected speech is stored on optical disks.

3.1 Speech Recognizer

The speech recognizer is an HMM-based system with an N-best search algorithm incorporated with the LR-parsing technique. In the search algorithm, scores of the same phone models in different hypotheses at the phone-level are represented by one score of the best hypothesis and the differences. At each of the phone, word, and grammar levels, the history of hypotheses is stored as trees so as to prevent re-computations. Hypotheses of the histories in each level are linked as in Fig. 2. Word level and grammar level are separately implemented by two LR-tables for grammar and lexical access in order to deal with word-level phenomena

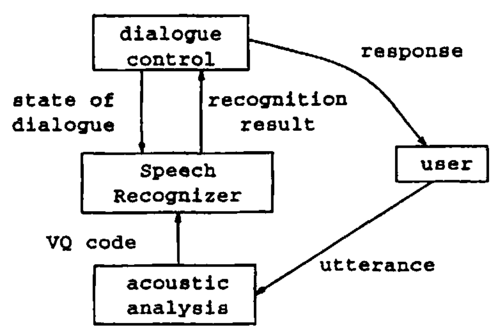

Fig. 1 Overall structure of the dialogue system.

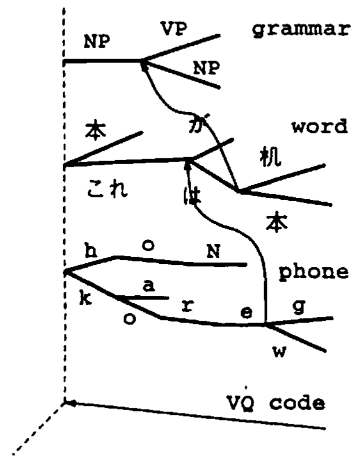

Fig. 2 Links between several levels in the search space.

Table 1 Speech corpora for training phone models.

Male speakers

1) WD-II : 1542 words * 10 speakers

2) ATR Phonetically balanced sentences : 150 sentences * 2 speakers

3) interrogative sentences : 11 sentences * 13 speakers

4) task specific sentences : 64 sentences * 3 speakers

Female speakers

5) WD-II : 1542 words * 10 speakers

6) ATR Phonetically balanced sentences : 150 sentences * 2 speakers

such as unknown word processing.

The speech recognizer processes input speech time-synchronously up to and including the grammatical parsing level. A kind of pipeline processing is performed so that acoustic analysis and the speech recognition process can start from the very beginning of the user's utterance.

The phone models are 43 context-independent discrete HMMs. The frame shift is 10 msec and the sampling frequency is 15 kHz. Two sets of phone models are made, one for males and the other for females or children. The speech corpora used for training models are shown in Table 1. Other details of the recognizer can be found in Refs.(10)-(12).

3. 2 Grammar and Lexicon

The grammar and lexicon are usually limited in continuous speech recognition systems, where recognition accuracy and efficiency are kept at acceptable levels. Consequently, the definition of the grammar and lexicon become an important issue in constructing actual operating systems. However, to the authors' knowledge, reports of the theory of or discussions about their definitions are very few.

Table 2 Example of templates for the grammar.

<station> <kara> <station> <made> <dorekurai> <kakaru>

<station> = Tokyo | Arakawaoki | Ueno | O-okayama | ...

<kakaru> = kakarudesho- | kakarimasuka | ...

The authors have defined them using the following process.

(1) Construct a prototype dialogue system with a keyboard input according to the specifications of the dialogue. Then record the utterances between the system and testees (A human operator listens to the user's utterances and enters them with a keyboard in a form which the prototype system can process correctly).

(2) The recorded speech is transcribed. Grammar and lexicon are defined according to the transcriptions in such a way that as many utterances as possible are covered by them. At this step, the grammar has 26 sentence templates and the lexicon has 70 words.

(3) This initial version is used for constructing the speech dialogue system. Then while the user tests the system with the speech input, the utterances are recorded again.

(4) These utterances are transcribed, and used to revise the grammar and lexicon. At this step, the grammar has 33 sentence templates and the lexicon has 85 words.

The grammar is defined as templates shown in Table 2. These templates are actually implemented in the form of context-free grammar. Words in the ⟨⟩ are nonterminal symbols. The flow of the dialogue is modeled in terms of state transitions of four states. Grammar and lexicon are defined for each state separately in order to reduce the complexity.

Interjections and mispronunciations are contained in the recorded data. However, these are excluded in defining the grammar and lexicon.

Comparing the first grammar and the second, most of the new templates differ from old templates with respect to the word order. The increase in the number of words is caused by the addition of destinations that are dealt with in the system.

3. 3 Semantic Processing

Speech is processed sentence by sentence. The speech recognizer regards each successive utterance as one sentence. It sends a simple semantic structure to the dialogue manager. An example of semantic structure is as follows.

Recognition result:

to-kyo-made dorekurai jikaNwa kakari masuka

Table 3 Items in the database.

Destination	Route
Tokyo	Arakawaoki – Ueno – Tokyo
O-okayama	Arakawaoki – Ueno – O-imachi – O-okayama
Ueno	Arakawaoki – Ueno
	⋮

Section	Line	Fare	Time
Arakawaoki – Ueno	Joban	1000	60
Ueno – Tokyo	Yamanote	200	10
Ueno – O-imachi	Keihintouhoku	300	30
	⋮		

(How long does it take to Tokyo?)
Semantic structure:
[name_of_station(Tokyo), how([fare, time]), theme(time)].

The semantic structure sent from the speech recognizer is processed by the dialogue manager. The dialogue manager queries the database. Table 3 shows the items in the database.

Semantic processing in the dialogue manager proceeds as follows.
(1) Resolving ambiguities in the semantic structure if possible.
(2) Querying the database.
(3) Response generation using a template with the information from the database.

There are cases when the response can be generated without difficulty from a single utterance. On the other hand, there are also cases when the ambiguity of the utterance cannot be resolved even by using the dialogue history and background knowledge. In the latter case, the system generates a question to obtain the required information for resolving the ambiguity.

3.4 Dialogue Control

In general, the dialogue manager controls the dialogue flow, understands the user's purpose and intention, and responds accordingly. This system maintains the dialogue histories in the form of pairs of the input semantic structure and the state. They are used to reject some utterances or recognition errors which are inconsistent with the expected dialogue flow.

The histories are also used to estimate incomplete information in the ellipses of some cases. Control of the dialogue flow includes rejections, escapes, and termination of the session.

3.5 Scenario and Sample Dialogue

The authors determined the specifications of dialogues as follows.

Table 4 Guidance sheet for the user.

```
This task is a simple transportation guidance from Tsukuba
to Tokyo. The system can provide information about seven
stations (destinations), Shinjuku, Ginza, Kasumigaseki,
Tokyo, O-okayama, Ueno and O-imachi. It can answer the
name of the railway, transfer stations, the time and the fare
to one station for each session.

Please respond appropriately to the output message. First,
the system greets.

System: こんにちは。 (Hello)
        どこから来ましたか。(Where did you come from ?)
[your utterance -- example: つくばから来ました
                            (I came from Tsukuba).]
System: どこに行きますか? (Where are you going ?)
[your utterance]
......
```

Table 5 A sample dialogue.

```
System: どこに行きますか? (Where are you going ?)
User  : shiNjukuni ikitaiNdesukedo
        [I would like to go to Shinjuku.]
System: 新宿ですか。新宿に行くには、常磐線と山手線に乗ります。
        この経路について、のりかえ、所要時間、費用をお尋ね下さい。
        (Shinjuku. You should take Joban-line and Yamanote-line
        to go to Shinjuku. Please ask about transfer stations,
        the time or the fare for this route.)
User  : dokode norikaeruNdesuka [Where do I transfer ?]
System: 上野です。(At Ueno.)
        他に何かございますか。
        (What else would you like to ask ?)
User  : donokurai kakarimasuka
        [How much does it cost ? / How long does it take ?
        (this utterance is ambiguous)]
System: 値段ですか、それとも時間をお尋ねですか。
        (Are you asking about the time or the fare ?)
User  : jikaN desu [time.]
System: 1時間20分です。他に何かございますか。
        (It takes one hour and twenty minutes. What else would
        you like to ask ?)
User  : uenokara shiNjukumadewa naNpuN kakarimasuka
        [How long does it take to go from Ueno to Shinjuku ?]
System: 上野から新宿までは20分です。
        (It takes twenty minutes from Ueno to Shinjuku.)
......
```

(1) Goal of dialogue
-The goal of the dialogue is to obtain information about railway transportation from Tsukuba to Tokyo.
(2) Form of dialogue
-User utterances and system responses are made in turn and do not overlap. If a user utterance occurs during the recognition process, it will be treated as if the utterance has occurred in turn.

-User's utterances are limited to one sentence for each utterance.
-The system first shows the recognition result and then displays its response in Kanji-kana (Japanese) text.
-The system responds after the end of the user's utterance.
-The system assumes only one topic in one session.
-The system takes the initiative.
-The user is assumed to speak cooperatively.
-The system does not show the state of the dialogue flow to the user.

(3) Answerable questions

The items covered in the defined task are as follows.
-Names of the railways to the destination.
-Time needed to arrive at the destination.
-Fare to the destination.
-Transfer stations on the way.

(4) Assumed situation

-User is at Arakawaoki station of the JR-line.
-User does not know the way to the destination.

A guidance sheet (Table 4) is provided to the user for the information about the system. Sample sentences, grammar and lexicon are not shown to the user. This is because the collection of spontaneous speech and the study of the manner of speaking and vocabulary are an important issue for the authors.

Table 5 shows a sample dialogue of the system.

4. Evaluation

The authors collected data on dialogues between humans and computers with the above-mentioned system and evaluated them. The subjects consisted of a variety of people, 33 males and 5 females, for a total of 38 speakers between the ages of twenty and sixty. Among them, only three males were included as speakers of training speech data and the other 35 speakers were not included. Some children, of about age ten, were also tested, but their data were not recorded. Most people had no prior knowledge of the system. Some were computer scientists; some were not reasearchers.
The system was tested in the laboratory office. Some utterances included laughter, clearing of throats and overlapping with others' voice.

The collected dialogue data consisted of 44 sessions. They include three sessions where the subject accidentally stopped talking with the system. Sessions consisted of 312 utterances in total. They included eight utterances which were out-of-grammar for the correct state of the dialogue flow and thus were invalid for the system.

The number of utterances per session is shown in Fig. 3. The mean number of utterances per session is 7.25. The number of bunsetu per utterance is shown in Fig. 4. The mean number of bunsetu per session is 2.08.

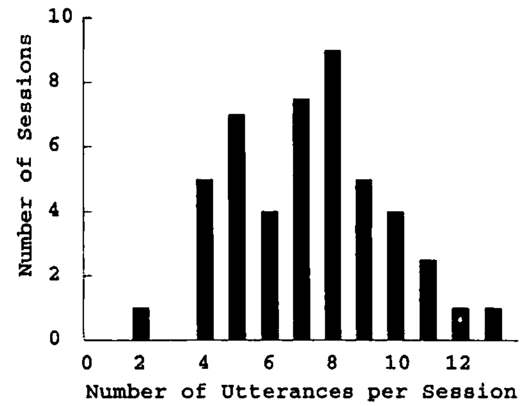

Fig. 3 Number of utterances per session.

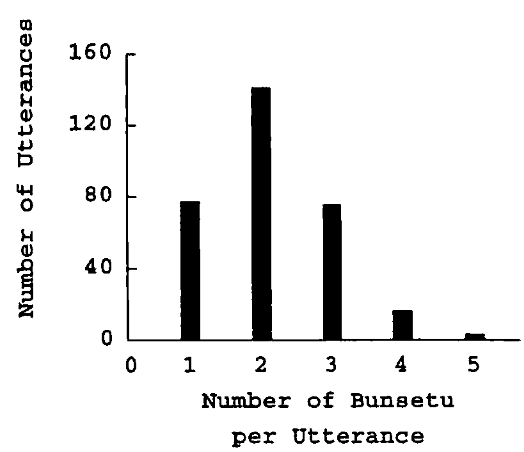

Fig. 4 Number of bunsetu per utterance.

The aim of collecting dialogue data is to study the phenomena that occur in spontaneous speech. The following four types of grammatical ill-formedness were studied: unknown words, out-of-grammar utterance, interjections, and restart and repeat phrases.

The data have 37 utterances that include some unknown word. This is 11.9% of the total utterances. The total number of unknown words is 39, which consists of 19 different words. Three cases were variations of pronunciation of known words. For example, /yoi/ is a variation of /i-/ (good). Two cases were nominal synonyms of known words. For example, /okane/ is synonymous with /ryo-kiN/ (fare). Three cases were verbally synonymous phrases of known phrases and another three cases were modifying synonyms of known words. There were five new words that comprised the phrases indicating that the user has no more questions. There was also one word that was an out-of-domain utterance; this word is /kakuniN/ (confirm). The system could not handle the utterance of confirmation.

There were three types of out-of-grammar utterance: changes in the order of the phrases from well-formed sentences, unknown types of phrase endings, and unknown combination of phrases. The total

Table 6 Occurrence of interjections.

Interjection	Number of Occurrences
/e-/	4
/e-to/	4
/eto/	3
/a/	2
/a-/	1
/e/	1

number of out-of-grammar utterances was 77. This is 25.3% of the total utterances.

There were fifteen utterances that had an interjection. This is 4.8% of the total utterances. The six types of interjection, /e-/, /e/, /eto/, /a-/, /a/, /e-to/, were found. The numbers of occurrences of these interjections are shown in Table 6. All these interjections occurred at the beginning of the utterance. There were no utterances that had more than one interjection. In four sessions, more than one utterance had interjections. These sessions had a total of eight utterances.

There were seven utterances that had repeat/restart phrases. This is 2.2% of the total utterances. There were five simple repeat phrases. There was one phrase that was restarted for correction and there was one phrase that was repeated for emphasis.

There were 183 utterances that are dealt with the grammar and the lexicon of the system. This is 60.2% of the total utterances. Among them, the system correctly recognized 117 utterances. The recognition rate was 63.9%. The system could make the response required by the user for 154 of these utterances. The understanding rate was 84.2%. Among total valid utterances, the system could understand 208 utterances. This understanding rate was 68.4%. The understanding rate of the utterances including the unknown words or the unknown phrases was 41%. The understanding rate of the utterances with interjections was 26%, and the understanding rate of the utterances with repeat phrases was 33%. These rates are lower than the rates of the valid utterances.

5. Conclusion

This paper described issues of speech dialogue system design, spontaneous speech collection and system evaluation. The authors constructed a speech dialogue system which operates in almost real time, with acceptable recognition accuracy and flexible dialogue control.

The system was used for spontaneous speech collection in a transportation guidance domain. The system performance evaluated in the domain is the understanding rate of 84.2% for the utterances within the predefined grammar and the lexicon. Also some statistics of the spontaneous speech were given.

For future research topics, the following are thought to be important.

(1) More flexible dialogue—processing of unknown words,[(13)-(16)] parsing partial sentences, etc.

(2) Integration of knowledge sources—incremental disambiguation,[(17)] cooperative processing of various levels of knowledge sources

(3) More intelligent dialogue—plans and intentions, dynamic models and adaptation of language

The authors plan to pursue these in the hope of realizing natural conversations between humans and computers.

Acknowledgement

The authors would like to thank Mr. M. Tamoto for his technical assistance, and the members of the Speech Processing Section of ETL for their help. The "ASJ Continuous Speech Corpus for Research" was used for the training data.

References

(1) Niimi, Y. and Kobayashi, Y., "Topic management and keyword prediction in a speech dialog system," *IEICE Technical Report*, SP87-103, Dec. 1987.

(2) Murayama, H., Amano, A., Kitahara, Y. and Ichikawa, A., "Information retrieval system using continuous speech recognition," *Proc. ASJ spring meeting*, 3-6-11, pp. 131-132, Mar. 1991.

(3) Tsuboi, H., Hashimoto, H. and Takebayashi, Y., "Continuous speech understanding based on keyword-spotting," *IEICE Technical Report*, SP91-95, Dec. 1991.

(4) Price, P. J., "Evaluation of spoken language systems: the ATIS domain," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 91-95, Jun. 1990.

(5) Hemphill, C. T., Godfrey, J. J. and Doddington, G. R., "The ATIS spoken language systems pilot corpus," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 96-101, Jun. 1990.

(6) Zue, V., Glass, J., Goodine, D., Hirshman, L., Leung, H., Phillips, M., Polifroni, J. and Seneff, S., "The MIT ATIS system: preliminary development, spontaneous speech data collection, and performance evaluation," *Eurospeech 91*, pp. 537-540, 1991.

(7) Hayamizu, S., Itou, K. and Tanaka, K., "A spoken language dialog system for spontaneous speech collection," *IEICE Technical Report*, SP91-101, Dec. 1991.

(8) Itou, K., Hayamizu, S. and Tanaka, H., "Processing unknown words in speech dialog system," *JSAI Technical Report*, SIG-SLUD-9201-1, Apr. 1992.

(9) Itou, K., Hayamizu, S. and Tanaka, H., "Continuous speech recognition by the generalized LR parsing," *IEICE Technical Report*, SP90-74, Dec. 1990.

(10) Itou, K., Hayamizu, S. and Tanaka, H., "Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses," *IEEE Proc.* ICASSP-92, vol. I, pp. 21-24, 1992.

(11) Hayamizu, S. and Tanaka, K., "Prediction about unknown phonetic contexts by tree-based phone modeling," *IEICE Technical Report*, SP90-64, Dec. 1990.

(12) Hayamizu, S., Sanaka, K., Yokoyama, S. and Ohta, K.,

"Generation of VCV/CVC balanced word sets for speech data base." *Bul. Electrotech. Lab.*, vol. 49, no. 10, pp. 803-834, 1985.

(13) Itou, K., Hayamizu, S. and Tanaka, H., "Processing unknown words in continuous speech recognition," *IEICE Technical Report*, SP91-96, Dec. 1991.

(14) Asadi, A., Schwartz, R. and Makhoul, J., "Automatic detection of new words in a large vocabulary continuous speech recognition system," *IEEE Proc.* ICASSP-90, pp. 125-128, 1990.

(15) Gorin, A L., Levinson, S. E. and Gertner, A. N., "Adaptive acquisition of spoken language," *IEEE Proc.* ICASSP-91, pp. 805-808, 1991.

(16) Tsukada, S., Watanabe, T. and Yoshida, K., "Evaluation of adaptive likelihood normalization for speech recognition," *Proc. ASJ spring meeting*, 1-P-11, pp. 125-126, Mar. 1992.

(17) Okumura, M. and Tanaka, H., "Incremental disambiguation model of natural language analysis," *Journal of JSAI*, vol. 4, no. 6, pp. 687-694, 1989.

Katunobu Itou is currently pursuing his studies toward the D.E. in Tokyo Institute of Technology. His areas of interest include speech understanding system, and natural language processing. He received the B.E., M.E. degrees in computer science from Tokyo Institute of Technology in 1988, 1990. He is a member of the Japanese Society for Artificial Intelligence, and the Acoustical Society of Japan.

Satoru Hayamizu received the B.E. and M.E. degrees from the University of Tokyo in 1978 and 1981, respectively. Since 1981, he has been working on speech recognition, speech understanding, and dialogue processing at Electrotechnical Laboratory (ETL). From 1989 to 1990, he was a visiting scholar in Carnegie Mellon University. He is a member or the Japanese Society for Artificial Intelligence, the Acoustical Society of Japan, the Japan Society of Mechanical Engineers, the Information Processing Society of Japan, the Japanese Cognitive Science Society, the Institute of Electrical and Electronics Engineers, and the European Speech Communication Association.

Kazuyo Tanaka received the B.E. degree from the Yokohama National University, Yokohama, Kanagawa, in 1970, and the Dr.Eng. degree from Tohoku University, Sendai, Miyagi, in 1984. In 1971, he joined the Electrotechnical Laboratory (ETL), Tsukuba, Ibaraki, where he has been working on speech analysis, recognition and understanding. He is now a chief of the Speech Processing Section, Machine Understanding Division, ETL. Dr. Tanaka is a member of the Acoust. Soc. Japan and the IEEE. His research interests include speech processing, dialogue understanding, machine learning and pattern recognition theories.

Hozumi Tanaka received the B.E. and M.E. degrees from Tokyo Institute of Technology in 1964 and 1966, respectively. In 1966, he joined the Electrotechnical Laboratory (ETL). In 1983, he joined the Faculty of Engineering, Tokyo Institute of Technology, and now he is a Professor of the Faculty of Engineering. He has been engaged in researches on natural language processing and artificial inteligence. Dr. Tanaka is a member of the Information Processing Society of Japan, the Japanese Cognitive Science Society, the Japanese Society for Artificial Inteligence, Japanese Society for Software Science and Technology, and Mathematical Linguistic Society of Japan.