ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HANDLING RICH TURN-TAKING IN SPOKEN DIALOGUE SYSTEMS

Mikio Nakano, Kohji Dohsaka, Noboru Miyazaki, Jun-ichi Hirasawa,
*Masafumi Tamoto, Masahito Kawamori, Akira Sugiyama, Takeshi Kawabata**
NTT Laboratories
3-1 Morinosato-Wakamiya, Atsugi 243-0198, Japan
nakano@atom.brl.ntt.co.jp, dohsaka@atom.brl.ntt.co.jp, nmiya@atom.brl.ntt.co.jp,
jun@idea.brl.ntt.co.jp, tamoto@idea.brl.ntt.co.jp, kawamori@atom.brl.ntt.co.jp,
sugiyama@atom.brl.ntt.co.jp, kaw@nttspch.hil.ntt.co.jp

10.21437/Eurospeech.1999-299

ABSTRACT

This paper discusses how to build a system that can engage in a mixed-initiative human-machine spoken dialogue in which system utterances sometimes overlap with user utterances and *vice versa*. In the method, a module that incrementally understands user utterances and another module that incrementally generates system utterances work in parallel, and the timing of taking and releasing the dialogue initiative is decided according to the understanding of user utterances and the content of the system utterances. This method enables the system to respond when the user holds the dialogue initiative and is speaking, and enables the system to react to the user's barge-ins when it holds the initiative and is speaking. An experimental system called DUG-1 is also presented.

1. INTRODUCTION

Recent advances in speech recognition technology have made it possible to build spoken dialogue systems that anonymous users can use. For these systems to become commercially viable, however, their being able to complete a task is not sufficient; they must be usable. The achievement of *rich turn-taking* in a spoken dialogue system is crucial to the system's usability. Here, rich turn-taking refers to the phenomena where utterances of the two dialogue participants overlap.

Previous spoken dialogue systems do not start speaking unless the user explicitly notifies the system of the end of his/her utterance by using clues such as long pauses or special keywords or the mouse [2, 3, 13, 17]. In addition, these systems cannot recognize and understand user utterances made before their utterances have finished. These limitations prevent rich turn-taking, and thus the systems are not easy to use. Although some systems that can either barge into the user's utterances or accept the user's barge-ins have been developed [1, 6, 7, 8, 16], a system has yet to be developed that can both make barge-in responses and accept the user's barge-ins.

This paper proposes a method for handling rich turn-taking in a mixed-initiative spoken dialogue system. It also presents an experimental system called DUG-1. DUG-1 achieves rich turn-taking through the incremental understanding of user utterances and the incremental generation of system utterances. Rich turn-taking is also achieved by having the understanding module and the generation module work in parallel.

We first discuss the necessity of handling rich turn-taking, and then present a method for handling it. Finally, we present an experimental system DUG-1.

2. NECESSITY OF HANDLING RICH TURN-TAKING

Utterances by one participant in human-human dialogues sometimes overlap with the other participant's utterances [12]. This phenomenon is called *rich turn-taking* in this paper. Needless to say, human-machine dialogues do not have to be like human-human dialogues from the standpoint of building a spoken dialogue system as a user-friendly human-machine interface; rich turn-taking does not have to be dealt with by spoken dialogue systems only because utterances overlap in human-human dialogues.

Nevertheless, handling rich turn-taking is crucial to the usability of the spoken dialogue systems for a number of reasons. First, let us consider the system's responses during user utterances. With rich turn-taking, the system could acknowledge the user's utterances by making backchannel utterances while the user is speaking, enabling the user to speak easily [9]. In addition, when the system cannot understand what the user is saying, it could interrupt the user's utterance so that the user need not uselessly finish the utterance. Next, let us consider the user's barge-ins. The user could save time by interrupting the system to move the dialogue forward. In addition, the user could barge into the system's utterance to request the system to clarify or repeat something and the system would comply.

Rich turn-taking has not been implemented in most previous spoken dialogue systems; they can only deal with orderly turn-taking [2, 3, 13, 17]. They do not have mechanisms for producing barge-in responses because

* Current address: NTT Laboratories, 1-1 Hikarino-oka, Yokosuka 239-0847, Japan

they determine the timing of taking turns by monitoring the length of pauses or detecting a special word at the end of the user's utterances [15]. In addition, they cannot accept the user's responses while they are speaking. Although some systems that can either barge into the user's utterances [1, 7, 16] or accept the user's barge-ins [6, 8] have been developed, there has been no system that can both make barge-in responses and accept the user's barge-ins.

A *mixed-initiative spoken dialogue system* must deal with both types of barge-in. In this paper, we use the term *mixed-initiative* in contrast to *fixed-initiative*. In a fixed-initiative dialogue, the participant who holds the initiative asks questions of the other participant who only answers the questions or only explains things to the participant who holds the initiative. In a mixed-initiative dialogue, on the other hand, the initiative moves among the participants and each participant can ask questions and explain things. If a mixed-initiative dialogue system is to handle rich turn-taking, it must be able to respond when the user is speaking and holds the dialogue initiative, and it must be able to react to the user's barge-ins when it is speaking and holds the initiative.

3. A METHOD FOR HANDLING RICH TURN-TAKING

This section proposes a method for handling rich turn-taking in mixed-initiative spoken dialogue systems and presents a system architecture that implements the method. It is based on concurrent processing and utilizes an incremental utterance understanding method [10] and an incremental utterance generation method [5, 6]. The architecture consists of the speech recognition module, the speech production module, and the language processing module, which comprises two submodules called the understanding module and the generation module. Figure 1 depicts the relationship among these modules.

In this architecture, the language understanding module and the language generation module work in parallel so that utterance generation is possible even while the system is listening to user utterances and that utterance understanding is possible even while it is speaking [14]. The understanding module incrementally understands user utterances by updating the partial results of understanding every time a word hypothesis is obtained from the speech recognizer, and the generation module incrementally produces system utterances phrase by phrase by referring to domain knowledge. The result of utterance understanding and the content of system utterances are written in shared memory so that both the understanding and generation modules can access them.

When the user holds the initiative in the dialogue, the generation module can respond based on the partial results of incremental understanding. It can make backchannels and confirm when it concludes from understanding results that these actions are required. The timing of responses is determined according to partial understanding results as well as user pauses. The speech recognizer detects pauses and notifies the language generation module.

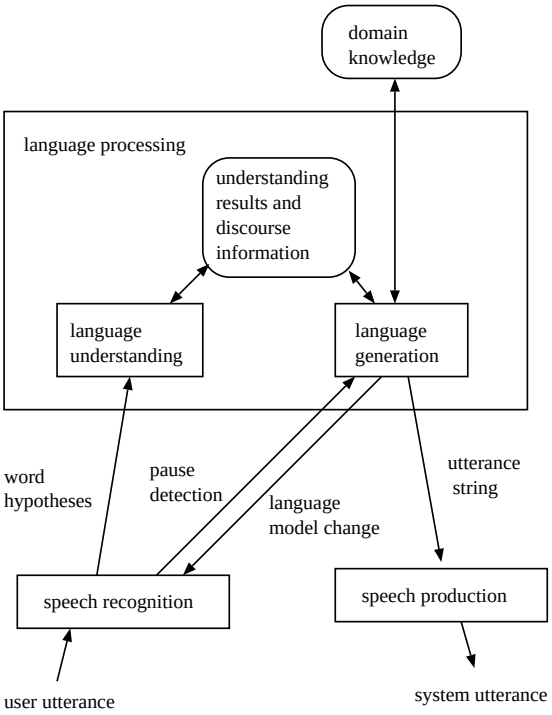

Figure 1: Spoken dialogue system architecture

When the system holds the initiative, since the generation module incrementally produces system utterances, the system can dynamically change the content of its utterances according to the user's backchannels, confirmations, and requests for re-explanation regardless of whether they overlap with system utterances or not.

The system takes or releases the initiative in the dialogue according to the result of user utterance understanding and the content of system utterances. For example, when the system thinks that it has understood the user's request, it takes the initiative to answer, and when the system finishes a question to the user, it releases the initiative to wait for the user's answer.

The language model for speech recognition can be dynamically switched according to the result of understanding and the content of system utterances to increase the accuracy of the recognition.

4. IMPLEMENTATION

4.1. System Features

We have developed a Japanese spoken dialogue system called DUG-1 based on the architecture explained in the previous section.

The speech recognition module in DUG-1 is a phoneme-HMM-based speaker-independent continuous speech recognizer [11] that can incrementally output word hypotheses using the ISTAR (Incremental Structure Transmitter And Receiver) protocol [7]. This recognizer out-

puts at each time frame the word hypothesis on the search path with the highest score. The speech recognition is directed by network grammars. The constraints posed by these grammars are weak enough to capture spontaneously spoken utterances, which sometimes include fillers and self-repairs. The grammars allow each speech interval to be an arbitrary number of arbitrary *bunsetsu* phrases. A *bunsetsu* phrase is a phrase that consists of one content word and a number (possibly zero) of function words. The grammars are switched according to the result of understanding and the content of system utterances to increase the accuracy of the recognition. When to switch the grammar is determined by the language generation module. The vocabulary size of each grammar is less than one hundred words so that the speech recognizer can work in real time.

When the user holds the initiative, the results of the utterance understanding are represented by a frame (i.e., attribute-value pairs with some procedural constraints) [4]. The frame is updated word by word by incremental understanding. The language generation module makes utterances including backchannels and confirmation based on the content of the frame.

When the system holds the initiative, the language generation module incrementally generates utterances in short units (such as clauses) using hierarchical planning. The current system understands user utterances using *bunsetsu* phrases and does not use parsing. It changes the content of the explanation according to the user's questions and backchannels based on the cooperative dialogue principles established based on the observation of human-human dialogues [6].

The speech production module outputs pre-recorded voices of bunsetsu phrases according to the requests of the language generation module. For the user to be able to speak at ease, a human-like face is displayed, which nods and moves its lips according to the content of system utterances.

Although this paper focuses only on rich turn-taking, DUG-1 has other unique features. For instance, when the user has the dialogue initiative, it can reply immediately after detecting a pause based on the partial result of understanding because it features an incremental understanding method.

4.2. Dialogue Task

One of the tasks of DUG-1 is to arrange the videorecording of TV programs. Before the dialogue, the user is assumed not to know the title of the program but to know some fragmentary information such as its category, who appears in it, and when it is being televised. The system has the complete timetable and it suggests a program according to the user's request. This task can be considered to be a kind of cooperative decision task.

DUG-1 completes this task as follows. At the first stage, the user has the initiative and informs DUG-1 of fragmentary information about the program he/she wants to record. DUG-1 then takes the initiative and tells the user the titles and features of the candidate programs. Then it releases the initiative. After the user chooses a program from the selection of candidates, the dialogue goes back to the first stage.

S1 *hai dôzo*
(May I help you?)
U2 *êtodesune*
(well)
S3 *hai*
(uh-huh)
U4 *supîdo ga deru asa janakute yoru no ⟨hai⟩ utabangumi o yoyaku shitaindesu keredomo*
(I want to record a music program in the morning, no, at night ⟨uh-huh⟩ where *supîdo* (a chorus group) appear)
S5 *supîdo ga deteru yoru no utabangumi desuka*
(is it a music program at night where *supîdo* appears)
U6 *hai*
(yes)
S7 *êtto poppujamu toyû bangumi ga ⟨hai⟩ arimasu*
(well there is a program whose title is *poppujamu* ⟨uh-huh⟩)
S8 *yoru no utabangumi de ⟨poppujamu deshita kke⟩ supîdo ga*
(it's a music program at night, and ⟨it was *poppujamu* wasn't it?⟩ *supîdo* is)
S9 *sôdesu*
(right)
U10 *hai*
(uh-huh)
S11 *poppujamu toyû bangumi ga arimasu*
(there is a program called *poppujamu*)
U12 *hai*
(uh-huh)
S13 *supîdo ga shutsuen shi masu*
(*supîdo* will appear in the program)
U14 *hai*
(uh-huh)
S15 *poppujamu toyû bangumi o yoyaku shimasu ka*
(do you want to record *poppujamu*?)
U16 *hai yoyaku shimasu*
(yes I do)
S17 *kashikomarimashita hoka ni gozai masu ka*
(all right, anything else?)
U18 *êto ⟨hai⟩ soredakede îdesu*
(well ⟨uh-huh⟩ that's all)
S19 *arigatô gozaimashita*
(thank you)

Figure 2: Example dialogue

4.3. Example Dialogue

Figure 2 shows an example dialogue between a user and the system. U means a user utterance and S means a system utterance. Utterances in angle brackets are barge-ins by the hearer. The system took the initiative at S7 and released it after S15.

Let us focus on turn-taking in this dialogue. First, the system made a backchannel during the user's utterance U4 based on the result of understanding of the phrase *yoru no* (at night). The system considered the user's backchannel *hai* during utterance S7 to indicate the user's understanding and thus it continued the explanation. When the system detected the user's barge-in during utterance S8, it stopped the subsequent explanation, answered *sôdesu* (right) in S9, and repeated the explanation. As this example shows, DUG-1 can handle rich turn-taking.

5. CONCLUSION

This paper proposed a method for handling rich turn-taking in mixed-initiative spoken dialogue systems, and also reported an experimental system called DUG-1.

ACKNOWLEDGMENTS

We would like to thank Dr. Yoh'ichi Tohkura, Dr. Ken'ichiro Ishii, Dr. Norihiro Hagita, and Dr. Kiyoaki Aikawa for their encouragement and helpful comments. We used in this research the speech recognition engine REX developed by NTT Cyber Space Laboratories and would like to thank those who helped us use it.

REFERENCES

[1] G. Aist. Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption. *ICSLP-98*, pp. 413–416, 1998.

[2] J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. *ACL-96*, pp. 62–70, 1996.

[3] H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The Philips automatic train timetable information system. *Speech Communication*, 17:249–262, 1994.

[4] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. GUS, a frame driven dialog system. *Artificial Intelligence*, 8:155–173, 1977.

[5] K. Dohsaka and A. Shimazu. A computational model of incremental utterance production in task-oriented dialogues. *COLING-96*, pp. 304–309, 1996.

[6] K. Dohsaka and A. Shimazu. System architecture for spoken utterance production in collaborative dialogue. *Working Notes of IJCAI 1997 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*, 1997.

[7] J. Hirasawa, N. Miyazaki, M. Nakano, and T. Kawabata. Implementation of coordinative nodding behavior on spoken dialogue systems. *ICSLP-98*, pp. 2347–2350, 1998.

[8] C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. Evaluating spoken dialog systems for telecommunication services. *Eurospeech-97*, pp. 2203–2206, 1997.

[9] M. Kawamori, A. Shimazu, and K. Kogure. Roles in interjectory utterances in spoken discourse. *ICSLP-94*, pp. 955–958, 1994.

[10] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka, and T. Kawabata. Understanding unsegmented user utterances in real-time spoken dialogue systems. *ACL-99*, 1999.

[11] Y. Noda, Y. Yamaguchi, T. Yamada, A. Imamura, S. Takahashi, T. Matsui, and K. Aikawa. The development of speech recognition engine REX. *Proceedings of the 1998 IEICE General Conference D-14-9*, pp. 220, 1998. (in Japanese).

[12] N. Osaka. An analysis of address-response relation in conversational speech centering on listening response. *SIG-SP87-107, Institute of Electronics, Information and Communication Engineers*, 1987. (in Japanese).

[13] J. Peckham. A new generation of spoken language systems: Results and lessons from the SUNDIAL project. *Eurospeech-93*, pp. 33–40, 1993.

[14] A. Shimazu, K. Kogure, M. Kawamori, K. Dohsaka, and M. Nakano. Internal communication in dialogue processing systems. *Proceedings of the Second Annual Meeting of the Association for Natural Language Processing*, pp. 333–336, 1996. (in Japanese).

[15] R. W. Smith and D. R. Hipp. *Spoken Natural Language Dialog Systems*. Oxford University Press, 1994.

[16] N. Ward. Using prosodic clues to decide when to produce back-channel utterances. *ICSLP-96*, pp. 1728–1731, 1996.

[17] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goodine, D. Goddeau, and J. Glass. PEGASUS: A spoken dialogue interface for on-line air travel planning. *Speech Communication*, 15:331–340, 1994.