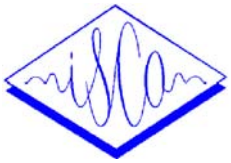

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

EFFECTS OF SYSTEM BARGE-IN RESPONSES ON USER IMPRESSIONS

Jun-ichi HIRASAWA, Mikio NAKANO, Takeshi KAWABATA, Kiyoaki AIKAWA
NTT Laboratories
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, Japan
jun@idea.brl.ntt.co.jp

10.21437/Eurospeech.1999-359

ABSTRACT

When designing a spoken dialogue system, in particular a real-time one, not only what the system responds but also when it responds need to be considered. This paper focuses on when the system should appropriately respond with backchannels, and reports an experiment that compared two response-time conditions: the immediate response and the orderly response. The results of the experiment show that the immediate barge-in backchannels can cause some subjects to have negative impressions of the system. This implies that the orderly response strategy is better and less risky than the simple immediate strategy and that some considerations might be required in designing a better barge-in response.

1 INTRODUCTION

When designing a spoken dialogue system, in particular a real-time one, we need to consider not only what the system responds but also when it responds. The system response time is an important factor in designing the system. If the system is not designed to respond at an appropriate time, the user might have uncomfortable impressions, misinterpret the response, or be delayed in completing a task. The appropriate system response time, however, has never been clear for system designers.

There are a wide variety of the possible system responses, and here we focus on backchannel responses. Backchannels are supposed to contribute to the comfort of communication between the user and the system. This is because backchannels have a function to acknowledge a user utterance, and this acknowledgment leads to mutual understanding between the user and the system [2].

In this paper, we investigate when the system should respond with a backchannel at the appropriate time. In human-human dialogues we can find frequent barge-in backchannels. Watanuki et al. [5] report that 67% of listener's responses overlap with a speaker's utterance during human-human dialogues, and backchannels of a human respondent are uttered 0.35 seconds after the end of each keyword in the speaker's utterance on average. This fact leads us to the hypothesis that the most appropriate time for the system backchannel should be immediately after the system's acceptance of important information from the user utterance. This immediate acknowledgment can appear as the system barge-in backchannel even during the user utterance.

If the target information is followed by such a system backchannel, it should be clear which information is acknowledged. Thus, immediate system barge-in backchannels might cause the users to have favorable impressions of the system. However, the opposite is possible. If a system barge-in backchannel disturbs a user utterance, users might have negative impressions of the system.

We conducted a dialogue experiment by using the Wizard of OZ (WOZ) method to investigate the effect of immediate responses with barge-in backchannels. In our experiment, the immediate response strategy was compared with a response in orderly turn-taking, which never overlaps with the user speech intervals.

2 RELATED WORK

There is a great amount of research on backchannels. However, most of it tends to focus on human-human dialogues and not on human-computer dialogues.

Okato et al. [3] conducted a WOZ experiment to investigate the effect of system backchannels. Their system makes a backchannel near the end of a user utterance if the utterance includes any keywords about executing a telephone shopping task. They did not control the time of the backchannels. Ward [4] built a system that can produce a barge-in backchannel even during a user utterance, but his backchannel is not content-driven but prosody-driven, so the function of his backchannel in the system only follows the pace of the user speaking. Aist [1] built a system that can handle content-driven interruption. Although his system can produce not only a backchannel but also a corrective utterance, it is made for a reading-tutor task, in which a user (child) reads aloud a sentence on the display. This is quite different from a spontaneous utterance in actual dialogues. We have built a system that can produce content-driven barge-in backchannels [2]. We have not yet evaluated the effect of immediate barge-in

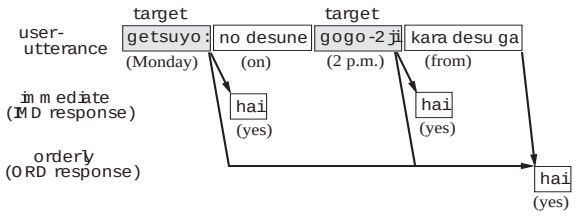

Figure 1: The difference between two response-time conditions, IMD and ORD.

backchannel responses in human-computer dialogues.

3 EXPERIMENT

We conducted an experiment with the purpose of investigating how the response times with barge-in backchannels affect user impressions.

Design. The experiment was arranged as a 2×2 mixture design. The first factor was the **response-time condition**, which had two different backchanneling times (Figure 1). One system made a backchannel immediately after the acceptance of the important information in a user utterance. This could be a barge-in backchannel uttered even during a user utterance. This is the **immediate (IMD) condition**. The other system responded with a backchannel just after the end of a user utterance. This backchannel kept orderly turn alternation and never overlapped with the user speech interval. This is the **orderly (ORD) condition**. A speech interval detector controlled the response times. The details will be described later.

The second factor was the **subject-sensitivity** to the system response time. Two subject groups were used. When a subject finished all the dialogues with the two conditioned systems, the experimenter asked whether the subject had noticed any differences among the dialogues. Subject who referred to the difference in the time of the backchannel in the interview, were put into one group (the **conscious group**). Subject who did not notice any difference were put into another group (the **unconscious group**).

The purpose of this experiment was to confirm whether the backchannels in the immediate (IMD) condition could improve the user impressions compared to the orderly (ORD) condition, no matter how conscious or unconscious the subject is.

Apparatus. We adopted the WOZ method for this experiment. The reason why we did not use a system with a continuous speech recognizer was to exclude any differences other than the system response time. A system with a continuous speech recognizer might have some recognition errors, which could cause the subject to have a negative impression. The system with the CSR might also have some response delays, which can prevent us from controlling the response time.

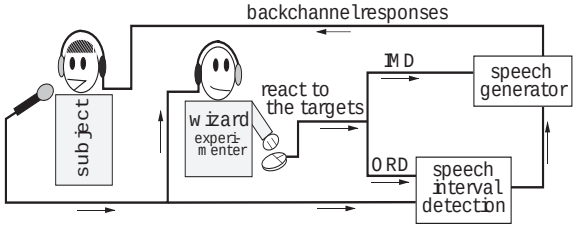

Figure 2: The Wizard of OZ experimental setup.

In our WOZ experiment, the wizard (experimenter) listened to the user's speech and reacted to target expressions in a user utterance (Figure 2). This enabled the system to produce backchannels according to the wizard's correct detection of the target expressions without any misrecognition. In the immediate (IMD) response condition, the system uttered a backchannel at the same time that the wizard detected the target. The wizard caused only 0.37-second delay on average. In the orderly (ORD) response condition, the reaction of the wizard went through a speech interval detector (Figure 2). This architecture made it possible for the backchannel in the ORD condition to be uttered only after the end of the user utterance without any barge-ins. Both in the IMD and in the ORD conditions, the response time could be controlled as expected.

The important target information that the wizard should react to was defined in advance. The task of the dialogues was making a reservation for a meeting room. The day, the room name, and the beginning and end times for the reservation were defined as the important target information.

Modality. The subject input and the WOZ system output were comprised of speech only. The output speech from the system was made up of pre-recorded voices, because synthesized speech might cause the subject to misunderstand.

Subjects. Ten subjects (five males and five females from 24 to 35 years old) took part in the experiment. They were not accustomed to using a spoken dialogue system. They were divided into the two sensitivity groups: four were in the conscious group and six in the unconscious group.

Procedure. The procedure of the experiment was as follows: (1) Each subject was instructed "to check whether the specified meeting room is vacant or occupied by talking with the dialogue system"; (2) the subject started a dialogue; (3) after each dialogue, the subject evaluated his/her impressions about the system on a scale one to five. Subjects answered questions about the system. (4) After evaluating the system, the subject went on to the next dialogue. Each subject had eight dialogues. After all the dialogues, the experimenter interviewed the subjects about whether they had noticed any differences among the dialogues.

Table 1: Results of the evaluation.

subject's sensitivity	conscious 4 subjects		unconscious 5 subjects		
response time	**(A)** immediate 10 dialogues	**(B)** orderly 10 dialogues	**(C)** immediate 9 dialogues	**(D)** orderly 9 dialogues	
items	mean (S.D.)	mean (S.D.)	mean (S.D.)	mean (S.D.)	effects
speed	*4.20*(0.92)	3.10(0.57)	*4.00*(0.71)	3.56(0.88)	$*$
easy to speak to	2.50(1.58)	*4.00*(0.82)	3.56(0.73)	*3.89*(0.60)	$*$
easy to use	2.70(1.16)	*3.70*(0.48)	*3.78*(0.44)	3.44(0.73)	$***$
likes the system	2.70(0.67)	*3.50*(0.53)	*3.56*(0.53)	3.44(0.73)	$***$
comfortable to use	2.60(1.07)	*3.40*(0.52)	3.78(0.67)	3.78(0.67)	$**$
easy to understand	3.50(0.71)	*3.90*(0.57)	4.11(0.33)	*4.22*(0.67)	$**$
friendly	3.30(0.67)	*3.50*(0.85)	*3.78*(0.44)	3.56(0.53)	n.s.
human-like	3.20(1.14)	*3.80*(1.03)	2.89(0.93)	*3.00*(1.12)	n.s.

Main effect in the response-time ($*$), in the subject's sensitivity ($**$) and in the interaction ($***$).

4 RESULTS

Among the 80 dialogues collected, 26 dialogues did not satisfy the response-time control and were discarded. In addition, 16 were discarded at random so that the number of the dialogues in each of two response-time conditions within a subject would be equal. As a result, 38 dialogues by nine subjects were analyzed (Table 1): 20 (ten in the IMD and ten in the ORD) by four time-conscious subjects, and 18 (nine in the IMD and nine in the ORD) by five time-unconscious subjects. For each dialogue, the subjects were provided eight items and were asked to give their impressions on a scale of five. The mean and the standard deviations of the evaluation for each item are shown in Table 1.

Two items, "speed" and "easy to speak to", were similar for the two subject groups (conscious and unconscious), and significantly different between the response-time conditions (IMD and ORD).

Speed. The mean score of "speed" in the IMD condition was evaluated to be significantly higher than in the ORD condition ($F_{(1,34)}$=9.27, p <.01, Figure 3, left). The interaction had no significant effect between the response-time and subject-sensitivity factors. The subjects of both sensitivity types indicated that the immediate responses had been quicker than the orderly responses, which demonstrated that the response-time was successfully controlled in the experiment.

Easy to speak to. The ratings of "easy to speak to" also showed the effectiveness of the response-time factor ($F_{(1,34)}$=7.60, p <.01, Figure 3, right). The ORD system was found to be easier to speak to than the IMD system. Thus, the IMD responses might provide difficulties in speaking for both types of subject.

The next two items, "easy to use" and "likes the system", were effective in interaction between the response-time conditions and the subject-sensitivity group.

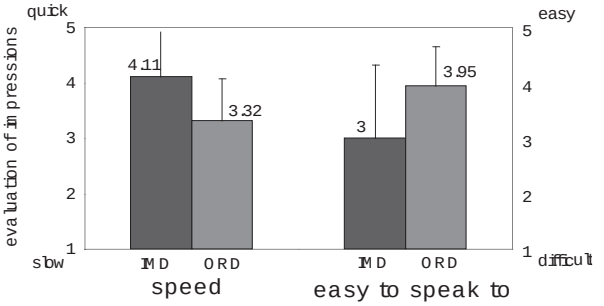

Figure 3: The evaluation of "speed" and "easy to speak to" in the immediate and orderly conditions.

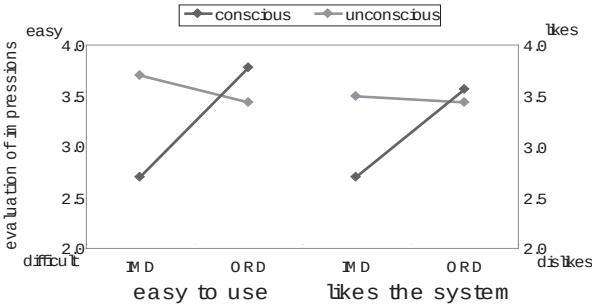

Figure 4: The evaluation of "easy to use" and "likes the system" in the response-time $\times$ sensitivity factor.

Easy to use and likes the system. "Easy to use" and "likes the system" measurements had the interaction effect (Figure 4, easy to use: $F_{(1,34)}$=7.17, p <.05; likes the system: $F_{(1,34)}$=5.12, p <.05). The simple main effect of the response-time was significant at the conscious group both in "easy to use" ($F_{(1,18)}$=6.34, p <.05) and in "likes the system" ($F_{(1,18)}$=8.73, p <.01), but not significant at the unconscious group. The simple main effect of the subject-sensitivity was significant at the IMD condition both in "easy to use" ($F_{(1,17)}$=6.85, p <.05) and in "likes the system" ($F_{(1,17)}$=9.32, p <.01), but not significant at the ORD condition. Thus, the orderly responses were favorably evaluated by all subjects, but the immediate responses were favorably evaluated only by the

unconscious subjects. The conscious subjects considered the immediate responses "more difficult to use" and caused them to "dislike the system".

The next two items, "comfortable to use" and "easy to understand", were significantly different only between the two subject groups, and the last two items, "friendly" and "human-like", were not effective between the response-time conditions and the subject-sensitivity group.

Comfortable to use and easy to understand. Regardless of the response-time conditions, the unconscious group evaluated the system better for "comfortable to use" ($F_{(1,34)}$=9.79, $p < .01$) and "easy to understand" ($F_{(1,34)}$=5.92, $p < .05$).

Friendly and human-like. "Friendly" and "human-like" showed neither the main effects nor the interaction effect. These two items seem independent of the response-time and of the subject-sensitivity, or are inappropriate because they had large variances.

5 DISCUSSION

Immediate barge-in responses. The results of the experiment demonstrated that simple immediate barge-in backchannels made it more difficult for both sensitivity groups to speak to the system. It also makes only the conscious group more likely to dislike the system and find it difficult to use.

Although not all immediate backchannels caused damage, some immediate barge-in backchannels stopped the user from speaking or made the user hesitate in speaking. These results lead us to think that the immediate response strategy may lower the user's impressions, and that the orderly turn-taking strategy does not cause negative impressions because the orderly alternation never disturbs a user's utterance. We can conclude that the orderly (ORD) turn-taking response strategy would be better and less risky than the immediate (IMD) response strategy for both types of subject.

Even though the IMD strategy is presently too simplistic, we believe that it can be further developed and improved upon. To design a better system barge-in response, some aspects need to be taken into consideration. For example, a user utterance can include some parts that are proof against barge-in interruption and the other parts that are vulnerable to it. Use of prosodic information in the user utterance may help the system discriminate between these two parts. Intonation of the system response should also be considered. The same utterance varies with its intonation patterns in spoken dialogue. *"Hai"* ("yes" in Japanese) with a high pitch might disturb a user utterance, while *"hai"* with a low pitch might not. The analysis of the effects of these factors needs further research.

Subject Sensitivity. The results of "comfortable to use" and "easy to understand" demonstrated that the time-conscious group tended to provide a poorer system evaluation than the time-unconscious group, regardless of the system response-time conditions. While we were unable to determine whether the subject sensitivity (conscious and unconscious) was due to an innate cause, the results of the experiments are still quite intriguing.

6 CONCLUSION

This paper reported our investigation of when a system should respond with backchannels appropriately. We compared two response-time conditions: the immediate response and the orderly response conditions. The results show that the immediate backchannels could cause some subjects to have negative impressions and find the system difficult to use and speak to. This implies that the orderly response strategy would be better and less risky than the simple immediate strategy and that some considerations might be required in designing a better barge-in response.

ACKNOWLEDGMENTS

We would like to thank Dr. Norihiro Hagita, the executive manager of the Media Information Laboratory, for his encouragements and comments. We also would like to thank the members of the Dialogue Understanding Research Group for their valuable discussions.

REFERENCES

[1] Aist, G. "Expanding a Time-Sensitive Conversational Architecture for Turn-Taking to Handle Content-driven Interruption", *Proc. of International Conf. on Spoken Language Processing*, vol. 2, pp. 413-416, 1998.

[2] Hirasawa, J., Miyazaki, N., Nakano, M. and Kawabata, T. "Implementation of Coordinative Nodding Behavior on Spoken Dialogue Systems", *Proc. of International Conf. on Spoken Language Processing*, vol. 6, pp. 2347-2350, 1998.

[3] Okato, Y., Kato, K., Yamamoto, M. and Itahashi, S. "System-User Interaction and Response Strategy in Spoken Dialogue System", *Proc. of International Conf. on Spoken Language Processing*, vol. 2, pp. 495-498, 1998.

[4] Ward, N. "Using Prosodic Clues to Decide When to Produce Back-channel Utterances", *Proc. of International Conf. on Spoken Language Processing*, pp. 1728-1731, 1996.

[5] Watanuki, K., Sakamoto, K. and Togawa, F. "Analysis of Multimodal Interaction Data in Human Communication", *Proc. of International Conf. on Spoken Language Processing*, vol. 2, pp. 899-902, 1994.