SYSTEM-USER INTERACTION AND RESPONSE STRATEGY IN SPOKEN DIALOGUE SYSTEM

Yohei Okato *Keiji Kato* *Mikio Yamamto* *Shuichi Itahashi*
University of Tsukuba

ABSTRACT

There are a number of restrictions in human-machine interactions, which continue to warrant a better control of response utterances in spoken dialogue systems. Indeed, the human user often has to deal with unnatural responses, and therefore requires some experience with such systems in order to improve interactions. This problem is re-examined here, with the aim of evaluating how human users are influenced by utterances built into a system based on the Wizard-of-Oz method. We report results which show that back-channel responses and brief confirmations from our system, have the effects of prompting human spoken interactions and providing more human satisfaction.

1. INTRODUCTION

Some spoken dialogue systems[1, 2] and dialogue strategies[3, 4] have recently been described as useful, because they prompt more spontaneous utterances from the user. However, there are a number of restrictions in human-machine interactions which still warrant further improvements. Indeed, users often have to deal with unnatural responses, and therefore require some experience with such systems in order to improve interactions. In addition, system responses are commonly determined in a heuristic manner, and their timing is still not sufficiently taken into account. Clearly, systems' spoken responses are expected to influence human-machine interactions[5, 6], and therefore it is important to improve the usability of spoken dialogue systems.

To this end, it is pertinent to observe that, in a completely human dialogue, the listener can play an influential role in the style of utterances chosen by the speaker. For example, the speaker is likely to pay attention to the listener's reactive speech and bodily gesture, in order to gauge the listener's degree of understanding. Similarly, in a human-machine dialogue, it is expected that the human user would feel more comfortable, if the system involved were to select an appropriate style of utterances and manage real-time responses such as a back-channel feedback[7].

In this paper a spoken dialogue system is presented, which involves the Wizard-of-Oz(WOZ) method[8] and incorporates a response strategy aimed at securing confirmation. The style of system utterances and timing behaviours of the responses will be our focus in outlining our response strategy. Experiments are also described, which were carried out to investigate the effects of our response strategy.

2. RESPONSE STRATEGY OF THE SYSTEM

As pointed out earlier, the handling of response utterances in current spoken dialogue systems is subject to a number of restrictions. Responses are generated upon detection of the user's end-of-utterance, and grammatical correctness is considered as important in many spoken dialogue systems. Yet, in completely human dialogues, confirmation is often conveyed in real time and in a way similar to back-channel feedback. Thus, response time and semantics would seem to be relatively more important than syntactic exactness, and to have the potential of inducing more liveliness in the responses of a spoken dialogue system.

In order to investigate ways in which system-response strategies influence users, we focused on two types of responses: (1) interjectory responses to the user's utterances; and (2) verbose responses which vary between brief and detailed expression used for confirmation. Assuming four qualitative degrees for each type, we then defined a system's response strategy as outlined in Table 1.

Keywords were also used which, by definition, cannot be removed in the communication process. Thus, back-

	STRATEGY			
	A	B	C	D
Interjection	no	yes	no	yes
Verbosity	much	much	little	little

Table 1: Response Strategy of the System

channel feedbacks are provided by the system if the user's phrase contains keywords and, in brief-response mode, the system simply confirms the keywords contained in the user's speech. Figure 1 gives an example of a keyword, and Figure 2 illustrates a back-channel feedback.

S: You / order/ a bag.
(Keyword: a bag)

Figure 1: Example of a "keyword"

U:The order is a key holder,please.
S: {Uh-huh}
({Uh-huh} is a listener's back-channel feedback)

Figure 2: Example of a "back-channel feedback", i.e., an utterance with no explicit meaning.

3. DATA COLLECTION

As stated in the introduction, we designed a spoken dialogue system based on the WOZ method, in order to investigate system responses to users. The dialogue task selected consists of telephone shopping, but the scenario adopted is a simplified version of real telephone shopping. Every user orders items from a list known beforehand and, for each item on the list, the user has to specify name, ID and quantity. The system then confirms the item ordered by the user and, upon completion of the order, it queries the user as to the form of payment (e.g., bank/card). Figure 3 illustrates both the flow of the entire dialogue and the flow of each sub-dialogue (ordering items and form of payment).

As for our data collection system, a WOZ operator's assistance was used rather than modules for speech recognition and back-channel feedback generation. We adopted this approach partly because the evaluation of our system performance would be compounded by speech recognition accuracy, and partly because automatic generation of back-channel feedbacks would require reaction at any time. Data were recorded in a soundproof room. The "MILES" [9] communication architecture designed to handle timing relations between

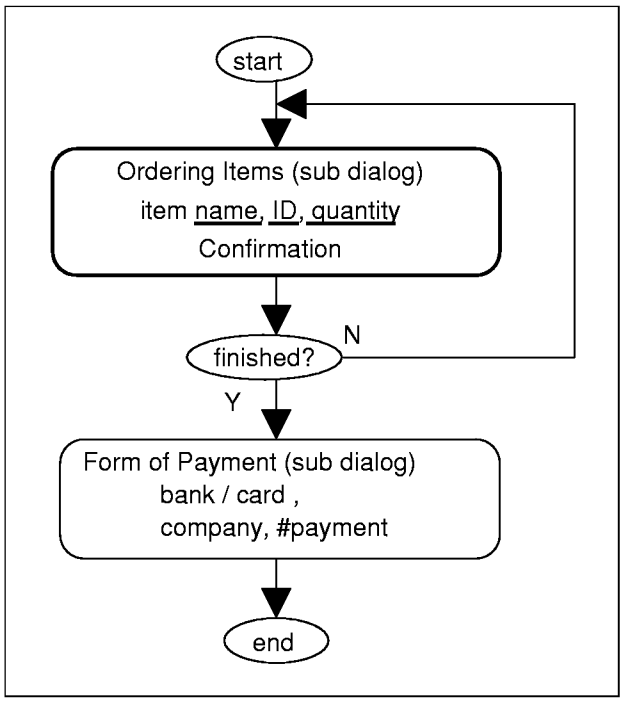

Figure 3: Dialogue flow

events and communicative elements, was used for the dialogue management module of our spoken dialogue system. System responses were generated using an NTT "Shaberinbo" synthesiser.

As for the data used, we first collected human-human dialogues, in which we brought situations as close as possible to real conversations. Then, we collected human-machine dialogues using the WOZ system; one WOZ operator handled the system with consistency during data collection. Users did not know who the operator was, nor were they aware of differences between sessions. All users were students who had not previously used spoken dialogue systems but who were given some brief instructions on usage. In all, we collected 64 human-machine dialogues from 16 users and in 4 sessions for each user. The actual speech data were segmented in utterance units (**UU**), which are defined as prosodic phrases preceded or followed by 100-msec pauses. 3797 UUs obtained from human-human dialogues(170-minute duration), were used for analysis. 1235 UUs obtained from human-human dialogues(21-minute duration) described previously, were used for comparison.

4. DIALOGUE ANALYSIS

In our dialogue analysis we first considered duration differences among the four strategies defined earlier, and

STRATEGY	#sub-dialogues	#UU MEAN	#UU SD	UU-duration(sec) MEAN	UU-duration(sec) SD
A	34	6.6	2.4	5.8	3.3
B	18	6.9	2.3	5.4	2.4
C	27	6.8	2.5	5.3	2.2
D	30	7.9	2.8	6.6	3.0

Table 2: Characteristics of user utterances

the number of occurring UUs. In order to normalise differences other than those among response strategies, we analysed only the sub-dialogues related to ordering items for which the system requires name, ID and quantity. Nor did we retain the sub-dialogues with recognition error. Table 2 gives a profile of the characteristics of users' utterances, which are interpreted in terms of the number of UU's and their duration. In particular, these tabulated results indicate that most users' utterances are prompted by the system with brief expressions and back-channel feedbacks.

In experiments B and D, back-channel feedbacks are given to users and therefore their timing is an important factor which we have also analysed. Figure 4 shows a histogram distribution of the times elapsed between users' utterances and back-channel feedbacks given by the system. The distribution's mean is 0.25 sec compared to 0.0 in human-human dialogues, owing to delays caused by the WOZ operator and the system. However, relative to other slower responses by the system, the delays contributed by back-channel feedbacks are permissible.

Figure 5 shows a bar-graph of the rates of interjection, which includes fillers, as well as back-channel feedbacks and repair, ranging from 14% for strategy A to 20% for strategy B. Note that the rate obtained for human-human dialogues peaks at 29%, and that such a rate can reach 50% for various types of Japanese dialogues[10]. We conclude that our response strategy tends to promote users' utterances, thereby rendering dialogues more natural.

While fillers are more common in users' interjections, the latter can also consist of a greater number of back-channel feedbacks depending on the user and on whether the system itself gives such feedbacks. Thus, in order to investigate the relationship between the system's response strategy and the rate of back-channel feedbacks by users, our data were merged pairwise (A+C) and (B+D) in order to secure more statistical significance. The relationship obtained is significant at the 5% level and can be depicted in Figure 6, which shows a 3% and a 1% rate of back-channel feedbacks by users when such feedbacks are respectively given and not given by our

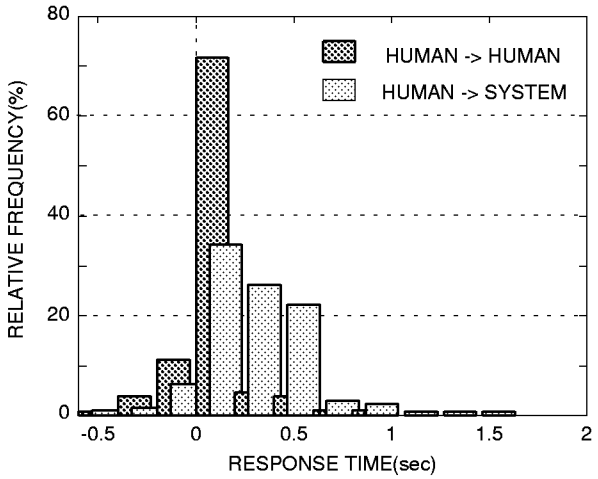

Figure 4: Response times of back-channel feedbacks

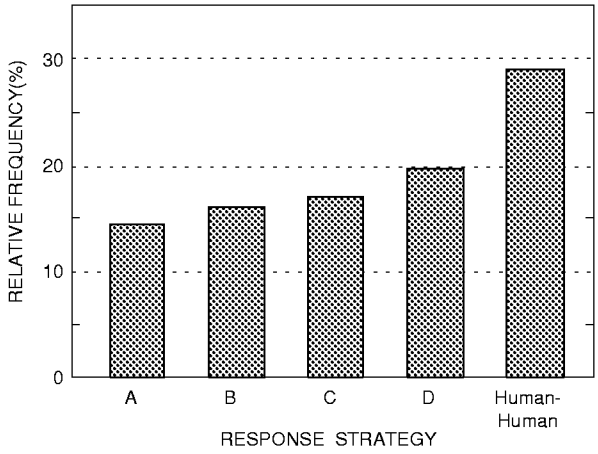

Figure 5: Interjection rates in users' UUs

system. It can then be argued that the system's behaviour in back-channel feedbacks prompts similar responses by users and that, given the 6.8% rate obtained for human-human dialogues, the system's performance is very encouraging.

5. USER SATISFACTION

After each dialogue session, users were asked to evaluate the system by answering a questionnaire on dialogue fluency and user satisfaction, which were to be graded from 1 to 5. The results shown in Figure 7 indicate that strategy D fares the best overall, in agreement with our analysis. As far as fluency is concerned, strategies A and D fare slightly better than strategies B and C.

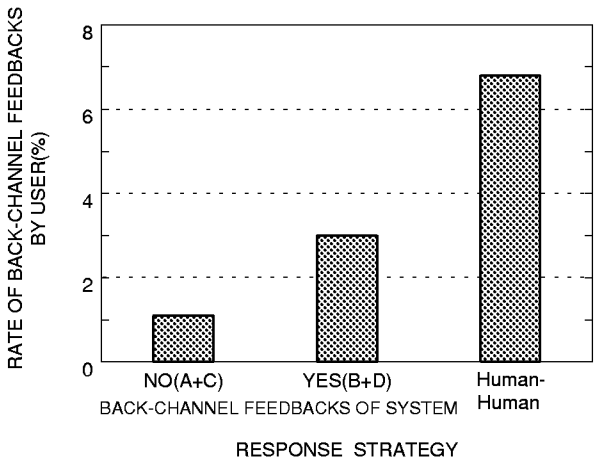

Figure 6: Rate of back-channel feedbacks by users

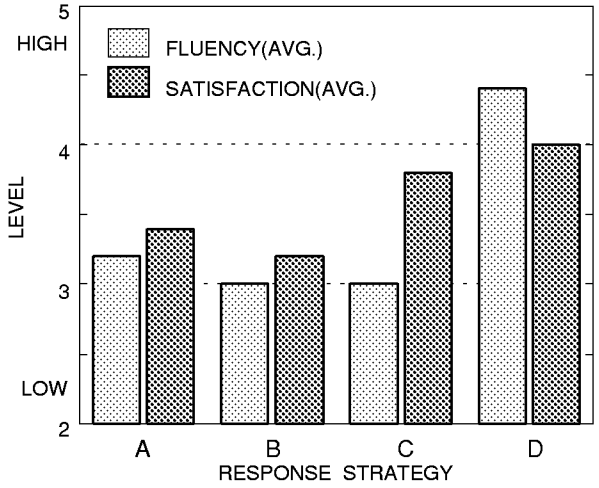

Figure 7: Subjective system evaluation

6. CONCLUDING DISCUSSION

In this paper we investigated how system utterances influence users using the WOZ method. Our results first show that back-channel feedbacks and brief confirmations by our spoken dialogue system, prompt more utterances from and give more satisfaction to users. However, this spontaneity effect tends to increase the occurrence of fillers and back-channel feedbacks in users' utterances. As a result, the dialogue system is forced to handle users' spontaneous speeh, so long as the system's response strategy is itself endowed with the ability of generating brief utterances and back-channel feedbacks.

Second, users tend to give back-channel feedbacks whenever the system gives them feedbacks. It follows that users are able to gauge the system's ability to handle interjections and to predict its behaviour. An intriguing question arises regarding users' behaviours with no preliminary instructions at all. Beyond this study, future work is planned to further evaluate our response strategy using a more automated, spoken dialogue system which has no WOZ operator.

7. REFERENCES

[1] Larsen, L. B. "A Strategy for Mixed-Initiative Dialogue Control", *Eurospeech 97*, Rhodes, Greece, pp. 1331-1334, 1997.

[2] Pieraccini, R., Levin, E. and Eckert, W. "AMICA: the AT&T Mixed Initiative Conversational Architecture", *Eurospeech 97*, Rhodes, Greece, pp. pp. 1875–1878, 1997.

[3] Levin, E. , Pieraccini, R. and Eckert, W. "Learning Dialogue Strategies within the Markov Decision Process Framework", *IEEE ASRU97 Workshop*, pp.72–79, 1997.

[4] Niimi, Y., Nishimoto, T. and Kobayashi, Y. "Analysis of Interactive Strategy to Recover from Misrecognition of Utteances Including Multiple Information Items", *Eurospeech 97*, Rhodes, Greece, pp.2251–2254, 1997.

[5] Itou, K., Akiba, T., Hasegawa, O., Hayamizu, S. and Tanaka, K. "Collecting and Analyzing Nonverbal Elements for Maintenance of Dialog Using a Wizard of OZ Simulation", *Proc. ICSLP-94*, Yokohama, Japan, pp. 907–910, 1994.

[6] Marcus, S. M., Brown, D. W., Goldberg, R. G., Schoeffer, M. S., Weizel,W. R. and Rosinski, R. R. "Prompt Constrained Natural Language-Evolving the Next Generation of Telephony Services", *Proc. ICSLP-96*, Philadelphia, U.S.A, pp. 857–860, 1996.

[7] Okato, Y. , Kato, K., Yamamoto, M. and Itahashi, S. "Insertion of Interjectory Response Based on Prosodic Information", IVTTA-96, pp. 85–88, New Jersey, U.S.A., 1996.

[8] Fraser,N. M. and Gilbert, G. N. "Simulating speech systems", *Computer Speech and Language*, 5(1): pp. 81-99, 1991.

[9] Akiba, T., Kamishima, T. and Itou, K. "MILES: Multimodal Interaction LEading Script, which can express time relations between events and communicative elements in dialogues", *IEICE Tech. report*(SP97-86),1997 (in Japanese).

[10] Murakami, J. and Sagayama, S. "A Discussion of Acoustic Problems in Sponaneous Spech Recognition", *IEICE Trans. Electronics Information and Communication* D-II, Vol. J78-D-II No.12 pp.1741–1749, Dec., 1995 (in Japanese).