SYSTEM-USER INTERACTION AND RESPONSE STRATEGY IN SPOKEN DIALOGUE SYSTEM

Yohei Okato          Keiji Kato          Mikio Yamamoto          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-
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.

**Figure 1:** Example of a “keyword”

**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

### 4. DIALOGUE ANALYSIS

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