1-1-2004

Acknowledgment Use with Synthesized and Recorded Prompts

Karen Ward
University of Texas at El Paso, kward@cs.utep.edu

Tasha Hollingsed
University of Texas at El Paso, tasha@cs.utep.edu

Javier A. Aldaz Salmon
University of Texas at El Paso, jaldaz@cs.utep.edu

Published in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004* pp. 85-88.

Acknowledgment Use with Synthesized and Recorded Prompts

Karen Ward, Tasha Hollingsed, Javier A. Aldaz Salmon
The University of Texas at El Paso
El Paso, Texas USA 79968
{kward,tasha,jaldaz}@cs.utep.edu

Abstract

Acknowledgments, e.g., "yeah" and "uh-huh," are ubiquitous in human conversation but are rarer in human-computer interaction. What interface factors might contribute to this difference? Using a simple spoken-language interface that responded to acknowledgments, we compared subjects' use of acknowledgments when the interface used recorded speech with that seen when the interface used synthesized speech. Contrary to our hypothesis, we saw a drop in the numbers of subjects using acknowledgments: subjects appeared to interpret the recorded-voice interface as signalling a more limited interface. These results were consistent for both Mexican Spanish and American English versions of the interface.

1 Introduction

In previous studies, we showed that subjects use acknowledgments and politeness words when interacting with a simple spoken-language application even when the interface does not offer such behaviors itself (Ward and Heeman, 2000; Ward et al., 2003). In post-experiment interviews conducted as part of that study, 50% of the subjects (11 in the English-language condition, 9 in the Spanish) had thought that they might be more likely to use acknowledgments if the interface had a more human-like voice. In this study, we tested that hypothesis: we examined the effect of changing the interface prompts from synthesized speech to recorded speech.

The term "acknowledgment" is from Clark and Schaefer (1989), who describe a hierarchy of methods by which one conversant may signal that another's contribution has been understood well enough to allow the conversation to proceed. Acknowledgments often appear in English as "uh-huh" and in Spanish as "ajá." Acknowledgments, also called "back-channels" by some researchers (e.g., Chu-Carroll and Brown, 1997), are one of several meta-dialogue behaviors that people use to control the flow of conversation.

Meta-dialogue behaviors such as acknowledgment are of interest because of their role in managing turn-taking: although acknowledgments may preface a new contribution by the same speaker (Novick and Sutton, 1994), often they occur alone as a single-phrase turn that appears to serve the purpose of explicitly declining an opportunity to take a turn (Sacks et al., 1974). If acknowledgment behavior is incorporated in spoken-language systems, it may offer a more fluid and adaptable means of managing turn-taking and pacing in human-computer interaction.

Although some research systems incorporate acknowledgments (e.g., Aist, 1998; Iwase and Ward, 1998; Okato et al., 1998), real-world spoken-language interfaces generally don't allow acknowledgments to serve their turn-taking purpose. Turn-taking is completely controlled by one conversant, usually the system. To reduce errors, designers of spoken-language systems create prompts that guide the user toward short, focused, in-vocabulary responses (e.g., Basson et al., 1996; Cole et al., 1997). In many systems, the use of barge-in defeats the common interpretation of an acknowledgment: if the user speaks, the system quits speaking and begins interpreting the user utterance. If the user intended to signal that the system should continue, the effect is exactly the opposite of the one intended. Thus, current design practices both discourage and render meaningless the standard uses of acknowledgments.

2 Experiment

The study design, described below, is identical to that used in our baseline study (Ward et al., 2003) except that the interface prompts and messages were delivered using recorded human voices instead of synthesized voice. These studies were conducted in both American English and Mexican Spanish.

2.1 Method

We did not want to explicitly instruct or require subjects to use acknowledgment behavior, as that would tell us nothing about their preferences. Instead, we wanted to create a situation in which subjects would have a reason to use acknowledgments, perhaps even gain an advantage from doing so, while still keeping the behavior optional. Conversants are likely to offer acknowledgments and repetitions when complex or important information is being transcribed, especially when the cost of making an error may be high. Acknowledgments in this context may serve a dual purpose of conveying understanding and of controlling the pace of the interaction. Furthermore, there may be more verbal acknowledgments offered during telephone-based interaction than during face-to-face interaction (Cohen and Oviatt, 1993). We therefore designed a task in which the subject is asked to make written notes of information presented verbally over the telephone.

We selected the domain of a telephone interface to E-mail. Subjects were told that the computer system would read E-mail messages to them over the telephone and that their task was to locate and transcribe particular items of information contained in the messages, e.g., "How do you get to the coffee house?" The messages included both "interesting" information that was to be copied and "uninteresting" information that was not, so that subjects would want to move through the "uninteresting" material more quickly. In this way we hoped to motivate subjects to try to control the pace at which information was presented.

The E-mail was presented in segments roughly corresponding to a long phrase, with each segment followed by a pause of about five seconds. Five seconds is a long response time, uncomfortably so for human conversation, so we hoped that this lengthy pause would encourage the subjects to take the initiative in controlling the pace of the interaction. If the subject said nothing, the system would continue by presenting the next message segment. Subjects could reduce this delay by acknowledging the contribution, e.g., "okay," or by commanding the system to continue, e.g., "go on" or "continuar." The system signalled the possibility of controlling the delay by asking the subject the question "Are you ready to go on" or "Estas listo(a) para continuar" after the first pause. This prompting was repeated for every third pause in which the subject said nothing. In this way we hoped to suggest to the subjects that they could control the wait time without explicitly telling them to do so.

On the surface, there is no functional difference in system behavior between a subject's use of a command to move the system onward (e.g., "go on," "next," "continue") and the use of an acknowledgment. In either case, the system responds by presenting the next message segment, and in fact it eventually presents the next segment even if the subject says nothing at all. Thus, the design allows the subject to choose freely between accepting the system's pace, or commanding the system to continue, or acknowledging the presentations in a fashion more typical of human conversation. In this way, we hoped to understand how the subject preferred to interact with the computer.

Subjects were told that the study's purpose was to assess the understandability and usability of the interface, and that their task was to find the answers to a list of questions. They were given no instructions in the use of the program beyond the information that they were to talk to it using normal, everyday speech.

We tested a total of 40 subjects, balanced for gender and language. Subjects were solicited from the University of Texas at El Paso campus. They ranged in age from 18 to 65, with most being between 20 and 25. Each subject was paid $10.00 for participating in the study.

We used a Wizard of Oz protocol as a way to allow the system to respond to acknowledgments and to provide robustness in handling repetitions. The wizard's interface was constructed using the Rapid Application Developer in the Center for Spoken Language Understanding Toolkit (Sutton et al., 1998). A simple button panel allowed the wizard to select the appropriate response from the actions supported by the application. The application functionality was limited to suggest realistic abilities for a current spoken-language interface. The subject could request a message by message number, for example, but not by content or sender.

The interface prompts and messages were presented using recorded human voices. The message texts were presented in a male voice, and the control portions of the interface were in a female voice. It was hoped that the two voices would help the subjects determine the state of the interface: delivering message text vs. controlling the interface functions.

2.2 Measures

In comparing the strategies used to control the length of the pauses (acknowledgment or command use or none), the dependent variable was the number of times each strategy was used to control the pacing of the interface. The total number of turns varied between subjects because some subjects listened to each message only once while others went through messages multiple times. We therefore normalized the counts by dividing the number of times each strategy was used by the number of turns where the subject had had an choice of strategies. We considered the possibility that subjects who completed the task in only one pass though the messages might show a preference for a different strategy than

those who required multiple passes through the messages, thus creating a bias in the normalized statistic. A preliminary analysis showed no significant difference, so we did not consider this possibility further.

The determination as to whether a particular utterance constituted an acknowledgment or a command was based primarily on word choice and dialogue context; this approach is consistent with definitions of acknowledgment, e.g., (Chu-Carroll and Brown, 1997). Immediately following a system inform (presentation of a segment of an E-mail message), the words "yes," "sí," "uh-huh," "ajá," and "okay" or a repetition of part or all of the system inform were considered acknowledgments. Phrases such as "go on," "continue," "next," "continuar," or "siguiente" following an inform were considered commands. The interpretation was confirmed during the post-experiment interview by questioning the subjects about their word choice.Transcriptions and categorizations of the subject utterances were checked by a second person for accuracy.

Some subjects (one in the Spanish-language condition and eight in the English-language condition) combined acknowledgments and commands in a single utterance, e.g., "okay, go on." If an acknowledgment was the first part of the phrase, then it was included in the analysis as an acknowledgment and if a command was the first part, then it was included as a command. Most subjects did this only once (the single subject in the Spanish-language condition and three of the eight in the English-language condition), and one speaker (English) produced as many as six combined-type responses.

A post-experiment interview was conducted to determine each subject's impression of the system. Several of the questions were drawn from the PARADISE model (Walker et al. 2000). The experimenter also explained the true purpose of the experiment and answered subjects' questions. This interview was taped and the experimenter took notes. Data from subjects who had realized that they were interacting with a human instead of a completely-automated system were excluded from the study because of the well-verified tendency for people to speak differently when they believe that they are speaking with a human instead of a computer (e.g., Brennan, 1991).

3 Results

We hypothesized that subjects would use acknowledgment behaviors to control the recorded-voice version of the interface than they did with the synthesized-voice version. We expected this increase to be seen in both Spanish and English conditions and across both female and male speakers. The results were contrary to our expectations.

When interacting with the recorded-voice interface, commands and acknowledgments were preferred as a strategy by 15% and 17.5%, respectively, of all subjects. This result was not significantly different than that seen in the synthesized-voice study, as confirmed by the Wilcoxen-Mann-Whitney test ($z = -0.5041$, $p = 0.0139$ for commands, $z = 1.686$, $p = 0.0465$ for acknowledgments).

Contrary to our expectations, the numbers of subjects using either acknowledgments and commands actually dropped. This was due to the fact that the numbers of subjects who used waiting as their sole strategy rose sharply, from 9 subjects in the synthesized-voice study to 19 in the recorded-voice study ($\chi^2 = 14.34$, $p<0.001$).

Forty percent of the subjects used a command at least once, and 45% used an acknowledgement at least once. Seven subjects seemed comfortable with both commands and acknowledgments, using at least five examples of each. When acknowledgments were used, the most common word choice was "okay" (both languages). When commands were used, the most common word choices were "go on" in English, and "continuar" in Spanish.

We found no significant difference between the recorded and synthesized-voice conditions when comparing male and female speakers nor when comparing English and Spanish speakers.

Politeness behaviors were common. These included the use of the phrases "thank you" or "gracias" and "please" or "por favor" as well as a responding "goodbye" or "adiós" to the system. Many subjects (7 Spanish-language and 8 English-language, 37.5% total) used a politeness behavior at least once and a few subjects (1 Spanish-language and 5 English-language, 15% total) used them more than once. One English-speaking female used politeness behaviors with almost half of her interactions with the system. One subject, when asked in the post-experiment interview why he chose to use this behavior, responded "I don't know, it's just habit I guess." Three other subjects made similar statements.

We believe, and some subjects confirmed, that some subjects in the recorded version assumed that they were listening to recordings similar to voice-mail messages on their telephones. They believed that the pauses were part of the message and so did not realize that the system was awaiting their response.

4 Conclusions

We compared subjects' use of various strategies for controlling the pacing of information presentation in a simple spoken-language interface using synthetic speech with one using recorded speech. We had hypothesized that subjects would offer more acknowledgments in the recorded-voice condition. In fact, we saw no differences

in the numbers of subjects using acknowledgment as a preferred strategy. We also saw a significant increase in the number of subjects who made no attempt to control the pacing of information presentation at all. We conclude that, in this case, use of a human voice in the interface misled subjects into assuming a more limited capability based on their previous experience with existing technology. In future work, we plan to move to a richer domain that will support a more complex interaction, one in which the system will have more opportunities to signal its interactive capabilities to the user.

Acknowledgments

This work was partially supported by a gift from Microsoft Corporation and by the National Science Foundation's Model Institutions for Excellence Initiative EEC 9550502. The authors thank David Herrera, Christian Servin, Tyler Smith, and Pauline Williamson for their assistance with the study. We also thank the anonymous reviewers for their helpful comments and suggestions.

References

Gregory Aist. 1998. "Expanding a Time-Sensitive Conversational Architecture for Turn-Taking to Handle Content-Driven Interruption," in *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 413-417.

Sara Basson, Stephen Springer, Cynthia Fong, Hong Leung, Ed Man, Michele Olson, John Pitrelli, Ranvir Singh, and Suk Wong. 1996. "User Participation and Compliance in Speech Automated Telecommunications Applications," in *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, 1676-1679.

Susan E. Brennan. 1991. "Conversation With and Through Computers," *User Modeling and User-Adapted Interaction*, 1:67-86.

Jennifer Chu-Carroll and Michael K. Brown. 1997. "Tracking Initiative in Collaborative Dialogue Interactions," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 262-270.

Herbert H. Clark and Edward F. Schaefer. 1989. "Contributing to Discourse," *Cognitive Science*, 13:259-294.

Ron A. Cole, David G. Novick, P.J.E. Vermeulen, Stephen Sutton, Mark Fanty, L.F.A. Wessels, Jacques de Villiers, J. Schalkwyk, Brian Hansen and D. Burnett. 1997. "Experiments with a Spoken Dialogue System for Taking the U.S. Census," *Speech Communication*s, Vol. 23.

Phillip Cohen and Sharon Oviatt. 1994. "The role of voice in human-machine communication," *Voice Communication between Humans and Machines* (ed. by D. Roe and J. Wilpon), National Academy of Sciences Press, Washington, D. C., Ch. 3, 34-75.

Tatsuya Iwase and Nigel Ward. 1998. "Pacing Spoken Directions to Suit the Listener," in *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, Vol. 4, 1203-1207.

David G. Novick and Stephen Sutton. 1994. "An Empirical Model of Acknowledgment for Spoken-Language Systems," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 96-101.

Y. Okato, K. Kato, M. Yamamoto and S. Itahashi. 1998. "System-User Interaction and Response Strategy in Spoken Dialogue System," in *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, Vol. 2, 495-498.

H. Sacks, E. Schegloff and G. Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-Taking in Conversation," *Language*, 50:696-735.

Stephen Sutton, Ron Cole, Jacques de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E.Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro and M. Cohen. 1998."Universal Speech Tools: the CSLU Toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 3221-3224.

Marilyn A. Walker, Candace A. Kamm and Diane J. Litman. 2000. "Towards Developing General Models of Usability with PARADISE," Natural Language Engineering.

Karen Ward and Peter A. Heeman. 2000. "Acknowledgments in Human-Computer Interaction," in *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, April 29-May 4, 280-287.

Karen Ward, Tasha Hollingsed, and Javier A. Aldaz Salmon. 2003. "Toward Building Conversational Spoken-Language Interfaces: Acknowledgment Use in American English and Mexican Spanish," *Proceedings of the Fourth Mexican International Conference on Computer Science*, September 10-12, 10-17.