

---

# The Effect of Miscommunication Rate on User Response Preferences

**Hua Ai**

Intelligent Systems Program  
University of Pittsburgh  
5113 Sennott Square  
Pittsburgh, PA 15260  
[hua@cs.pitt.edu](mailto:hua@cs.pitt.edu)

**Thomas Harris**

Language Technologies Institute  
Carnegie Mellon University  
5000 Forbes Avenue  
Pittsburgh, PA 15213 USA  
[Thomas.harris@cs.cmu.edu](mailto:Thomas.harris@cs.cmu.edu)

**Carolyn Penstein Rosé**

Language Technologies Institute/  
HCI Institute  
Carnegie Mellon University  
5000 Forbes Avenue  
Pittsburgh, PA 15213 USA  
[cprose@cs.cmu.edu](mailto:cprose@cs.cmu.edu)

---

Copyright is held by the author/owner(s).  
CHI 2006, April 22–27, 2006, Montréal, Québec, Canada.  
ACM 1-1-59593-298-4/06/0004.

**Abstract**

We report results from a small Wizard-of-Oz study investigating user responses to miscommunications in speech dialogue systems. We explore the separate and joint effects of miscommunication rate and system response to miscommunications on the likelihood that users choose to resort to direct manipulation, to repeat, or to rephrase. While we predicted that users would be more likely to resort to direct manipulation as miscommunication rate increased, our surprising finding was that users were most likely to resort to direct manipulation where communication success was least predictable, i.e., in the middle of the range, rather than at either extreme.

**Keywords**

Multimodal system, dialogue systems, error recovery

**ACM Classification Keywords**

H5.2. User interfaces: Interaction styles.

**Introduction**

The concept of a “Smart home”, which began with the idea of home networking and security systems is growing now to include voice controlled appliances, self-adapting environmental controls, and other technologies for home automation. Computers, mobile

phones, and cameras are increasingly commonplace, often working in a coordinated manner. Such an environment, characterized by a high level of connectivity between devices and the ability to maintain continuous interaction with users even across multiple devices, is called an Intelligent Environment [2]. Since in an Intelligent Environment users may be interacting with multiple devices in parallel, a multimodal interface is a natural choice. Speech, direct manipulation, gesture, and remote control are common examples of input modalities. Among these, speech input is a promising option for the home. Brumitt et al. [1] present empirical evidence of the desirability of speech interaction from a usability standpoint. Among its commonly cited advantages over alternatives such as direct manipulation are its naturalness and its potential to enable hands and eyes free interaction.

Although the research area of speech recognition has made great strides in the past decade, the most serious usability issue with speech dialogue systems remains speech recognition errors and the resulting miscommunications. The undesirable error rate and the high cost of error resolution degrade system performance, frustrate users, and limit serious commercial potential. Although the error rate of research systems after being trained on the voices of the users can achieve a word error rate as low as 10% [6, 7], this rate doubles if the system is not trained with those specific users [4], and is even worse in field settings than in the lab [3, 9]. Note that a word error rate of 10% means that 1 in 10 words is incorrectly recognized. If typical utterances are 2 or 3 words long, then every 3 to 5 utterances may be incorrectly recognized. However, although training with target users can significantly improve the system's

performance, this approach typically places unrealistic burdens on users, leaving users unlikely to choose to interact with home appliances through speech rather than through direct manipulation. Without providing users with a way to overcome this hurdle, users will not have the opportunity to experience the advantages of speech based interaction discussed above.

Our proposed solution is to provide a natural mechanism whereby users can train the system during actual use, which requires much less time and effort on their part. Our hypothesis is that when encountering frequent speech recognition errors, it will be faster and less frustrating for users to demonstrate what they were trying to communicate by means of direct manipulation rather than to engage users in a lengthy error recovery dialogue. We expect that this approach will more quickly allow users to obtain their desired response from appliances while providing the system with training data about how they express their preferences, and thus eventually leading to improved speech recognition. As pointed out by Shin et al. [8] when they analyzed the users' behaviors under error conditions in spoken dialogs, the user's behavior became less amenable to accurate speech recognition as the length of the error recovery dialogue increased. Thus, we expect that this approach will be more productive. Moreover, we expect it to be a comfortable "fallback" interaction. In support of this, Margono et al. [5] show that direct manipulation can make users feel that they are in control, which increases the user's satisfaction and reduces frustration.

An effective design for a "Show and Tell" approach to training would balance the benefits of pure speech based interaction with the benefits of direct

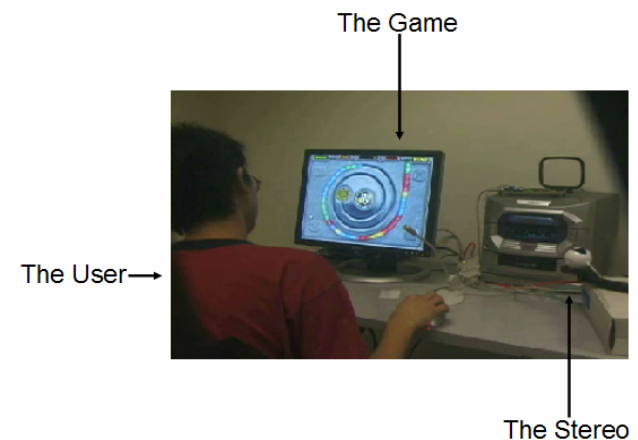
manipulation. In order to strike this balance, it is necessary to determine at what point users cease to feel that they are making sufficient progress in their speech interaction and chose to make a modality switch from speech to direct manipulation, when offered both of these options. One way of using this understanding to drive design would be for the system to take the initiative to suggest a modality shift just before the point where a lengthy recovery dialogue would have likely lead to a user taking that course of action out of frustration. The goal is for this approach to reduce frustration long enough for the system to be trained effectively during use. In order to implement the "Show and Tell" approach, we must influence whether people choose to continue with speech or resort to direct manipulation. As a first step towards accomplishing our goal, we evaluated the effect of system response to miscommunication and miscommunication rate on the likelihood that users choose to resort to direct manipulation, to repeat, or to rephrase, both in the case where the system requests a repeat and in the case where the system requests a direct manipulation.

### Method

*Experimental Setting.* The experiment was conducted in a Wizard-Of-Oz setting where users were required to complete 10 tasks with a shelf stereo while playing an eyes-and-hands busy video game named Zuma®. See Figure 1. Subjects were told that the stereo is an intelligent stereo that can accept commands either via direct manipulation or through speech commands. The subjects were not instructed how they should talk to the stereo, and no vocabulary list was given. Thus, users were free to behave naturally with the system and were free to choose speech based interaction or direct manipulation at any time.

*Experimental Procedure.* An experimenter remained in the same room with the user and read the instructions for each task out loud one by one. The tasks included things like "Can you turn the stereo on?", "Can you skip over this song?", "Can you play the Sonic Youth CD?", or "Turn it off, please."

Subjects were required to complete each task within 2 minutes while staying "alive" and active in the video game. After each task, the game was restarted to ensure that the game behaved consistently, in a distracting manner, across all tasks and users. Users were not allowed to pause the game. A wizard in a different room acted as the speech interaction component of the stereo system. System feedback, which consisted of pre-recorded canned messages, was selected by the Wizard and played for the user as though the system itself was responding.



**Figure 1.** Experimental setting

*Participants.* 9 subjects, 4 females and 5 males, aged from 18 to 33, were recruited from Carnegie Mellon University and assigned randomly to conditions.

*Experimental Design.* Using a 3X2 factorial design, we explore the separate and joint effects of miscommunication rate and system response to miscommunications on user behavior. We used three different levels of simulated miscommunication rates (MR) as a between subjects factor (30%, 50%, 70%), where 30% signifies that a randomly selected 30% of the users' utterances would be treated as not understandable by the system. We used two different system responses to miscommunications (SRM) as a within subjects factor, specifically one where the system asks for a repeat (askRepeat) and one where the system asks the user to use direct manipulation (showMe).

To control for ordering effects, we constructed three random sequences with half as many requests for direct manipulations as requests for repeats. We made requests for direct manipulation less frequent because we expected them to more quickly lead to the conclusion of the interaction with the user. The sequence of system responses that each user experienced was determined by a combination of which of three response sequences they were assigned to (which determined the sequence of system responses to miscommunications they would see) and which of three miscommunication rates they were assigned to (which determined how frequently their utterances were treated as miscommunications). We use the video game as a source of distraction to provide users with a motivation for using speech as a modality. The simulated speech recognition errors provide a

competing motivation to select direct manipulation. We hypothesized that as miscommunication rate increased, users would more quickly choose to use direct manipulation over speech. Similarly, we predicted that users would be more likely to cooperate with the system's request after a miscommunication in a low miscommunication rate condition than in higher ones.

### Collected Corpus and Coding Scheme

To quantify differences in user behavior, we coded videos of the participation of the 9 participants. We coded each user utterance as well as each instance where a user chose to use direct manipulation:

- **DM** - the user chose direct manipulation
- **Rephrase** – user repeated the utterance using different words than before
- **Repeat** – user repeated the utterance using the exact same words as before

Tags used to code system prompts corresponded to the pre-recorded prompts used:

- **ok** – “OK. I will do that for you”
- **askRepeat** – “I am sorry, I cannot understand you. Could you repeat that?”
- **showMe** – “I am sorry. I do not think I can understand that. Could you show me what you mean? ”
- **done** – “OK, I have done the task for you.”

Using this coding we constructed a matrix in which we recorded an observation for each simulated miscommunication, noting the user number, task number, miscommunication number, system response type (askRepeat or showMe), and 4 binary flags corresponding to the three user behavior tags

mentioned above, plus an additional one that indicated that a user continued with speech interaction either through repeating or through rephrasing. Each binary flag was coded as true if the associated behavior was observed at least once in between the system's response to a miscommunication and the next one. There were a total of 95 miscommunications spread over the 9 sessions.

### Results and Conclusions

As mentioned, in order to implement the "Show and Tell" approach, we must influence whether or not people choose to continue with speech. We took a simplistic approach in our manipulation of System Response to Miscommunication (SRM) where we either requested that the user repeat (askRepeat) or requested that the user show the system what was meant through direct manipulation (showMe).

Out of 95 miscommunication episodes, there were 32 showMe episodes and 63 askRepeat episodes. See Table 1 for counts. Using a binary logistic regression separately for each of the 4 binary user response variables, with SRM as the predictor, we found a significant effect of SRM on the likelihood of Direct Manipulation ( $p < .05$ ) and Repeat ( $p < .05$ ). Overall, users were significantly more likely to continue with speech interaction in the askRepeat condition ( $p < .01$ ). There was no significant interaction with Miscommunication Rate (MR). Thus, while users did not always follow the system's advice, it did have an effect on their choices regardless of MR. However, we consider the conclusion of no interaction with MR to be preliminary because of the small number of data points. Note that none of the 4 binary predictor variables are mutually exclusive.

**Table 1** Effect of SRM on frequency of alternative types of user responses.

SRM	Direct Manipulation	Repeat	Rephrase	Continue Speech
askRepeat	10 (16%)	26 (41%)	34 (54%)	59 (94%)
showMe	11 (33%)	6 (19%)	20 (63%)	23 (72%)

**Table 2** Effect of Miscommunication Rate (MR) on frequency of alternative types of user responses.

Misc. Rate	Direct Manipulation	Repeat	Rephrase	Continue Speech
30%	7 (28%)	10 (40%)	14 (56%)	23 (92%)
50%	9 (36%)	5 (29%)	7 (41%)	11 (65%)
70%	5 (10%)	17 (33%)	33 (65%)	48 (94%)
High Certainty (30% or 70%)	12 (16%)	27 (36%)	47 (62%)	71 (93%)
Low Certainty (50%)	9 (36%)	5 (29%)	7 (41%)	11 (65%)

We predicted that as Miscommunication Rate (MR) increases, users would more frequently resort to direct manipulation out of frustration. However, the surprising finding was that likelihood of choosing Direct Manipulation was more related to uncertainty of communication success, where the condition with 50% miscommunications presents the highest degree of uncertainty. See Table 2. In fact, the condition in which participants were most likely to continue with speech interaction was the 70% MR condition. Out of 95 episodes, 25 were in the 30% MR condition, 17 were in the 50% condition, and 51 in the 70% condition. Using a binary logistic regression separately for each of the 4 binary user response variables, with MR as the predictor, we found a significant effect of High versus Low Certainty on the likelihood of Direct Manipulation ( $p < .005$ ) and likelihood of continuing with speech ( $p < .01$ ). Thus, users in the Low Certainty condition (50% MR) were significantly more likely to resort to direct manipulation and significantly less likely to continue with speech interaction than users in the High Certainty condition (30% MR and 70% MR).

While the results from this small user study are just a first step towards the design and implementation of an effective "Show and Tell" recovery strategy, the results are interesting in that they suggest that the condition that is most important to avoid placing users of speech dialogue systems into is one in which the system's behavior is unpredictable to them. The main effect of SRM suggests that a simple approach to directing users to interact with speech enabled appliances in one modality versus another can be effective, although the results show that there is room for improvement.

## References

- [1] Brumitt, B., Cadiz, J.J. Let There Be Light: Examining Interfaces for Homes of the Future. In *Proc. Interact '01*, IOS Press (2002), 375-382.
- [2] Cohen, M. Design principles for intelligent environments. In *Proc. AAAI Symp. on IE 1998*, AAAI Press (1998), 36-43.
- [3] Karis, D., Dobroth, K. M. Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE J. of Selected Areas in Communications* 9, 4 (1991), 574-585.
- [4] Keizer, G. 1998. *The gift of gab: CNET compares the top speech recognition apps.*  
<http://204.162.80.182/Content/Reviews/Compare/Speech/>.
- [5] Margono, S., Schneiderman, B. A study of file manipulation by novices using commands versus direct manipulation. In *Proc. 26th Annual Technical Symp.*, ACM Press (1987), 154-159.
- [6] Robinson, T., Christie, J. Time-first search for large vocabulary speech recognition. In *Proc. of ICASSP* (1998).
- [7] Renals, S., Hochberg, M. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *Proc. ICSLP* (1996), 149-152.
- [8] Shin, J., Narayanan, S., Gerber, L., Kazemzadeh, A., Byrd, D. Analysis of user behavior under error conditions in spoken dialog. In *Proc. ICSLP* (2002), 2069 – 2072.
- [9] Spitz, J., Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proc. of the 4th Darpa Workshop on SNL*, Morgan Kaufmann Publishers (1991), 19-22.