



Constraining User Response via Multimodal Dialog Interface

KIRK BAKER

Linguistics Department, The Ohio State University, 222 Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210-1298

kbaker@ling.ohio-state.edu

ASHLEY MCKENZIE AND ALAN BIERMANN

Department of Computer Science, Duke University, P.O. Box 90129, Durham, NC 27708-0129

GERT WEBELHUTH

Seminar für Englische Philologie, Georg-August-Universität Göttingen, Käte-Hamburger-Weg 3, 37073 Göttingen

Abstract. This paper presents the results of an experiment comparing two different designs of an automated dialog interface. We compare a multimodal design utilizing text displays coordinated with spoken prompts to a voice-only version of the same application. Our results show that the text-coordinated version is more efficient in terms of word recognition and number of out-of-grammar responses, and is equal to the voice-only version in terms of user satisfaction. We argue that this type of multimodal dialog interface effectively constrains user response to allow for better speech recognition without increasing cognitive load or compromising the naturalness of the interaction.

Keywords: constrain user response, multimodal dialog interface, speech recognition

1. Introduction

Accurate interpretation of acoustic input remains a performance-limiting factor for speaker-independent, continuous speech recognition applications. Processing difficulties inherent to the underlying variability of the signal are exacerbated in large vocabulary, high perplexity domains. Traditional speech interface design strategies aimed at achieving an acceptable measure of recognition performance under these conditions often rely on constraining user responses via list or menu-based interactions, explicit instructions for answering prompts and frequent response confirmation (Boyce, 1999; Gardner-Bonneau, 1999; Novick et al., 1999; Boyce, 2000). These methods ultimately succeed in reducing the likelihood of recognition errors occurring or persisting undetected, but do so at the expense of natural and efficient dialog flow. While judicious prompt design alleviates to some extent the characteristic tedium of such system-oriented approaches,

user frustration frequently leads to speech behavior that further reduces intelligibility (Baca, 1998). Furthermore, because users are forced to remember sets of speaking directions orthogonal to the task at hand, these systems increase cognitive load and are unnecessarily dependent upon short-term memory (Grasso and Finin, 1997; Shneiderman, 1997; Balentine, 1999).

This paper addresses the issue of how to influence user response for better system performance without compromising the system-user interaction in terms of its naturalness or short-term memory requirements. Baddeley's (1992) tripartite model of short-term memory holds that separate neurological components simultaneously store and process phonological and visual information, while a third controls attention. Working within this theoretical framework, a number of previous studies investigating the effects of information processing on cognitive load demonstrated that the concurrent presentation of information via distinct

modalities does not adversely affect performance on cognitively complex tasks (see, e.g., Mousavi et al., 1995; Mayer et al., 1999; Mayer and Moreno, 1998; Goolkasian, 2000). Conversely, split-attention scenarios, which require attending to disparate sets of information accessed via a single modality, are associated with decreased task performance (e.g., David and Hirshman, 1998; Yeung, 1999; Velayo and Quirk, 2000). With this framework and results in mind, we decided to approach the problem of processing information related to the speech interface in addition to processing information required by the task itself as precisely the type of cognitively complex, two-in-one scenario amenable to multimodal presentation.

The basic idea underlying our application involves utilizing the multiple information channels enabled by broadband connectivity to split the speech recognition interface into separate modalities: the meta-information constraining users' linguistic behavior may be presented visually in a device-appropriate manner (e.g., cell phone display, PDA, voice-enhanced web access), while the task itself is conducted via voice interaction. The visual display may serve to inform users of the system's capabilities, whether as a menu-style presentation of options to open-ended prompts such as "How may I help you?", or a restricted list of valid responses to a particular question. It can also be used to display information obtained from a user as a dialog progresses. Because users have access to immediate visual feedback, the need for verbal confirmation of responses is obviated; similarly, time-consuming response instructions or explanation of system capabilities can be encapsulated textually and visually processed by the user while they are listening to a voice prompt. A limited recognition vocabulary coordinated to the text display can be employed to ensure higher recognition rates, and because the display persists at least until the user responds to a given prompt, he/she is not required to keep lists of response options in memory.

People adapt their conversational style automatically to their model of a listener's capabilities (Baber et al., 1997; Walker et al., 1998; Balentine, 1999), so we expect that making the system's recognition abilities explicit and maintaining a naturally-structured dialog will simultaneously improve system performance and reduce user frustration. However, there is no *a priori* guarantee that users' responses will conform to the format suggested by the text display; in this case a limited recognition grammar would result in a greater percent-

age of out-of-grammar responses, and usability would decrease.

In order to test the hypothesis that a multimodal dialog interface can reduce cognitive load effectively and simultaneously constrain user responses for improved speech recognition, we designed a simple dialog application that optionally utilizes text displays coordinated with the speech interface and conducted an experiment with two groups of users. The remainder of this paper covers the setup and results of that experiment. In Section 2, we describe the application in detail, beginning with an overview of the domain and continuing with a presentation of the system design. Section 3 reports the results of an experiment comparing user behavior and system performance between the bimodal and speech-only interfaces to the dialog application. Finally, in Section 4, we discuss the results of our experiment in the broader context of natural language computer-human interfaces.

2. Application

We started the design process by analyzing 650 transcribed dialogs supplied by the UK-based GE Consumer Finance's customer service call center as a member of the AMITIÉS (Automated Multilingual Interaction with Information and Services) consortium. Each dialog consisted of a conversation between a customer, calling to resolve some issue regarding their credit card account, and a financial services representative. Typically, customers wished to complete such tasks as making a debit card payment on their account, changing their address, reporting a lost or stolen card, or various other account inquiries. We analyzed the dialogs in terms of the data keys needed to complete each transaction (e.g., things like name, address, phone number, date of birth), vocabulary, the grammatical structure of utterances, and the dialog structure of different tasks (for a more detailed description of the analysis procedure see Hardy et al., 2002).

A typical dialog began with the customer service representative greeting the caller, who stated the reason for his/her call. After verifying the caller's identity, the dialog moved on to completing the primary task. Finally, the representative wrapped things up, thanked the caller, and the conversation was complete. Figure 1 illustrates a sample dialog for a customer calling to inform the representative of an address change.

Greet Caller	REP: Good afternoon customer services, you're speaking to Sam, how can I help? CUST: Hi, do you want me to give you my new address?
Verify Id	REP: Ok, can I take your name? CUST: It's Jane Smith REP: And your postcode and telephone number of your old address? CUST: A1 1CD and 1111 111 1111
Task	REP: Thank you and your new house number? CUST: 85 The Lane REP: And the postcode? CUST: Err N, oh sorry, A1 err 1EF REP: The Lane did you say? CUST: Yeah REP: And is there a new telephone number? CUST: Err yeah no
Task Summary	REP: Ok, that's been updated for you to erm 85 The Lane is there anything else I can help you with? CUST: No that's it thanks
Close	REP: Thanks for calling then CUST: Bye REP: Goodbye

Figure 1. Sample dialog for address change.

Our application was designed to handle two of the transactions found in the natural dialogs: change of address and lost card. Dialog flow was controlled via a set of dialog states corresponding to the observed task structure for the two topics. Figure 2 illustrates these states and transitions.

Each state included a pre-recorded voice prompt to indicate to the user the information they should provide. The verbal prompts were designed to be fairly open-ended, allowing for a mixed-initiative conversation similar to the customer-agent interactions in the dialog database. More importantly for our study, the prompts were designed to be unrestricted enough to elicit a range of responses sufficient for a meaningful between-group comparison given our small subject pool.

In the text-coordinated version of the application, each state also included a text cue to prompt the user. As opposed to the verbal prompts, the text cues suggested that the user produce responses constrained to what they saw onscreen. The verbal prompts did not refer explicitly to the text cues; users were expected to be able to integrate the visual information passively into the formulation of their responses. The voice prompt for each dialog state and corresponding text cue are detailed in Table 1.

Each dialog state was also associated with a recognition grammar defining licit utterances for a particular user turn. The grammars served to map user utterances to the particular data key we needed to extract at that

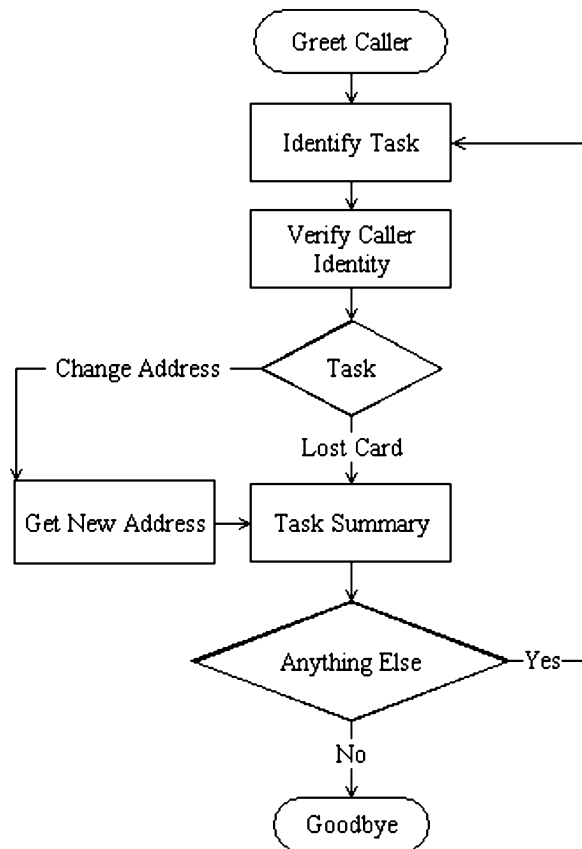


Figure 2. Dialog states and control flow.

Table 1. Spoken prompts and corresponding text cues for dialog states. Each text display was preceded by the words “You say:”

Spoken prompts	Text display
“Hello. This is Fred, Duke’s automated credit card help desk computer. How may I help you?”	Change Address Lost Card Card Balance Credit Limit Charge Denied
“What is your name?”	<Name>
“What is the zip code currently listed on your account?”	<Zip Code>
“May I have your date of birth?”	<Birth Date>
“Can I get the street and street number currently listed on your account?”	<Street Address>
“Can I get your new state?”	<New State>
“Can I get your new city?”	<New City>
“Can I get your new zip code?”	<New Zip Code>
“Can I get your new street and street number?”	<New Street Address>
“Your address has been changed. Is there anything else I can do for you?”	Yes No

point in the dialog. The recognition grammars for the version of our application not utilizing a text display were constructed on the basis of customer replies in the AMITIÉS database. These grammars were designed to incorporate the full range of observed responses and as broad a range of anticipated responses as possible, taking dialectal differences between the British English of the database and the American English of our users into account. The general form of a grammar for a particular prompt consisted of an optional list of filler words (um, uh, yes, etc.), response variants, and any optional markers of politeness (please, thank you, etc.). The grammars for the text-coordinated version accepted only utterances taken verbatim from the visual display (Table 1).

The application was built using the Nuance speech recognition system (v7.0.3). We created a component (The Recognizer Manager) that handled language and dialog processing using the Nuance Application Programming Interface. The Recognizer Manager also processed incoming phone calls and updated the text display of a Java Graphical User Interface on a remote computer via a socket-based network protocol. The voice signal was sent from the modem to the speech recognizer using Open H.323, an open source H.323 implementation.

Finally, we created a database of 20 “customers,” each associated with a first and last name, address

(street and number, city, state and zip code) and birth date (month, day and year). This customer list was linked to the dialog application such that slots in the recognition grammars were tied to the appropriate keys in the database.

Given this basic overview of application architecture, we turn in Section 3 to the description of an experiment that compared recognition performance and user behavior between the text-coordinated and no-text versions of the application.

3. Experiment

3.1. Subjects and Experimental Setup

Twenty undergraduate volunteers enrolled in an introductory linguistics course at UNC Chapel Hill were randomly assigned to evaluate the text-coordinated or no-text version of the application. Each group consisted of 7 female and 3 male participants. Subjects received \$5 for participating. This portion of the experiment was conducted in the phonetics lab at UNC Chapel Hill.

Subjects were seated individually at a desk with a computer monitor and telephone on it, and were given a sheet with a name, account number and address that corresponded to an entry in the customer database. Subjects were asked to assume the given identity while they completed the experiment. Once they dialed into the system, located in the language lab in the Computer Science Department at Duke University, they received pre-recorded instructions over the phone describing a scenario involving their credit card that they had to work through. An experimenter at Duke monitored the dialog and manually selected the next dialog state if recognition remained unsuccessful after three attempts by the system to understand the answer to a given prompt. Each participant made two phone calls corresponding to the scenarios shown in Table 2.

Table 2. Initial prompts to describe experimental scenarios.

Scenario one	“You tried to purchase a new microwave, and your wallet was not where you expected. Now you have looked everywhere and are worried that your credit card might fall into the wrong hands. Would you like for me to repeat the scenario?”
Scenario two	“You have a new job in Austin, Texas and wish to let your credit card agency know. Would you like for me to repeat the scenario?”

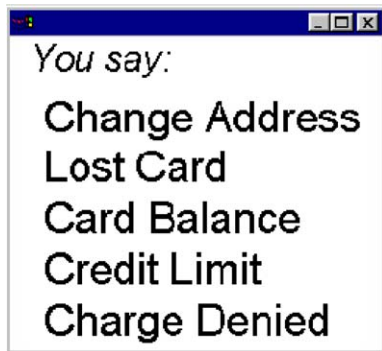


Figure 3. Text display for greeting prompt: "Hello. This is Fred, Duke's automated credit card help desk computer. How may I help you?"

After completing the phone calls, participants filled out a subjective evaluation form. The total procedure took approximately 15 minutes per subject.

The text-coordinated group of subjects was told to watch the computer screen for directions on how to respond to each of the computer's questions as the dialog progressed. The text cues were presented using large black letters in a Java window with a white background centered in a white screen. Figure 3 shows the text cues coordinated to the initial greeting prompt and is representative of the appearance of the text cues for the other prompts as well.

The no-text group was presented a blank white screen and was asked to answer the computer's questions as the dialog progressed.

3.2. Results

The dialogs were recorded as part of the experiment and later transcribed by one of the experimenters. On the basis of this information, we calculated values for a number of variables reflecting both user behavior and system performance between the two versions of the application. Users of the text-display version of the application averaged 2.9 words per turn versus 4.7 words per turn for the users who did not see the text prompt, a difference that a one-factor ANOVA run on a per-utterance word count between the two groups indicates as significant ($F(1,403) = 13.9, p < .0001$). However, the text cues had an even bigger impact on the length of responses to the initial open-ended prompt "How may I help you?" Users without text cues averaged 16.9 words per response vs. 3.5 words for the text-cued

group in answering that prompt, again a significant difference ($F(1,34) = 17.6, p < .0001$). Users without the text prompt tended to provide more scenario-related background for their call, i.e., "Um, I moved to Austin Texas, I have a new job, and I wanted to let you know so I could change the address on my account.", whereas the text-cued group successfully mapped the reason for their call onto the limited set of in-grammar responses indicated onscreen. Out-of-grammar responses (utterances whose syntactic structure was not parsable by our grammar) constituted 6% of the text-cued group's total, versus 13% for the other group. Binary logistic regression on a per utterance out-of-grammar measure between groups indicates that this difference is significant as well ($F(1,403) = 7.06, p < .03$). Most of the text-cued group's out-of-grammar responses consisted of some type of conversation management, i.e., "Um, I'd like to **change** my **address**," versus the type of extra background information that the uncued group tended to provide.

Table 3 contains measures of system performance related to the recognition of user utterances. These variables were calculated on the basis of several standard evaluation metrics (Becchetti and Ricotti, 1999): substitution errors (the recognizer mistakes the actual word for another in its vocabulary), insertion errors (the recognizer hypothesizes a word when none was actually spoken), and deletion errors (the recognizer fails to register a spoken word). As expected given the user behavior noted above, the overall accuracy ($1 - \frac{\text{[#insertions + #deletions + #substitutions]}{\text{#total words}}$) of the text-display version of the application was 30% higher than for the no-text version, presumably a direct consequence of the limited vocabulary and simplistic recognition grammar employed by the former. The results of a one-factor ANOVA run on a per-utterance accuracy

Table 3. Comparison of system performance for the two versions of the application. The text display version is significantly more accurate than the no-text version: $F(1,403) = 19.70, p < .0001$. Accuracy measured by $1 - \frac{\text{[#insertions + #deletions + #substitutions]}{\text{#total words}}$.

	Text display	No text
# Words spoken	512	1081
# Words correctly recognized	270	279
Insertions # (%)	1 (0%)	9 (3%)
Deletions # (%)	226 (44%)	810 (75%)
Substitutions # (%)	21 (4%)	34 (3%)
Accuracy	52%	21%

rate between the two groups indicate that this difference is significant ($F(1,403) = 19.70, p < .0001$). For both versions of the application, deletions accounted for the majority of recognition errors (text-cued: 44%, no-text: 75%).

Although we observed a significantly higher accuracy rate for the text-coordinated version of the application, the actual accuracy rates for both versions (text-cued: 52%, no-text: 21%) are notably below recognition rates commonly reported for tasks of similar complexity and environmental conditions (see, e.g., Martin and Przybocki, 2001). At least part of the explanation for the overall poor recognition results may lie in the fact that the H.323 implementation we used for the experiment dropped packets, causing the voice signal to be degraded. Unfortunately, this fact was not revealed until post-experiment testing, at which point we switched over to a different method of transporting the audio signal to the recognizer.

After completing both calls, users filled out a subjective evaluation asking them to rate their experience with the system based on the seven statements shown in Table 4. Subjects indicated how strongly they agreed or disagreed with each statement by circling a number

Table 4. Average user responses to subjective evaluation form. Scale: 1 (=strongly disagree) to 5 (=strongly agree). Differences between the two groups for each statement are not significant.

Statement	Dep. Var.	Mean	Std. Dev.	Std. error mean
1. I was able to complete the required task.	Text	5.00	.000	.000
	No Text	4.55	1.023	.324
2. I found the system intuitive and easy to use.	Text	4.6	.505	.16
	No Text	4.2	.781	.247
3. The system recovered gracefully from errors.	Text	4.7	.483	.153
	No Text	3.9	1.101	.348
4. The system understood what I said.	Text	4.2	.844	.267
	No Text	3.9	.962	.304
5. The length of the dialog was not excessive.	Text	4.6	.667	.211
	No Text	4.7	.644	.204
6. Given a choice between using this system or waiting five minutes for a human operator, I would choose to use this system.	Text	4.7	.469	.148
	No Text	3.75	1.4	.443
7. The system was easier to use the second time when compared to the first interaction.	Text	4.10	1.101	.348
	No Text	4.10	.994	.314

ranging from 1 (=strongly disagree) to 5 (=strongly agree). Although user responses for the text-cued group were slightly higher on average, an independent samples *t*-test (2-tailed) indicated no significant difference between the responses of the two groups.

4. Discussion

In general, our results are in line with other studies showing that giving users an explicit model of system capabilities improves the quality of the interaction in terms of performance and user satisfaction (e.g., Walker et al., 1998; Karsenty, 2002). We interpret the similar average scores of the two groups to the subjective evaluation favorably, on the assumption that frustration with the conversational requirements imposed by the text cues would be reflected in lower scores. The fact that the scores for the text-cued group were not lower may be taken to indicate that those users did not feel any more constrained by the nature of their interaction with the system than did the un-cued group. Furthermore, we may speculate that had the voice-only version of the application contained lists of valid responses or other explicit speaking instructions, the scores for the uncued group would have been correspondingly lower; however, the obtained results do not speak directly to that claim.

Given that 94% of the text-cued responses were grammatical even within a simplistic recognition grammar suggests that this type of multimodal speech interface represents an effective way to constrain user response. Responses to the open-ended prompt were influenced notably by the text cues, suggesting that naturally structured voice interfaces can make efficient use of people's ability to incorporate visual information into the meta-task parameters of their interaction with the system. This type of dialog interface accommodates user expectations for conversational content of direct pragmatic relevance to the task and retains the system-oriented advantages obtained by explicitly stipulating licit user behavior. However, this type of multimodal design is no more flexible in terms of adapting to unanticipated user behavior than would be a similarly restrictive, single modality dialog interface. We would also expect considerably worse performance for the subset of users consistently not conforming to the text instructions than for a system that employed a more comprehensive grammar. Furthermore, the simplistic keyword matching type of recognition grammar that our particular results support is antithetic to intelligent

dialog applications capable of extracting semantic intent from grammatically complex user utterances.

Introducing a visual component into the dialog interface raises a number of additional issues not explicitly addressed by the current study. Cases requiring variable responses that cannot be represented literally onscreen (i.e., (Name) or (Zip Code) for the user's actual name or zip code) may be problematic for people conditioned to say exactly what they see in a text prompt, although none of the participants in this experiment had any apparent difficulty. A speech interface dependent on a text display would be considerably less accessible to visually impaired users than a voice-only interface. Also, text prompts are a distraction in environments where vision is preoccupied with some other task, and poorly literate users may find such systems unacceptable. We would expect that both of these situations would result in increased cognitive load.

The current study points to several further research questions involving this type of multimodal dialog interface. A comparison between this application and a traditional system in which user response options are presented via voice menus (versus the open-ended style of this application's voice-only prompts) may be instructive and would allow for a more direct assessment of cognitive load. Tests of the application in degraded environments (versus laboratory conditions) could examine how text cues affect system performance and user satisfaction in noisy conditions and altered visibility conditions (changing the size of the display screen or the size of the characters on the screen, for example). Finally, examining the effects of a text display on user experience may also be of interest; we would expect a carefully designed text component to reduce the number of interactions with an interface before users could navigate it quickly.

The results of this study demonstrate that a textual component of a dialog interface may be coordinated with spoken prompts to constrain user responses effectively. This allows for the use of a smaller recognition grammar, which in turn increases the likelihood of accurate speech recognition. Furthermore, taking advantage of bimodal information presentation should ease the amount of cognitive load and unnaturalness associated with voice-only interfaces that similarly constrain user response options. Users of our bimodal dialog application judged their interaction with it as favorably as users of a nonrestrictive, voice-only version did, suggesting that under the appropriate conditions, a multimodal speech interface represents a design that

constrains user response while maintaining natural and efficient dialog flow.

Acknowledgments

Funding for this research was sponsored by Space and Navel Warfare Systems Center (SSC) grant number N66001-01-1-8941. The results of this work are not necessarily endorsed by the U.S. government.

References

- Baber, C., Johnson, G., and Cleaver, D. (1997). Factors affecting users' choice of words in speech-based interaction with public technology. *International Journal of Speech Technology*, 2(1):45–49.
- Baca, J. (1998). Comparing effects of navigational interface modalities on speaker prosodics. *Assets '98, Proceedings of the Third International ACM Conference on Assistive Technologies*. Marina del Rey: ACM, pp. 3–10.
- Baddeley, A. (1992). Working memory. *Science*, 255(5044):556–559.
- Balentine, B. (1999). Re-engineering the speech menu. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston: Kluwer, pp. 205–235.
- Becchetti, C. and Ricotti, L.P. (1999). *Speech Recognition: Theory and C++ Implementation*. West Sussex, England: John Wiley and Sons.
- Boyce, S. (1999). Spoken natural language dialog systems: User interface issues for the future. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston: Kluwer, pp. 37–61.
- Boyce, S. (2000). Natural spoken dialog systems for telephony applications. *Communications of the ACM*, 43(9):29–34.
- David, P. and Hirshman, E. (1998). Dual-mode presentation and its effect on implicit and explicit memory. *American Journal of Psychology*, 111(1):77–88.
- Gardner-Bonneau, D. (1999). Guidelines for speech-enabled IVR application design. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston: Kluwer, pp. 147–162.
- Goolkasian, P. (2000). Pictures, words, and sounds: From which format are we best able to reason? *The Journal of General Psychology*, 127(4):439–459.
- Grasso, M. and Finin, T. (1997). Task integration in multimodal speech recognition environments. *Crossroads*, 3(3):19–22.
- Hardy, H., Baker, K., Devillers, L., Lamel, L., Rosset, S., Strzalkowski, T., Ursu, C., and Webb, N. (2002). Multi-layer dialogue annotation for automated multilingual customer service. *Proceedings of the ISLE Workshop on Dialogue Tagging for Multimodal Human Computer Interaction*. Edinburgh.
- Karsenty, L. (2002). Shifting the design philosophy of spoken natural language dialog: From invisible to transparent systems. *International Journal of Speech Technology*, 5:147–157.
- Martin, A. and Przybocki, M. (2001). Analysis of results. *2001 NIST Large Vocabulary Conversational Speech Recognition Workshop*.
- Mayer, R. and Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in

- working memory. *Journal of Educational Psychology*, 90(2):312–320.
- Mayer, R., Moreno, R., Borrie, M., and Vagge, S. (1999). Maximizing constructivist learning from multimedia communications by minimizing cognitive load. *Journal of Educational Psychology*, 91(4):638–643.
- Mousavi, S.Y., Low, R., and Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87(2):319–334.
- Novick, D., Hansen, B., Sutton, S., and Marshall, C. (1999). Limiting factors of automated telephone dialogs. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston: Kluwer, pp. 163–186.
- Shneiderman, B. (1997). *Designing the User Interface*. 3rd ed. Reading, MA: Addison-Wesley.
- Velayo, R.S. and Quirk, C. (2000). How do presentation modality and strategy use influence memory for paired concepts? *Journal of Instructional Psychology*, 27(6):126–135.
- Walker, M., Fromer, J., Di Fabbriozio, G., Mestel, C., and Hindle, D. (1998). What can I say?: Evaluating a spoken language interface to email. *Proceedings of the Conference on Human Factors in Computing Systems*. NY: ACM, pp. 582–589.
- Yeung, A. (1999). Cognitive load and learner expertise: Split-attention and redundancy effects in reading comprehension tasks with vocabulary definitions. *The Journal of Experimental Education*, 67(3):197–212.