ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

EFFECTS OF DIALOG INITIATIVE AND MULTI-MODAL PRESENTATION STRATEGIES ON LARGE DIRECTORY INFORMATION ACCESS

S. Narayanan, G. Di Fabbrizio, C. Kamm, J. Hubbell[1], *B. Buntschuh*[2], *P. Ruscitti, J. Wright*

AT&T Labs – Research
180 Park Avenue, Florham Park, NJ, 07932
{shri,pino,cak,ruscitti,wright}@research.att.com, hubbell@hfi.com, bb@tellme.com

ABSTRACT

This paper compares the effects of three different dialog initiative strategies (system initiative, mixed initiative and user initiative) on system performance and user acceptance on a large directory information access task. We used a personnel directory query application that could be accessed from a voice-only (telephony) and a multi-modal (kiosk) interface. Although the user initiative condition resulted in a lower proportion of in-grammar utterances, no significant effects of dialog initiative were observed for concept accuracy, perceived task completion, ease of use or user satisfaction. Dialogs were significantly shorter with the kiosk interface than with the telephony interface, and users preferred the kiosk interface and found it easier to use.

1. INTRODUCTION

Several recent studies of spoken-dialog systems have compared the effects of dialog initiative on system performance and user satisfaction [1],[2],[3],[4]. Depending on the complexity of the task, these studies have generally concluded that user acceptance and system performance (in terms of task success or concept accuracy) for systems using system-driven dialog strategies is equal to or better than that of systems using mixed-initiative strategies. In contrast, how dialog initiative strategy affects the performance and user acceptance of a multi-modal speech-enabled application has not been widely explored. This study compares the effects of three different dialog initiative strategies on system performance and user acceptance for a large directory information access application. Two different access modes were tested: a telephony-only system and a multi-media multi-modal kiosk.

2. A CONVERSATIONAL MULTI-MODAL-MULTIMEDIA SYSTEM FOR DIRECTORY ACCESS

2.1 mVPQ APPLICATION

The mVPQ (Multi-modal Voice Post Query) application provides access to contact information of AT&T personnel, as well as call completion. As described in [5], the goal of the dialog is to obtain the specific listing information requested by the user. This requires that the system acquire two information elements (i.e., “concepts”) from the user: a) the name of the person and b) the type of information/action desired (e.g., phone number, email address, place a call). If the name is not unique in the database, the system engages in a disambiguation sub-dialog to determine a unique solution, if one is possible.

For the experiments reported here, a subset of the entire AT&T corporate directory encompassing AT&T Labs was used. This directory listed 4361 employees, but with expansion to include nicknames, last names only, multiple pronunciations, and the possessive form, the resultant lexicon for mVPQ has 54,626 distinct entries.

2.2 MULTI-MODAL SYSTEM ARCHITECTURE

The application was developed on a standards-compliant computer telephony-IP telephony architecture that was extended to support multi-modal services. The system architecture is described in detail in [6]. The system used an ECTF compliant CT server [7] that included an Application Resource Manager (ARM) that controlled several resources performing media operations such as Text-To-Speech Synthesis (TTS), Automatic Speech Recognition (ASR), signal detection and generation, playing and recording prompts, database query (LDAP), and graphical user interface (GUI) management. The current instantiation of the platform uses AT&T’s Watson technology for speech recognition [8] and synthesis [9].

[1] J. Hubbell is now at Human Factors International.
[2] B. Buntschuh is now with TellMe Networks.

ARM sends information about user input and system status to the dialog manager and translates dialog manager directives to internal function calls and synchronizes the thread of execution. For example, each ASR result, including the associated lexical semantic tags and scores, are passed by the ARM to the dialog manager, which uses a natural language understanding (NLU) module to determine the meaning of the utterance. The dialog manager is responsible for determining the structure of the interaction with the user, based on the current context of the interaction and the most recent ASR/NLU result, dynamically adapting to the current conditions to resolve ambiguities, uncertainties and error conditions. The dialog manager determines both the action to take, based on the current dialog state, and the corresponding content for output presentation. The dialog manager (DM) was implemented using the DMD scripting language, following the AT&T Mixed Initiative Design Architecture (AMICA) [10].

GUI management is accomplished via the Generic Multi-modal Interface (GMMI) [11], which provides a graphical user interface with audio-visual I/O capabilities. Using dynamic HTML documents with embedded Java script functions, it is able to display information and capture mouse clicks, stylus taps and keyboard strokes. GMMI uses the Microsoft (MS) COM technology to integrate (1) a standard ITU H.323 terminal (MS NetMeeting), (2) a web browser with dynamic HTML support (MS Internet Explorer), and (3) 3D animations (MS Agent component). The GMMI logic acts like a message router, intercepting all the events from the underlying component modules and redirecting them to the ARM, and vice versa. Since ARM directly communicates with the dialog manager, this architecture gives full control to the dialog manager for content generation, including displaying DHTML pages, animating 3D agents and executing function scripts.

2.3 DEVICE DEPENDENT DIALOG STRATEGIES

The directory query application was available via two different access devices – a telephone, which allowed only voice input and voice output, and a kiosk interface, which accepted speech and touchscreen or typed input and provided both speech and graphical/text output. A screen shot of the opening screen of the kiosk interface is shown in Figure 1. Because of the persistence of visual information, the kiosk condition afforded the opportunity to present more information in a single presentation than was reasonable in the audio-only telephone system.

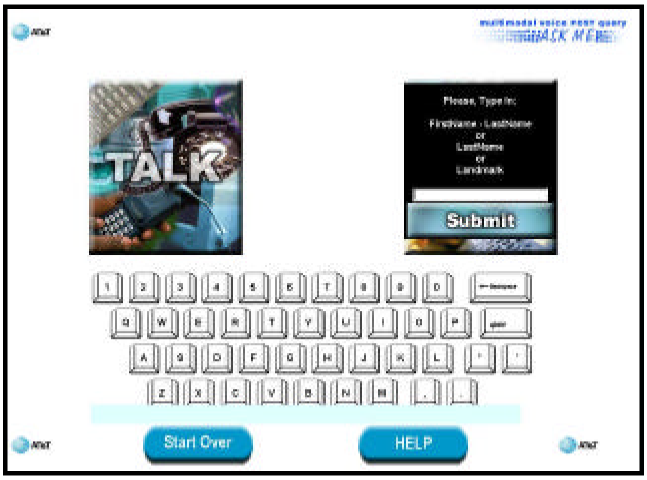


Figure 1. Opening GUI for mVPQ Kiosk

Besides information presentation, error control (in cases of ASR or understanding rejections) and disambiguation strategies (e.g., when there are multiple matching listings for a user's query) are also provided by the DM in a device-dependent fashion. For example, as shown in Figure 2, when name disambiguation was needed, the audio-visual capability of the kiosk allowed multiple disambiguating fields (e.g., the person's full name and picture and the location name and picture) to be presented simultaneously. If a user requests "room number for Rose", and there are 5 matching listings for the surname Rose, the kiosk will provide audio output stating "I have 5 listings for Rose. Choose the one you want," and display a list all the matches with additional information – complete name, photo, and location – to help resolve the request as shown in Figure 2.

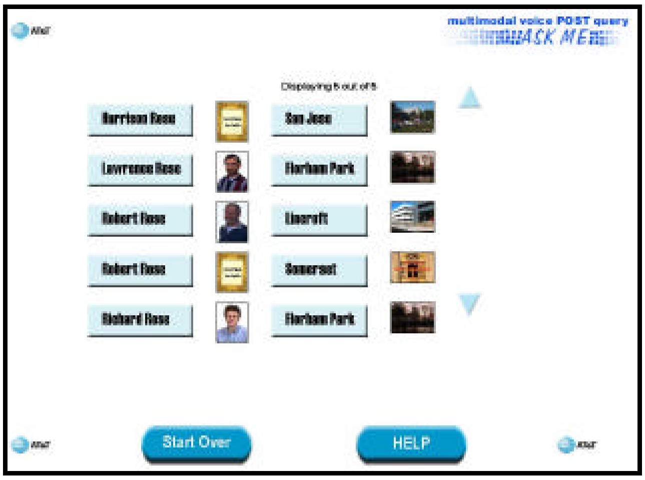
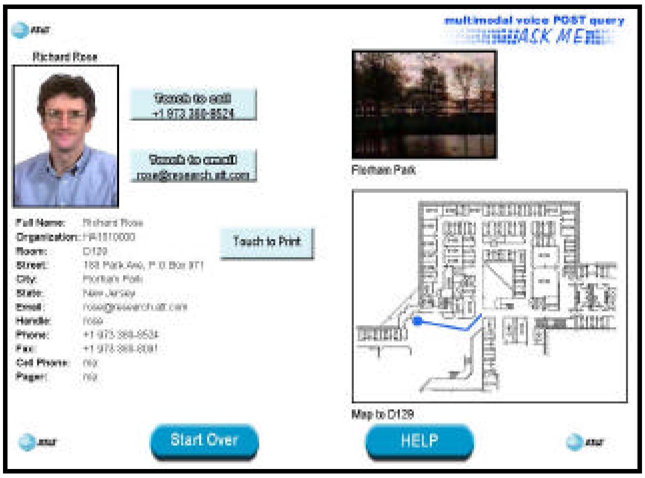


Figure 2. mVPQ Kiosk Disambiguation Display

The user can then speak or touch the desired listing to obtain the complete listing - in the example shown, the user could have said "Richard", "Richard Rose" or "Florham

Park" or touched the name, location, or picture associated with that entry to get the complete information shown in Figure 3, highlighting any specifically requested information elements. In contrast, the disambiguation strategy for the audio-only telephone interface solicits information for disambiguation on information element at a time (e.g., "I have 5 listings for Rose. Please say the first name."), with the sequence chosen automatically to minimize the number of questions required to achieve resolution. For the example above, the telephony system would solicit location information first, and should the user not provide that, it would then request first name information.


multimodal voice POST query
ASK ME
Richard Rose
Touch to call +1 973 360-8524
Touch to email rose@research.att.com
Touch to Print
Florham Park
Map to D129
Start Over
HELP

Full Name: Richard Rose
Organization: HA1510000
Room: D129
Street: 180 Park Ave, P O Box 971
City: Florham Park
State: New Jersey
Email: rose@research.att.com
Handle: rose
Phone: +1 973 360-8524
Fax: +1 973 360-8091
Cell Phone: n/a
Pager: n/a

Figure 3. mVPQ Kiosk Information Display

This ordering of the questions would be used because the locations for the 5 listings for "Rose" are distinct, whereas disambiguation cannot be guaranteed by first name alone (i.e., there are two different employees with the name "Robert Rose"). If the user is unable to provide any additional information to disambiguate, and if the total number of listings is deemed reasonable (a maximum of 5 listings in our design), the DM will report the requested information element (e.g., room number) for those listings as a final fallback option; otherwise, the interaction terminates as incomplete.

3. EXPERIMENTAL DESIGN

This experiment tested three different dialog initiative strategies, reflecting constraints on both the system prompts and on the ASR grammar. In the **System Initiative (SI)** condition, audio prompts were highly directive, and the ASR grammars were constrained to the set of valid responses to the prompts (plus several additional key words [e.g., help, cancel]). For example:

> ***SYSTEM:*** *VPQ. Please say the name of the person.*
>
> ***Acceptable Response from USER:*** *Larry Rabiner.*

This condition typically handles only a single input concept per user turn. Because the kiosk interface always allows some user initiative, the SI condition was only tested with the telephone interface.

The **Mixed Initiative (MI)** condition used directive prompts, but the ASR grammars were relatively unconstrained, so that, if the user provided extra task-related information, the system was capable of handling it:

> ***SYSTEM****: VPQ. Please say the name of the person.*
>
> ***Acceptable Response from USER:*** *Larry Rabiner's fax number please.*

In the **User Initiative (UI)** condition, the initial system prompt was open-ended, and the corresponding ASR grammar was identical to the mixed initiative task:

> ***SYSTEM:*** *VPQ. What can I do for you?*
>
> ***Acceptable response from USER:*** *I'd like the fax number for Larry Rabiner.*

Fifty employees of AT&T Labs Research (ten per initiative and interface combination) participated in a scenario-based experiment. Each subject performed six tasks and answered a brief survey after each task and a general survey after the completion of the experiment. Five scenarios were information query tasks, and one involved placing a telephone call. Three of the scenarios required disambiguation sub-dialogs to complete the task. The other three scenarios could also trigger a disambiguation dialog, depending on the subject's wording of the initial query. Two of the tasks (placing the telephone call to a colleague and finding out the organization number of the user's supervisor) allowed the subject to decide who to ask about. To access the telephony interface, subjects dialed a toll-free number from an ISDN telephone. Access to the kiosk interface was initiated by picking up an analog telephone handset at the kiosk. The off-hook signal generated a phone call to the service running on the CT server via a cable modem/IP telephony gateway system.

Speech utterances were transcribed and, along with detailed log files from each dialog, were used to derive a set of objective measures, which included concept accuracy, percent of in-grammar utterances, user turns, and number of concepts provided on the first query of a dialog.

Table 1. Mean Results

	Kiosk		Telephone		
	MI	UI	SI	MI	UI
Concept Accuracy (overall) (%)	69.8	60.9	73.3	75.6	63.5
Concept Accuracy (in grammar) (%)	80.0	80.4	81.0	79.7	74.9
In-grammar utterances (%)	82	72	89	94	88
User Turns	3.6	3.6	5.8	5.9	4.5
# of Concepts on First Query	1.4	1.9	1.0	1.0	1.9
Perceived Comp. (%) – 6 tasks	86.6	91.7	76.7	79.3	73.3
Perceived Comp. (%) – 5 tasks	90.0	94.0	87.5	82.0	84.0
Ease of Use	1.5	1.9	2.7	3.3	2.6
User Satisfaction	2.1	1.8	3.0	3.6	2.1

Subjective measures extracted from the surveys included perceived task completion and ease of use measures for each task, as well as an overall judgment of ease of use and user satisfaction on the general survey. Ease of use and user satisfaction were rated on a 7-point scale, with 1 corresponding to “very easy” and “very satisfied”, and 7 corresponding to “very difficult” and “very dissatisfied”, respectively.

4. RESULTS AND DISCUSSION

Mean results for each initiative condition are shown in Table 1. Concept accuracy is reported for in-grammar utterances only (in grammar) and for in-grammar and out-of-grammar utterances combined (overall). Overall concept accuracy can be viewed as concept accuracy from the user’s perspective, while in-grammar concept accuracy reflects system performance for that subset of the utterances that were within the constraints of the ASR grammars used for this task.

ANOVAs were run for each dependent measure to evaluate the effects of initiative strategy (SI, MI, UI) and interface type (kiosk, telephone). All effects were evaluated against a p<0.05 significance criterion. There were no significant effects of initiative strategy for concept accuracy, perceived task completion, ease of use or user satisfaction. The proportion of in-grammar utterances was significantly lower for the UI strategy than for the SI or MI strategies (F(2,45)=3.88).There was a significant interaction of dialog initiative and task for user turns (F(10,225)=3.41), with UI condition yielding the shortest dialogs on most tasks, and the SI condition generally yielding longer dialogs. In addition, there was a significant interaction of task, initiative and interface type for the number of concepts offered at the initial query (F(5,495)=4.63). The mean number of concepts offered by the subjects was significantly higher for the UI strategy than for SI and MI, suggesting that subjects generally complied with the constraints imposed by the wording of the opening prompt in the latter conditions. Coupled with the lack of significant differences in concept accuracy, this result suggests that increasing the size of the ASR grammars (from the SI to MI conditions) did not adversely affect performance for the directory query task. Our results for concept accuracy are not consistent with other studies [1] comparing SI and MI dialog strategies. We suspect that the relatively low complexity of this task domain in terms of number of required concepts to complete the task may be the reason that performance was not affected significantly across these dialog initiative conditions.

Comparing across interfaces, significantly higher mean user satisfaction (F(1,45)=4.05) and ease of use ratings were observed for the kiosk condition, despite a significantly lower mean proportion of in-grammar utterances (F(1,45)=22.6). There was a significant interaction between task and interface for the ease of use measure (F(5,180)=2.06). The scenario requiring subjects to place a phone call was rated easier to use with the telephone interface than with the kiosk interface, whereas for all the other scenarios (which were information query scenarios), the kiosk was rated as easier to use. Perceived task completion was significantly higher for the kiosk interface when all 6 tasks were considered (F(1,45)=7.1), but there were no differences when the analysis did not consider one of the tasks, which had a null outcome (i.e., the requested information was not available in the database, so the system responded with “I’m sorry, but there is no pager number for Cynthia Smith.”). In this case, the subjects interpreted the concept of “task completion” ambiguously. Interestingly, the subjects in the telephony condition were more like to consider this task as not completed than the subjects in the kiosk condition. It is possible that the more complete feedback about the requested listing that was provided in the kiosk condition gave subjects more confidence in judging that the task had in fact been completed than in the telephone condition, where users only heard about the particular information element they had requested.

Kiosk dialogs, with an average of 3.6 dialog turns per task, were significantly shorter than telephony dialogs, which averaged 5.4 dialog turns per task (F(1,45)=10.75, p<0.01). This reflected the frequent use of the touch screen for disambiguation. Fifteen of the 20 subjects using the kiosk interface used the touchscreen for at least one disambiguation turn. Overall, the touchscreen was used on 49.2% of the disambiguation turns. This result suggests that subjects found it easy to switch input modes from speech to touch during the course of a task.

5. SUMMARY

In summary, these experiments demonstrate that the wording of the initial query in a dialog has a significant impact on the number of concepts spoken by users of the system, and also influences the likelihood of obtaining in-grammar utterances. For this directory query application, concept accuracy and task completion were not affected by dialog initiative strategy The multi-modal kiosk interface, presenting richer output information as well as allowing either spoken or touch input, yielded shorter dialogs and was preferred over the telephony-only interface.

6. REFERENCES

[1] Danieli, M. and Gerbino, E. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[2] Walker, M., Fromer, J., Di Fabbrizio, G., Mestel, C. and Hindle, D. What can I say: Evaluating a spoken language interface to email. In *Proceedings of the Conference on Computer Human Interaction (CHI 98)*, 1998.

[3] Potjer, J., Russel, A., Boves, L, and den Os, E. Subjective and objective evaluation of two types of dialogue in a call assistance service. In *Proceedings of the 1996 IEEE Third Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA)*, 89-92, 1996.

[4] Billi, R., Castagneri, G, and Danieli, M. Field trial evaluations of two different information inquiry systems. In *Proceedings of the 1996 IEEE Third Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA)*, 129-134, 1996.

[5] Buntschuh, B., Kamm, C., Di Fabbrizio, G., Abella, A., Mohri, M., Narayanan, S., Zeljkovic, I., Sharp, R. D., Wright, J., Marcus, S., Shaffer, J., Duncan, R. and Wilpon, J. G. VPQ: A Spoken Language Interface to Large Scale Directory Information, *Proc. ICSLP 98*, 1998.

[6] Di Fabbrizio, G., C. Kamm, P. Ruscitti, S. Narayanan, B. Buntschuh, A. Abella, J. Hubbell and J. Wright. Extending a Standard-based IP and Computer Telephony Platform to Support Multi-modal Services, *Proc. of Interactive Dialogue in Multi-modal Systems*, Kloster-Irsee, Germany, pp.9-12, 1999.

[7] ECTF Architecture Framework - Revision 1.0. http://www.ectf.org/ectf/pubdocs/arch_fr.pdf.

[8] Sharp, R. D., Bocchieri, E., Castillo, C., Parthasarathy, S., Rath, C., Riley, M. and Rowland, J. The Watson speech recognition engine. *Proc. ICASSP 97*, 4065-4068, 1997.

[9] Beutnagel, M., Conkie, A., Schroeter, J., Styliano. Y, and Syrdal, A. The AT&T Next Gen TTS System. *Proc. Joint Mtg. ASA, EAA and DEGA*, Berlin, 1999

[10] Pieraccini, R., Levin, E. and Eckert, W., AMICA: the AT&T Mixed Initiative Conversational Architecture, *Proc. EUROSPEECH 97*, Rhodes, Greece, Sept. 1997.

[11] Di Fabbrizio, G., Narayanan, S., Ruscitti, P., Kamm, C. Buntschuh, B., Abella, A., Hubbell, J. and Hamaker, J. Unifying conversational multimedia interfaces for accessing network services across communication devices, *Proc. ICME 2000*, New York, New York, 2000.