DIALOG DESIGN FOR USER ADAPTATION

Esther Levin and Alex Levin

City College of New York Spacegate, Inc.

ABSTRACT

It is a common belief that repetitive users can adapt to a spoken dialog system. In this paper we describe a dialog design that allows experienced users to make their interactions with the system more efficient and present experimental evaluation of such adaptation. In our study we focused on the application of dialog systems as a tool for real-time data collection for healthcare. Specifically, we implemented a dialog system, Pain Monitoring Voice Diary, for monitoring chronic pain patients and conducted a usability study involving 171 dialog sessions with 24 users. Breakdown of the data according to the level of user experience indicates that experienced users adapt and take advantage of dialog design to make their interaction more efficient.

1. INTRODUCTION

Recent research shows the applicability of spoken dialog technology to healthcare related applications [1-3] where the users conduct dialogs with the system on a regular basis, often identifying themselves in the beginning of each session. Good dialog design in such applications should provide flexible level of user support to accommodate both novice callers and experienced callers: For the experienced caller, the system needs to provide short and effective call flow, without making the caller hear long and tedious prompts; For the novice caller, the system needs to provide enough information and help to guarantee question understanding and successful session completion. In this paper we describe a dialog system designed to provide such flexible level of support to the users, and show that with such system experienced users indeed are able to make their interactions more efficient.

We focused our research on the application of healthcare data capture[2,3], where spoken dialog system collects the data through an over the phone interaction with the subject, stores and analyzes it in real time. Traditional method for such data collection is paper-based questionnaires filled by the subjects. Using spoken dialog technology for this application has the following advantages:

- Speech is a natural modality of interactions for humans, and the input device – the phone – is user friendly and ubiquitous and no special training for its use is required - Compliance is monitored automatically: the calls can be initiated by a system following a prescribed protocol, and the system can report about any non-compliance to trial administrator in real time.
- Spoken automated dialog reaches much beyond voice-enabling static paper questioners: possible answers are not limited by number of check-boxes to fit on a piece of paper; question selection can be done dynamically based on previous answers; personalization of both content and style based on the patient's history is possible.
- The ability to transform the captured data into real-time reports, and further interface the information with other clinical or back-office systems and databases provides an unparalleled opportunity to enhance patient feedback and monitoring. Overall ASR based system offers the caregiver an extensive and practical tool to facilitate efficient and convenient patient communications, which saves time while increasing quality of care.

For this study we implemented a dialog system for chronic pain patient's assessment and monitoring, an application for which well established standard questionnaires [4–6] are available, and the vocabulary for potential answers can be established from the medical literature. Fig. 1 shows the dialog flow for Pain Monitoring Diary. The dialog flow is represented as a series of dialog units, where each unit comprises several caller-system exchanges designed to elicit one piece of information from the caller to fill a slot in the session report.

2. DIALOG DESIGN FOR FLEXIBLE LEVEL OF USER SUPPORT

We used the following mechanisms to provide for a flexible level of user support that is intended to satisfy both the novice and the experienced users:

- Prompt Design. The system prompts are designed to provide an appropriate level of support to the user. For example, the initial prompt for the 'Pain Location' dialog unit is "*Where does it hurt? <pause>. For example, your head stomach or back? <pause>. Remember, if you don't know how to answer this question, just say 'I need help' ".* The pauses in this prompt are designed to encourage the experienced user to barge in with the answer (most experienced users barge in after the initial *"where does it hurt"* portion of the prompt), while providing more information (in this case, examples of possible answers) for the inexperienced user who hesitates to answer immediately. It also reminds the user to ask for help if it is still not clear what can be said as an answer.

- Context sensitive help. Help information is provided on user's request, describing and clarifying the current question, and in some cases enumerating the possible answers the caller can choose from, while in other cases giving more examples of possible answers. For example, if the caller asks for help after the *"where does it hurt"* question, the system will provide a very elaborate help prompt that lists different body parts that the user can say (pausing shortly after each one to encourage the user to barge-in if the user knows what to say). It also reminds the user that they can choose the "none of those" option: "*Okay. Here is the help information. At this point I need to find out the part of your body that hurts the most. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of them matches, please say 'none of those'. Here is the list: abdomen <pause>, ankles <pause>, back <pause>,...(list continues) ..., toes <pause>. Which one is it?"* .

- Detecting speech recognition failures. Even when the user has not asked for help explicitly, the dialog is designed to detect user's repeated failures and provide more support. When the system experiences recognition problems such as rejection or silence, it will re-prompt the user again for the same question. The re-prompts are designed as an escalating list, providing increasingly more information and progressively constraining the user as more such errors are detected. For example, if the user's utterance is rejected by the recognizer after the initial prompt: "*Where does it hurt? <pause> For example, your head, stomach or back? <pause>. Remember, if you don't know how to answer this question, just say 'I need help"*., the system will re-prompt for the same information with "*I didn't get that. Please tell me the part of your body that hurts the most, Remember, you could always say 'I need help'* ", the second prompt skips the pauses and reminds the user to ask for help if needed, and also clarifies the question ("body part that hurts the most").

Another case where the system detects that something went wrong with speech recognition, is when the user says "no" to a confirmation question as in:

System prompt: *Was that your left shoulder?*

User: *No.*

System prompt: *Sorry about that. Let's try it this way. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of them matches, please say 'none of those'. Here is the list: abdomen <pause>, … (list continues). Which one is it?*

Since the user disconfirmed the recognized body part, the system detects a recognition problem and gives the user more information on how this question can be answered to minimize the out-of-grammar utterance rate.

- Dialog Personalization. The knowledge of caller identity (callers identify themselves in the beginning of each session) provides a system with an opportunity to personalize both the content of the current session (what is the data to be collected) as well as the style (how to ask for these data) based on the results of the previous sessions. As shown in fig. 1, in our system we took advantage of a larger inter-session context by designing two types of data collection sessions: *normal* and *follow up*. The follow-up session type is deployed if the subject reported a high level of pain in the previous session. The follow-up session differs from the normal one not by the additional questions the patient is asked such as if and when the subject took the medication, etc, but also by the format of the questions. If in the previous session the subject reported pain in left shoulder, in the follow up session the question will be "is the pain still in your left shoulder?". This format of "reminding" prompts was used for pain location and pain type dialog units, and it was designed to possibly shorten the dialogs and also provide the subject comfort and feeling of continuity in using the system.

3. CONTROLLING CAPTURED DATA ACCURACY

Data validity, accuracy and integrity in healthcare applications are very important, since the penalty for an erroneously filed final session report can be very high. We designed the system to take into account the known limitations of automated speech recognition technology and to be able to ensure the overall high accuracy of data capture and session completion rate by:

a) Improved rejection mechanisms for confirmation and other grammars. We incorporated a garbage model in the yes/no grammar used for confirmations in our application. The garbage model was designed to match out-of-vocabulary utterances [7, 8], specifically the corrections users are frequently providing instead of negative confirmation, e.g.,

System prompt: *Was that your left shoulder?*

User: *no, right shoulder*

We used rejection criterion based on combination of recognition score and garbage model scoring to control the overall accuracy of this grammar.

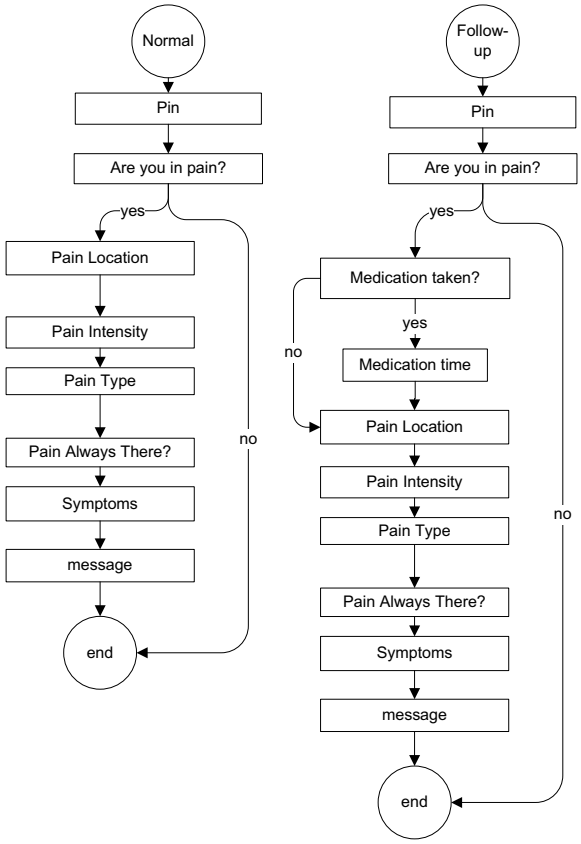

Fig. 1: Dialog flow is described in terms of dialog units for normal and follow-up session types.

b) Using confirmations as the way to control the larger grammar's accuracy. The grammars that are substantially larger than yes/no are also those for which we can expect more ASR errors and out-of-vocabulary utterances. Those are grammars like the body-part grammar, or the symptoms grammar when, without substantial data collection, we cannot accurately predict all possible ways the users will answer the "where does it hurt?" question or the "what's your most disturbing symptom?" question. For such grammars, we use the confirmation mechanism to control the overall accuracy of the data we capture. The result is considered captured only if the user answers "yes" to the confirmation question, reducing the error rate for the dialog units with larger grammars to the level of yes/no grammar.

c) Using recording to capture the out-of-grammar answers and problematic user inputs. In some cases, e.g. when the user is trying to answer the "where does it hurt?" question with a word that is not covered by "body part" grammar, the confirmation mechanism does not help. For cases like this one, we ask the user

to say a key-phrase like 'none of those' and then just record the user's input:
System prompt: "*Was that your left shoulder?"*
User: "*No"*
System prompt: "*Sorry about that. Let's try it this way. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of the locations match, please say 'none of those'. Here is the list: abdomen <pause>, ankles …"*
User (barges in):"*none of those"*
System prompt: *"Ok. Let me just record your answer. Please describe the location of your pain in your own words."*
User: <……>
System prompt (after recording is finished): "*Thanks, I got that. Let's move on."*
The recorded utterance is captured and flagged as "transcription is needed" for later processing. The same mechanism of fall-back to recording instead of recognition is used after several repeated recognition failures.

4. EXPERIMENTAL EVALUATION

Experimental evaluation of usability of the Pain Monitoring Voice Diary was performed with 24 volunteers, mostly students recruited on campus. The goal of this evaluation was to validate the assumptions underlying dialog design.
The volunteers were asked to contribute ten sessions with the system over a period of 2 weeks; in practice the number of sessions per subject ranged from 1 to 20. The subjects were asked to either relate to pain episodes in their past while answering the system's questions, or use as a guidance one of 9 provided medical scenarios compiled by a pain specialist, ranging from migraines and back pain to post-surgery pain (knee injury), and cancer and chemotherapy-related afflictions.

4.1 Dialog Evaluation

We collected the total of 177 dialog sessions: 171 sessions were completed, while in 6 the called hung up. Sixty six of the completed session were of the 'follow-up' type. There were a total of 2437 dialog turns, where dialog turn corresponds to one system prompt and one user utterance. The data capture rate, measuring the percentage of slots filled automatically was 98%, while the other 2% were flagged for transcription. Data capture rate is not a direct measure of ASR accuracy since slots are not necessarily filled after first attempt. Among the utterances sent to transcription, where the user had opted for the 'none of those' option, 70% corresponded to the type of pain slot, 20% to the symptoms slot, and 10% to the body part slot, indicating that those are the grammars with the highest out-of-vocabulary rate.
Table 1 shows other metrics derived from dialogs[9]: average session duration; number of dialog units per session; average duration of a dialog unit; average number of caller utterances in dialog unit; average duration of one dialog turn; percentage of barged-in prompts and percentage of task-oriented prompts. The high standard deviations of session duration and dialog units per session are due to the extensive variability of dialog sessions. Not only the sessions differ by type (normal and follow up), but also there is branching within the same type application (e.g., some of the subjects report symptoms, while others don't, some take medications, etc). In addition there is a great variability due to ASR errors and different possibilities inherent in the design of the call flow (e.g., caller initiated help requests, speech recognition error handling such as re-prompts, negative confirmations.)

Session duration (sec)	99.34(45.92)
Number of dialog units per session	7.65 (2.48)
Duration of dialog unit (sec)	12.99 (2.7)
Dialog turns per dialog unit	1.86 (0.43)
Percentage of task oriented turns	82% (15.4)
Percentage of barged-in prompts	68% (13)
Time duration of a dialog turn (sec)	6.97 (1.3)
Time duration of a dialog turn when barge-in was disabled	10.63(1.5)

Table 1: Dialog session statistics (figures in parentheses are standard deviations)

The high standard deviations in caller utterances per dialog unit and dialog unit duration are due to the fact that not all dialog units are created equal. For example, 'Are you in pain' dialog unit can fill a slot with a single 'yes/no' utterance, while 'Pain Location' unit requires at least 2 dialog utterances (body part and confirmation) if speech recognition does not fail, and more if it does.
Percentage of task-oriented dialog turns (82%) (those are dialog turns that are NOT due to speech recognition errors or caller help requests) is a measure of dialog efficiency: if there were no errors and help requests at all, it would be 100%. The prompts in the dialog were designed to be barged-in by experienced callers. To quantify the use of barge-in we computed the percentage of barged-in prompts (68%). To quantify how far in the prompts the barge-in occurs we computed the average duration of dialog turn (6.97 sec), and compared it to the reference of average duration of dialog turn (10.63 sec) when barge-in was disabled.

4.2 Evaluation of Flexible Level of User Support

One of the goals of the dialog design described above was to have a flexible and adaptive user support for different types of users, providing short prompts and efficient call flow for experienced users, while providing more detailed information in a troublesome situation and for novice users. To evaluate the efficiency of the dialogs as a function of user proficiency, we divided the sessions into seven classes according to the sequential order of the session with same user. Table 2 shows some statistics of the classes. For example, class A contains all the first sessions each of the 24 users had, with a total of 308 dialog turns; while class G contains all the sessions (whose ordinal number was ten and above) for which the users had had previously at least 9 sessions completed.
Figures 2,3 and 4 illustrate the average dialog turn duration, average percentage of barged-in prompts, and average percentage of task oriented prompts for the classes of Table 2 separately.
The differences between the 7 session classes for the three metrics shown are statistically significant, as tested by ANOVA [10, 11], with F measure of above 49 and P less than 0.0001 for all three metrics. The error bars in the figures indicate 95% confidence interval. The results in figures 2 and 3 confirm the assumptions of the dialog design: the prompts were designed to be barged in by experienced users, and indeed, the results indicate that the more experienced the user is, the more often and earlier she will barge in: the novice user barges in only on 59% of the prompts, with an

average turn duration of 7.7 seconds, while users that had more than 9 sessions completed in the past barge in on 73% of the prompts, with an average dialog turn duration of 6.5. Figure 4 shows that with experience the users become more efficient with the system, as measured by the percentage of task-oriented dialog turns: for novice users this percentage averages at 75%, increasing to around 81% after just one previous session was completed, and up to 86% after at least 9 sessions were completed previously.

Class Name	Call Order	# Sessions	# Turns
A	1	24	308
B	2	19	302
C	3	15	206
D	4, 5	27	380
E	6, 7	23	324
F	8, 9	21	305
G	10+	43	612

Table 2: dialog sessions divided according to the call order

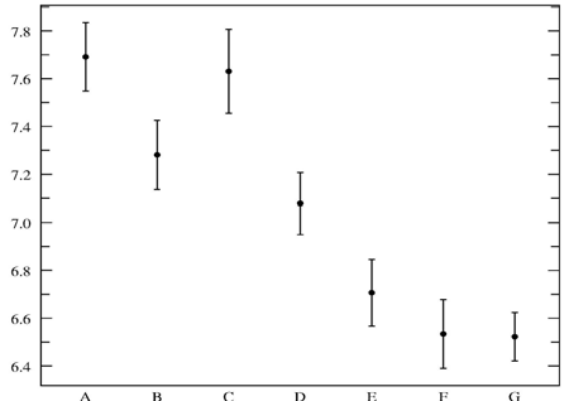

Fig. 2: Average turn duration [sec] for dialogs in classes A-G.

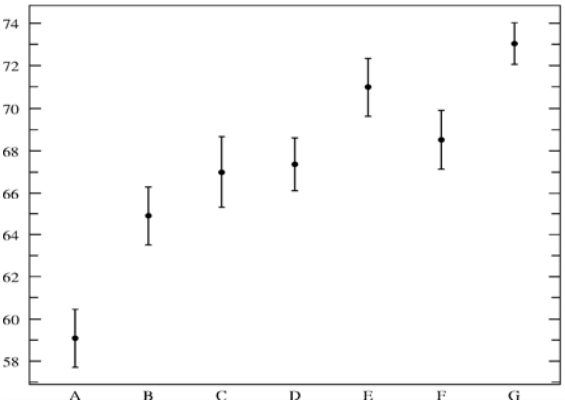

Fig. 3: Percentage of barged-in prompts for dialogs classes A-G.

5. SUMMARY

This paper describes a dialog system designed to provide a flexible level of user support that allows experienced users to make their interactions with the system more efficient, while providing novices sufficient support to complete their sessions. We present experimental results showing how users' effectiveness with the system changes as a function of their level of experience.

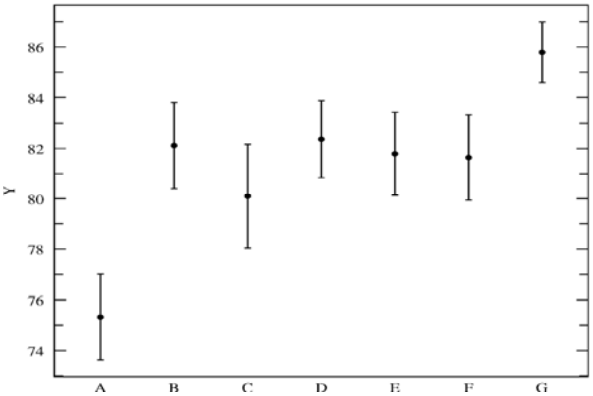

Fig. 4: Percentage of task-oriented turns for dialogs classes A-G.

6. ACKNOWLEDGEMENTS

The Pain Monitoring Voice Diary system (PMVD) developed by Spacegate, Inc. under brand name – SpeechMatrix is currently scheduled for validation trials with Beth Israel, NY Cancer Center. The project described was supported by grant "Automated Speech Real-Time Patient Data Collection" from NIH/NCI. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

7. REFERENCES

[1] Black, L, McTear,M. , Black, N. , Harper, R. and Lemon, M. "The Voice-Logbook: Integrating Human Factors for Chronic care System.", *ICSLP*, Jeju Island, Oct 2004.

[2] Black, L, McTear,M. , Black, N. , Harper, R. and Lemon, M. "Evaluating the DI@L-log System on a Cohort of Elderly, Diabetic Patients: Results from a Preliminary Study", Interspeech2005, Lisbon, Portugal.

[3] E. Levin, A. Levin, "Spoken Dialog System for Real-Time Data Capture" , in Proc. Interspeech2005, Lisbon, Portugal.

[4] Davis, Mellar P., Walsh, Declan "Cancer Pain: How to measure the fifth vital sign", *Cleveland Clinic Journal of Medicine*, vol 71, Num 8, August 2004.

[5] Daut R.L., Cleeland C.S., Flanery R.C. "Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases."*Pain* 1983 Oct;17(2):197-210

[6] Turk, C. and R. Melzack. (2001) Handbook of Pain Assessment, Second Edition, July 2001

[7] R. Rose and D. Paul, "A Hidden Markov Model Based Keyword Recognition System," *Proc. ICASSP*, Albuquerque, 129–132, 1990.

[8] Manos and V. Zue, "A Segment-based Spotter Using Phonetic Filler Models," *Proc. ICASSP*, Munich, 1997.

[9] M. A. Walker et al. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc ACL/EACL, San Francisco, pp. 271-280.*

[10] Eugene S Edgington, "Randomization Tests", July 1986, Marcel Dekker, Inc.

[11] http://www.physics.csbsju.edu/stats/Index.html