

Spoken Dialog System for Real-Time Data Capture

Esther Levin and Alex Levin

City College of New York, USA, Spacegate, Inc., USA

esther@cs.cuny.cuny.edu alex@spacegate.com

Abstract

This paper reports on the development of spoken dialog system as a tool for real-time data collection for healthcare, life and behavioral science. Specifically, we implemented a dialog system, Pain Monitoring Voice Diary, for monitoring chronic pain patients. We discuss the requirements and characteristics of this application, specifically, the needs for adaptive level of user support and high and controllable accuracy of data capture, and show their implications for the design of Pain Monitoring Voice Diary. In usability study involving 118 dialog sessions with 24 volunteers we measured 98% data capture and 80% dialog efficiency, as estimated by percentage of task-oriented prompts.

1. Introduction

The purpose of health, behavioral and life style surveys and questionnaires is to monitor and record quantitative and qualitative data and to identify patterns and changes *over-time* of health status and health-related behavior. While it is an art to design valid questionnaires, the tools and methods of data collection are not less important and often can influence the research outcome. Traditional data collection methods vary from paper-based diaries and reports, to video/audio recordings, to human observation. Recently, electronic data collection techniques, utilizing hand-held computer devices (PDAs), the Internet, or IVR, have been introduced. These methodologies enable collection of meta-data about the respondent's compliance and use of such data to measure and improve compliance, but those methods and devices also face many issues and challenges. Very often, the methods of data collection have large impact on data quality, timeliness and costs.

2. Pain Assessment via Spoken Dialog

To address the needs for innovative methods of data capture, we evaluated the applicability of Spoken Dialog Technology for real time data collection. Similarly to what was proposed in [1], dialog system collects the data through an over the phone interaction with the subject, stores and analyzes it in real time. The advantages of using this technology are:

- 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 (as opposed to PDA or computers)
- 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 questionnaires: 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.

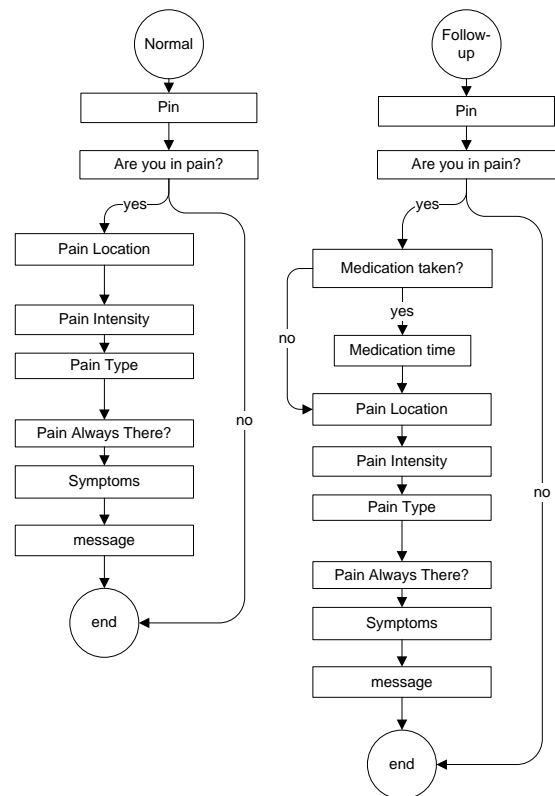


Figure 1: Dialog flow is described in terms of dialog units for normal and follow-up session types.

For this study we implemented a dialog system for chronic pain patient's assessment and monitoring. In US alone an estimate of over 10 million of individuals are living with chronic pain, and recently the Joint Commission on Accreditation of Healthcare Organizations called pain the fifth vital sign that providers should monitor in the care of patients[2], along with temperature, pulse, respiration, and blood pressure. Pain assessment is also an application for which well established standard questionnaires [2][3][4] are available, and the vocabulary for potential answers can be established from the medical literature. Figure 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. Figure 2 shows a transcribed session and its corresponding report automatically generated by the system in the end of the session.

3. Dialog Design

The characteristics and requirements of data capture task are different than those for other applications of spoken dialog technology. Successful dialog design needs to take the following specificities of this task into account:

- The subjects participating in data collection are enrolled through a personal face-to face interview at which they receive relevant information about the trial and guidance on the process of data collection. In the same opportunity the patients can receive some training, explanation and possibly a demo on how to use the spoken dialog system.

- Subjects call the system repeatedly according to the study protocol, and identify themselves in the beginning of the session. This provides an opportunity to use the knowledge accumulated across sessions for personalization.

- The system should accommodate both novice callers (in the beginning of the trial) and experienced callers (those who completed several sessions); 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.

- Data validity, accuracy and integrity in this application are very important, since the penalty for an erroneously filed final session report can be very high. Since the automated speech recognition technology is not perfect, the design has to take into account the possibility of speech recognition errors and improve the overall accuracy using dialog actions such as re-prompts, confirmations, error handling, and, if necessary, recording and flagging the unrecognized utterances for later transcription.

In the design of the dialog we addressed these task characteristics by providing an adaptive level of user support, and controlling the captured data accuracy.

3.1. Flexible Level of User Support.

The flexible level of user support that is intended to satisfy both the novice and the experienced users is achieved by deploying the following mechanisms.

- **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 is designed to remind the user to ask for help if it is still not clear what can be said as an answer.

- **Context sensitive help.** It is unreasonable to expect system users to retain the information provided in the training material or at the system orientation session for the whole duration of the trial that can last for months. Therefore, for every question in Pain Monitoring Voice Diary, 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?"* The information provided during these explicit requests for help closely follows the information the user received during the enrollment process.

- **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 the 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 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.** Data capture is a unique dialog application since not only the users call the system many times during the trial, but they also identify themselves in the beginning of each session. This 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 figure 1, in our system we took advantage of a larger inter-session context by designing two types 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.

3.2. Controlling the Accuracy of Data Capture.

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 [[5][6], 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.

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 offer 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.

System:	Please tell me your four digit pin.
Caller:	****
System:	Ok, *****, lets begin. Are you in pain now?
Caller:	yes
System:	Where does it hurt?<pause> For example, your head, stom...
Caller:	my left shoulder
System:	Was that your left shoulder?
Caller:	yes
System:	Please rate your pain on a scale of zero to ten, with zero being no pain and ten being the worst possible pain.
Caller:	it is seven
System:	Was that 7?<pause> Please say yes or no.
Caller:	yeah
System:	How does the pain feel?
Caller:	I don't know
System:	Okay, Here is the help information. At this point I need to find out the type of your pain. Please choose carefully a word from the following list that best describes your pain and just say it. If nothing matches, please say ‘none of those’. Here is the list: aching, <pause> burning, <pause> dull...
Caller:	it feels sharp
System:	I heard ‘sharp’. Is that right?
Caller:	yeah
System:	Is the pain constant or does it come and go?
Caller:	comes and goes
System:	Do you have any other symptoms? ...
Caller:	yeah
System:	Please tell me your most disturbing symptom. For...
Caller:	I have fever
System:	I heard ‘fever’. Is that right?
Caller:	yes
System:	Would you like to leave a message for the staff? ...
Caller:	no
System:	Thanks for your cooperation. I've got all the information that I need. I'll call you again in 4 hours. Goodbye.

	Captured Value	Confirmed (yes/no)	Confidence Score
Pin	****	no	66
Are in pain?	yes	no	80
Pain Location	left shoulder	yes	86
Pain Intensity	7	yes	88
Pain Type	sharp	yes	88
Pain constant?	pain comes and goes	no	47
Symptoms	fever	yes	86
Message	none	no	78

Figure2: Transcribed session with the system in normal mode and its report

4. Experimental evaluation

Experimental evaluation of usability of the Pain Monitoring Voice Diary was performed with 24 volunteers, mostly students recruited on campus. 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 12. There was no formal training session with the system provided, instead, once enrolled (through a website) the subjects received an email notification with their PIN and general information about the system. 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.

We collected the total of 118 dialog sessions: 113 sessions were completed, while in 5 the called hung up. 42 of the completed session were of the ‘follow-up’ type. There were a total of 1766 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, 80% corresponded to the type of pain slot, and 20% to the symptoms slot, indicating that those are the grammars with the highest out-of-vocabulary rate.

Session duration (sec)	105.6(46.78)
Number of dialog units per session	7.85 (2.6)
Duration of dialog unit (sec)	13.46 (4.54)
Dialog turns per dialog unit	1.88 (0.46)
Percentage of task oriented turns	80% (16)
Percentage of barged-in prompts	66% (13)
Time duration of a dialog turn (sec)	7.19 (1.10)
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)

Table 1 shows other metrics derived from dialogs[7]: 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.)

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) in the case speech recognition does not fail, and more if it does.

Percentage of task-oriented dialog turns (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 (66%). To quantify how far in the prompts the barge-in occurs we computed the average duration of dialog turn (7.19 sec), and compared it to the reference of average duration of dialog turn (10.63 sec) when barge-in was disabled.

5. Summary

In this paper we describe a new application of spoken dialog technology, i.e., real-time data collection for health, behavioral and life style studies. We discuss the characteristics of this application that are different from the common automation applications such as call routing or database retrieval, and show how these characteristics are taken into account in system design, to provide adaptive level of user support and high and controllable accuracy of data capture. Finally, we present the results of system usability study.

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] 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.
- [3] 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
- [4] Turk, C. and R. Melzack. (2001) *Handbook of Pain Assessment*, Second Edition, July 2001
- [5] R. Rose and D. Paul, “A Hidden Markov Model Based Keyword Recognition System,” *Proc. ICASSP*, Albuquerque, 129–132, 1990.
- [6] A. Manos and V. Zue, “A Segment-based Spotter Using Phonetic Filler Models,” *Proc. ICASSP*, Munich, 1997.
- [7] M. A. Walker et al. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc ACL/EACL*, San Francisco, pp. 271-280.