Evaluating the DI@L-log System on a Cohort of Elderly, Diabetic Patients: Results from a Preliminary Study

Lesley-Ann Black¹, Michael McTear¹, Norman Black¹, Roy Harper², and Michelle Lemon²

¹Faculty of Engineering, University of Ulster, Jordanstown, Northern Ireland
²Diabetes Specialist Care, Ulster Hospital, Dundonald, Northern Ireland
{l.a.black,mf.mctear,nd.black}@ulster.ac.uk

Abstract

DI@L-log is an automated medical spoken dialogue system designed to enable patients to regularly communicate health data to the point-of-care over the telephone. In order to investigate the performance of the system, a preliminary evaluation was conducted using 5 novice hypertensive diabetic patients from the Ulster Hospital in Northern Ireland. The purpose of the study was to assess several factors which need to be addressed when designing spoken dialogue systems for elderly, disabled users. We examined the performance of the system, the interaction preferences of the user and their usage patterns, the level of user satisfaction, and the impact of the system on the patient’s health over time. A revised version of the system, modified to take account of findings from this study, is currently undergoing testing with a larger group of subjects.

1. Introduction

Spoken dialogue systems provide access to services and information that have traditionally only been available to people with personal computers and internet access [1]. This paper presents results from a study that examined the use of a spoken dialogue system, known as DI@L-log [2], by elderly, infirm, hypertensive diabetes patients. The system was designed to take account of the difficulties that these patients might experience with an automated system and to provide support and feedback [3]. DI@L-log is a task-oriented system with a female persona that plays a role similar to a nurse. The system acts as an automated “voice logbook” that replaces the more passive paper logbook currently used by diabetes patients to record home monitored data. The system collects and analyses the patient data and communicates the results automatically to doctors at the hospital, alerting them to any abnormal developments and thus enabling them to manage and reduce complications associated with the disease more effectively.

2. Overview of the DI@L-log system

DI@L-log consists of a voice user interface, voice browser, web server, backend database, and graphical user interface. The system is written in VoiceXML using Voxpilot’s online development environment (Voxbuilder) and is hosted on the Voxpilot Website [4]. The voice browser, located in Monaco, acts as a gateway to the backend system to which it sends data as SQL queries using the Apache Tomcat Web server and JSP. Figure 1 illustrates the components of the system and the call management. When the patient places a call from their own home in the UK, it is transferred to the VXML interpreter in Monaco. A request is then submitted back to the DI@L-log files residing on the Ulster Hospital server in the UK.

DI@L-log enables patients to speak their vital signs health data (weight, blood sugar and blood pressure) using speech or DTMF (telephone keypad). Data is retrieved from the backend system during a call and at the end of the call the database is updated with the newly elicited readings. The system provides feedback by comparing the new data with previous call and target readings.

The graphical user interface, which is not the focus of the present paper, allows patient trends and alerts generated by the system to be viewed and manipulated by doctors at the hospital if individual or multiple readings were regarded as too high. This decision support feature is modelled on medical protocols for blood glucose, blood pressure, and the individual’s weight target, based on the body mass index (BMI). The decision support component aims to act as an intervention strategy in order to mitigate complications associated with diabetes and co-existing hypertension.

The following is an extract from an interaction between a patient and the system:

System: Welcome to the Dialogue Interactive Service. You can say help or exit at any time. Please now enter your four digit PIN.
Patient: One nine seven five
System: If your number is 1 9 7 5 say "yes", or press "1". To change, simply say "no," or press 2.
Patient: Yes
System: One moment. I’m looking up your details… Ok Joe, are you ready to give your measurements?
Patient: Yes
System: If your number is 1 9 7 5 say "yes", or press "1". To change, simply say "no," or press 2.
Patient: Yes
System: If your number is 1 9 7 5 say "yes", or press "1". To change, simply say "no," or press 2.
Patient: Yes
System: One moment. I’m looking up your details… Ok Joe, are you ready to give your measurements?
Patient: Yes
System: Thank you. What is your weight today?
Patient: I’m twelve stone and 6 pounds.
System: Please confirm your weight is twelve stone and six pounds. Say "yes", or press 1. Otherwise, say "no" or press 2.
Patient: Yes
System: Super. And what is your blood sugar?
Patient: Six point two
System: Did you say your blood sugar is six point two?
Patient: Yes
3. Design of the study

The study involved 5 male diabetes patients (mean age = 64.2; age range 57-70) who have essential or suspected 'white coat' hypertension. These patients are currently registered and treated at the Diabetes Specialist Clinic at the Ulster Hospital in Northern Ireland (UCHT). Patients were asked to contact DI@L-log once per week (every Monday) from their own homes for the duration of 8 weeks. All patients are native English speakers. Two additional health professionals also participated in the study acting as control subjects; however for the purposes of this paper their data is not included. Patients were supplied with blood sugar and blood pressure meters, courtesy of industrial sponsors. They were shown how to use the upper arm blood pressure monitor by a diabetes specialist nurse. A training session was established to provide patients with an overview of the system. To reduce cognitive load, a large-print help sheet with a sample document readings when contacting the system.

In order to examine the relationship between interaction modes and system performance, clients were asked to vary input modes by using speech, DTMF, or both (mixed modality) over the course of the study. Each DI@L-log call required the patient to log in to the system using a 4 digit PIN and then to submit and confirm their weight, blood sugar and blood pressure information in response to system prompts. Patients received spoken feedback regarding their health status in real-time at the end of the call.

In the current system, patients could use fairly free natural language phrases and sentences to speak their data. For example, weight could be expressed in a variety of ways, such as (optional items in brackets):

(Today) (my) (weight) (is) fourteen two.
(It’s) fourteen (point) two.
(I) (weigh) fourteen (stones) (and) two (pounds).
(I’m) fourteen stone.

Grammars were devised for each of the 3 measures to permit a range of natural input such as this.

4. Quantitative Analysis

4.1. Usage patterns and interaction modes

Statistics for usage patterns and interaction modes were derived from log files generated by the host platform (Voxpilot). A total of 42 calls were received, of which 38 (90.4%) were successfully completed. The 4 unsuccessful calls resulted from 2 hang-ups by users and 2 cases of bad fetch errors that caused the system to terminate the call. The main usage results are presented in Table 1.

As can be seen from Table 1, calls using speech took longer than those using DTMF, due to the occurrence of speech recognition errors and of failures to detect the user’s spoken input. Notwithstanding these problems, patients used speech more often than they used DTMF (52.6% compared with 36.8%), with the remaining usage being in mixed mode (speech and DTMF).

Indeed, it is interesting to note usage patterns over time, as patients tended to use speech increasingly over DTMF as they became more familiar with the system. Usage patterns for each patient are shown in Table 2 (NC = no call made).

<table>
<thead>
<tr>
<th>ID/ Week</th>
<th>1212</th>
<th>1418</th>
<th>4532</th>
<th>5567</th>
<th>7686</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DTMF</td>
<td>DTMF</td>
<td>DTMF</td>
<td>Voice</td>
<td>DTMF</td>
</tr>
<tr>
<td>2</td>
<td>Voice</td>
<td>DTMF</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
</tr>
<tr>
<td>3</td>
<td>Voice</td>
<td>DTMF</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
</tr>
<tr>
<td>4</td>
<td>Voice</td>
<td>DTMF</td>
<td>Voice</td>
<td>DTMF</td>
<td>Voice</td>
</tr>
<tr>
<td>5</td>
<td>NC</td>
<td>Mixed</td>
<td>DTMF</td>
<td>Voice</td>
<td>Voice</td>
</tr>
<tr>
<td>6</td>
<td>DTMF</td>
<td>Voice</td>
<td>Mixed</td>
<td>DTMF</td>
<td>DTMF</td>
</tr>
<tr>
<td>7</td>
<td>Mixed</td>
<td>DTMF</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
</tr>
<tr>
<td>8</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
<td>Voice</td>
</tr>
</tbody>
</table>

4.2. System performance

The most relevant feature of system performance was speech recognition accuracy. Recognition errors were identified by comparing the recognized results, as presented in the system.
log files, with audio files that recorded the patients’ spoken input. Overall word accuracy was 74.6%.

Given that there were 4 main inputs (not including confirmations using “yes” or “no”), and that the inputs all involved some form of numerical data, accompanied in some cases by additional words, it is beneficial to examine word accuracy rates across the different tasks.

Task 1: PIN Spoken input of the PIN involved a 4 digit string. Recognition accuracy was 100% for both speech and DTMF. The high rate of speech recognition accuracy for this measure is due to the more constrained built-in digit grammar used to collect the PIN value.

Task 2: Weight. As indicated earlier, in addition to the specification of stones and pounds, which were the required input elements, users could also speak extraneous words in order to make their speech feel more natural. The grammars were written in such a way as to return only the two required values. Recognition accuracy was 77.8% using speech.

Task 3: Blood sugar. This value consisted of a whole number or a whole number and a decimal number (for example, six point two). Recognition accuracy using speech was 50.3%. This was the most difficult parameter, due largely to ‘nomatch’ events involving the word “point”, which was pronounced by some speakers to sound more like “pint”.

Task 4: Blood pressure also caused several problems because of the multiple numbers (up to six) to be recognised in sequence, for example “one hundred and fifty five over seventy two”. This resulted in a 57.7% accuracy rate for speech. DI@L-log automatically partitioned these values into subcomponents if recognition was not achieved on the first attempt as highlighted in the dialogue in Section 2. If only one of the numbers was submitted or if the system misrecognised the complete input, the system would enter a subdialog and re-prompt saying either “Sorry I need two numbers” or “Let’s take this step by step… Firstly, what is the top number”, followed by “And what is the lower number?”

For this task some patients had problems when entering with DTMF. In particular, one patient would include the star key when the system had entered the subdialog (using DTMF “*78” for diastolic pressure instead of just entering “78”) and this would not be recognised by the system.

Other problems included “speech too early” errors (11 times), which then led to hyperarticulation, thus adding to the problems of accurate speech recognition.

5. User satisfaction analysis

Qualitative analysis of user satisfaction was based on a 20-item, 5 point Likert scaled questionnaire that was completed by the patients at the end of the study. The results of the qualitative analysis are shown in Table 3.

The findings in Table 3 indicate that users were not intimidated or lost in their interaction with the system – they responded positively to the voice, its pace, the level of personalisation, choice of modalities, and the call duration. The system scored a moderate 3.3 average rating, mainly due to problems with speech recognition. Over time, patients felt the system became easier to use, forming a mental model of the navigation and claiming that they knew what DI@L-log was going to say next. Patients also said that the system had distinct advantages over the paper diary (efficiency, regular contact) and all users reported that they became more aware of their health by using the system. In addition, participants agreed that the system generated a sense of care and interestingly only one patient said that they would prefer to talk to a real medical expert, and this was only in the situation when they needed particular assurance or advice. When asked about the feedback, patients felt the system could improve advice by setting more realistic user targets. One user would have liked the system to give more tailored advice on dietary issues specific to his needs by creating some sort of metabolic profile. Another said it would be reassuring if the system linked to a live agent when repeated readings triggered a warning. In other aspects of the feedback, patients appreciated the acknowledgment of increases or decreases in measurements since the previous call, again creating the sense of care and personalisation.

Table 3: User satisfaction analysis

<table>
<thead>
<tr>
<th>Factor</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriateness of female voice</td>
<td>3.6</td>
</tr>
<tr>
<td>Pace of voice appropriate</td>
<td>4.2</td>
</tr>
<tr>
<td>Length of call too long</td>
<td>2.6</td>
</tr>
<tr>
<td>Learnt more about system over time</td>
<td>3.6</td>
</tr>
<tr>
<td>System easier to use over time</td>
<td>4</td>
</tr>
<tr>
<td>Voice preferred modality</td>
<td>1.8</td>
</tr>
<tr>
<td>Too many interactions in the call</td>
<td>2.2</td>
</tr>
<tr>
<td>Prefer more dynamic feedback</td>
<td>2.8</td>
</tr>
<tr>
<td>Advantages over paper log book</td>
<td>4</td>
</tr>
<tr>
<td>Increased health awareness</td>
<td>5</td>
</tr>
<tr>
<td>Felt lost (navigation)</td>
<td>1.2</td>
</tr>
<tr>
<td>Intimidated by system</td>
<td>1.8</td>
</tr>
<tr>
<td>Surprised by system actions</td>
<td>2.2</td>
</tr>
<tr>
<td>Speech recognition difficult</td>
<td>4.4</td>
</tr>
<tr>
<td>System was frustrating</td>
<td>3.6</td>
</tr>
<tr>
<td>Dual choice of interaction beneficial</td>
<td>4</td>
</tr>
<tr>
<td>Level of personalisation good</td>
<td>4.6</td>
</tr>
<tr>
<td>System generated a sense of care</td>
<td>4.4</td>
</tr>
<tr>
<td>Prefer to speak to a real medic</td>
<td>2.8</td>
</tr>
<tr>
<td>Future benefit for chronic patients</td>
<td>4.4</td>
</tr>
<tr>
<td>Average system rating</td>
<td>3.3</td>
</tr>
</tbody>
</table>

6. Discussion

The conversational design of the version of DI@L-log used in this study was based on discussions with health professionals at the UCHT. The rationale for the more open-ended grammars for the weight, blood sugar and blood pressure measures was that this would give the impression of a friendly, conversational system. However, as the results indicate, the more open-ended input permitted by these grammars led to an unacceptable recognition performance. Accordingly the next version of the system has been updated with more restricted grammars to ensure that user input is more likely to be recognized on the first attempt. These design trade offs are important when considering the user audience, who are elderly, and infirm.

During the trial, only one patient asked DI@L-log for help, so we have embedded help within the error recovery...
loop. We also added mental cues and explicit statements which inform the user of what the system has heard, rather than simply playing a no-match prompt. For example:

"I'm sorry but that reading of 0.3 is not a valid input for your blood sugar reading. This value must be greater than 1. Please try again."

In addition, timeouts were increased to accommodate different interaction speeds required by the user audience who may take slightly longer to respond to prompts than younger users.

Regarding the comparison between speech and DTMF for input of numerical data, it was found that more turns were required when using speech (on average, 2.0 turns per task) compared with DTMF (average 1.1 turns per task). However, as can be seen from the call duration times reported in Table 2, the actual time required for speech input was not proportionately greater than the time required for DTMF input (average call duration using speech: 3m 35s as opposed to 3m 01s for DTMF). Indeed, mixed mode using a combination of speech and DTMF resulted in an average call duration of 3m 45s. This suggests that, for this cohort of elderly patients, DTMF input was less convenient due to problems of manipulating the telephone keypad and that speech would be a more rapid and more convenient method for the input of data if the recognition accuracy could be improved. This was shown in all 5 of the patient’s call completion times which displayed progressively quicker interaction times (on average 1m 5s) at the end of the trial than at the start. In any case, as seen from the usage patterns reported in Table 2, users tended to use speech rather than DTMF as they became more familiar with the system.

Looking at the results of the user satisfaction analysis (Table 3), users generally found speech recognition to be problematic and did not choose speech as their preferred modality, preferring the option of mixed mode input. It will be interesting to note whether better results for speech input will be obtained in a subsequent evaluation using the more constrained grammars and prompts.

7. Conclusions and future work

This paper has presented a preliminary evaluation of the DI@L-log system with a small group of patients. The findings from this study will be useful for future development of the system.

It is encouraging that the users emphasized the usefulness of the system over the current logbook used for home monitoring. There are indications that the system will provide benefits in terms of the healthcare of patients. Users remarked that the system had created a greater sense of care and that it made them more aware of their health. For example, in the current study 3 warnings about blood sugar levels and 1 warning about blood pressure were generated and presented to the patients concerned. The system also generated 1 blood pressure alert for the attention of the consultant. Measures such as these could play a huge role in reducing hospitalisation by facilitating earlier intervention to avoid complications.

The results also confirmed the findings from other studies (such as [3]) that spoken dialogue systems need to be carefully designed to take account of the difficulties that elderly and infirm users might have with new technologies, for example, with speaking to an automated system. The results from the current study indicate that a more constrained system with simple well-defined prompts and tasks is more likely to achieve satisfactory recognition rates, leading to greater transaction success as well as enhanced user satisfaction.

The next stage of the research is to evaluate a revised version of the system with a larger cohort of patients. The revised system makes the input of the data simpler. For example, to elicit the patient’s weight, the system first asks for the stones value before going on to ask for the pounds value. A built-in digit grammar is used to collect these values. Similarly, for blood pressure the values are divided into top value and bottom value and elicited separately using a built-in number grammar. Along with more explicit instructions it is expected that recognition rates with the revised system will improve. Initial results indicate that these patients have been interacting predominantly using speech with far greater success levels due to the more restrictive nature of the grammars and a more careful prompt design, with the main sacrifice being the system’s conversational style.

However, it is also possible that as patients become familiar with the system then they will not require more detailed instructions and will also wish to enter their data by combining values – for example, by saying both elements of their weight and both elements of their blood pressure within single utterances. A longer term goal of the research would be to develop a system that adapts automatically to the user over time in these respects, and that would require fewer prompts as users become more familiar with the system, or would make modifications if users have problems with particular aspects of the interaction.

Finally, a scheduled call-reminder service would be an appropriate extended feature of the system for the elderly who may, from time-to-time, forget to call the system.

8. References


9. Acknowledgements

We gratefully acknowledge sponsorship from the Health Informatics Society of Ireland (HISI), industrial partners Voxpilot, Omron UK, Lifescan (J&J). Warm thanks to Conor McMeel, John Kennedy, Paula Smeele and the patients and staff at the UCHT for their invaluable assistance in development of this project.