

Evaluation and Usage Patterns in the Homey Hypertension Management Dialog System

Toni Giorgino, MSc, Silvana Quaglini, PhD, Mario Stefanelli, PhD

Laboratorio di Informatica Medica, Università di Pavia, Pavia, Italy
{toni,sil,mstefa}@aim.unipv.it

Abstract

The Homey dialog system has been developed to supplement face-to-face encounters in the process of care of chronic diseases: the current prototype targets the care of Hypertension and integrates a general-purpose electronic patient record (EPR) and a state-of-the art dialogue engine. Patients report their health condition from their home via an intelligent dialog system (IDS), and care-giving professionals can receive timely information about their state, right at the point of care. Homey is currently ongoing clinical evaluation. This paper provides an overview of the system, and presents preliminary results obtained from the usage patterns collected in the ongoing clinical trial. Finally, conclusions are drawn on the utility and prospective uses of the data collected.

Introduction

The telephone has been employed in medicine since its appearance, and recently it is being integrated with more advanced telecommunication tools. For instance, Dual Tone Multi Frequency (DTMF) systems have been proposed and used as an alternative and supplement to a visit performed by a physician. The data that can be entered in DTMF systems is limited to numeric quantities or coded information; nevertheless, such systems turned out to be successful in home monitoring patients with chronic diseases, like hypertension (1).

Automatic speech recognition

The natural evolution of DTMF systems is *spoken* dialogue systems, which have been made possible by advances of research in the field of automatic speech recognition. Only recently, the *automatic* recognition of speech has reached an acceptable degree of reliability (e.g., fraction of words correctly recognized). Still, trying to have a computer able to extract words from a spoken audio segment yields results which are significantly different whether the recognition program can be previously trained on the particular speaker (*speaker dependent* technology) or not, as is in the case of telephone users (*speaker-independent*).

As soon as the computer recognizes a sentence, it must process it and extract the semantically relevant content, isolating other possible random words or noises. An appropriate answer, probably as well as the next question, is then constructed by a *dialogue manager*, and the conversation goes on. The computer's response should of course be appropriate to the user's intentions to be perceived as "intelligent" and to give her the feeling that the machine is "in control". The resulting system is called an Intelligent Dialog System (IDS).

Home monitoring

Hypertension is a chronic disease whose management requires diligence in monitoring the blood pressure values; it is also important that the patient follows the prescribed therapy, unless she is affected by side effects, which need to be reported as soon as possible. Thus, any new technology making data collection easier and faster may potentially increase the quality of the care of patients with hypertension.

Visiting a physician's office is the most common way of receiving health care for the majority of patients. Although effective, traveling to see the doctor may be inconvenient and costly for the treatment of some chronic diseases. Patients affected by essential hypertension, for example, should be monitored frequently to follow their blood pressure values, heart rate and physical activity, presence of side effects, etc.; the values recorded are taken into account by the physician, which decides whether to modify or suspend the therapy. Recent clinical trials have shown that home monitoring actually reduces mean blood pressure values with respect to usual care (5). Remote monitoring via telemedicine may possibly grant both cost savings and better quality of life.

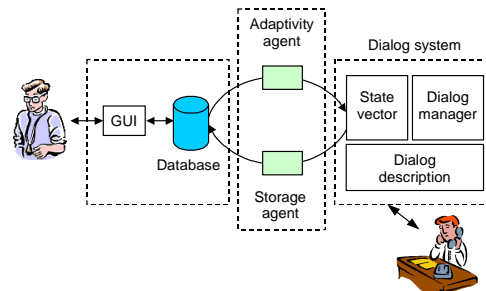


Figure 1: Architecture overview

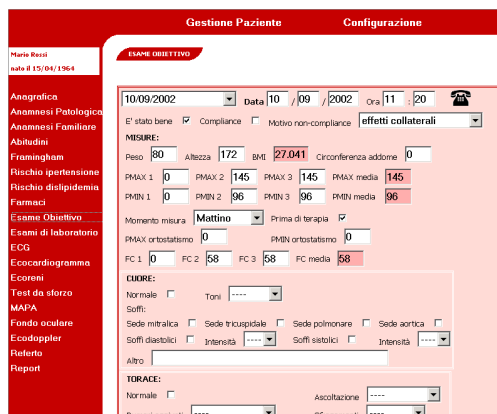


Figure 2: The graphical interface as shown to the physician. A telephone icon, which appears top-right, reminds that the data displayed were entered by the patient via the spoken dialog system.

The Homey Project

Within the E.U. research project “Homey”, we built an intelligent dialogue system to manage and monitor patients with essential hypertension. Several Italian hospitals were involved in the design of the system and its deployment and adoption. We tried to keep the dialogue structure close to the usual interaction between a physician and a patient; its scope is to acquire health status (measures, presence of side effects) and to give advice, by issuing alerts and prompts. The domain knowledge was derived from a set of world-widely accepted guidelines for the hypertension and dyslipidemia (2).

Architecture

The dialog system is interfaced to the electronic health record (EHR) of each monitored patient. Two actors are therefore allowed to enter data into the EHR. In the first place, the physician may use a conventional (web-based, keyboard and mouse) interface to store and update patient information. On the other hand, the patient is also allowed to enter the self-measured data by the means of a telephone. Figure 1 illustrates the data flow inside the application. Patients periodically call a dedicated telephone number and engage a dialogue with the system, which talks and interacts with them to acquire clinical data, monitor their style of life and ask about the occurrence of possible side effects.

Values entered by the patient are distinguished by those taken during encounters by a special icon which appears near them on the graphical interface. Data are entered in tables through a graphic user interface, designed keeping into account the suggestions of the clinicians (fig-

ure 2). The database holds several values that will affect the dialog (e.g.: whether the patient has been prescribed to follow a diet, or the date of the next visit). When a new call is set up for a specific patient, values are extracted from the database and the corresponding call is prepared. The software that performs this task is called “adaptation agent”; part of the adaptations are in fact performed by the agent to try to make the dialog more effective and friendly.

Evaluation

Despite the care put into the design of a dialogue application, questions and grammars, the fact that untrained users interact with a complex system poses remarkable usability challenges. The dialog application should be easy to use and understand, and robust. A number of issues in the design of our application were therefore analyzed by the means of on-field trials. The objectives of the evaluation phases were:

Testing the reliability of the system – Recognition errors, although annoying, should not affect the user’s ability of proceeding in the remainder of the dialog.

Extension of grammars and lexicon – Grammars should capture most of users’ answer schemes.

Reformulation of questions’ wording – Users’ answers wording is influenced by how questions are asked.

Extraction of patient’s learning curve – To address adaptability issues, one desires to put each patient into an ability class. It is then possible to study how quickly users learn to use the system, so they can be hinted about more advanced features, like mixed-initiative.

Assessing the clinical effectiveness – The ultimate goal of the project is both to raise patient compliance with the guideline and to expand the availability of data for the use of the clinician. Collecting quantitative information involving real patients may assess whether the system helps to achieve a better quality of life.

Internal trial

To debug our dialog application according to the goals listed, we designed an evaluation plan to happen in two phases. The first phase (*internal trial*) involved a group

Start date / time	
Call duration	Number of prompt sessions
Whether user was anonymous	Time spent in recognition mode
Patient code	Time user actually spent speaking
List of concepts acquired	Time for pronouncing prompts
Number of fields acquired	Whether call reached bye message
how many were unique	How many times help was requested
Number of recognition sessions	How many rejected utterances
how many were confirmation questions	How many times no answer detected
how many got answer YES	How many errors
how many got answer NO	How many DTMF digits pressed

Figure 3: list of low-level per-dialogue measures automatically computed.

of volunteers, which were assigned a realistic disease profile. They were asked to call the system at fixed schedules, pretending to be hypertensive patients. Profile assigned included therapy, average blood pressure values and so on; the profile, in turn, affected the inquiry of side effects during the dialogue. The volunteers were asked to annotate and report problems and obstacles found. The call collection ended after every user performed the assigned number of calls and therefore addressed technical and usability issues. It involved approximately 15 people and collected 150 dialogues, amounting to about 500 minutes of conversation, of which 150 were human speech, the rest synthesized by the text-to-speech component (TTS).

Data collection architecture

The procedure designed to collect and analyze data coming from the real users' calls keeps into account two sources of information: (a) call logs and (b) clinical information about the patients. The former is gathered from information generated by the computer telephony platform while it is running. By post-processing the logs, it is possible to automatically gather the detailed per-call measures listed in figure 3.

The other valuable source of information is the EHR itself, where a record is made every time a patient uses the voice system to enter his data, or doctors login via the web to store outcomes of real encounters. On the contrary to the telephony logs, information in the EHR has a coarser granularity as it only stores confirmed values, and not, for exam-

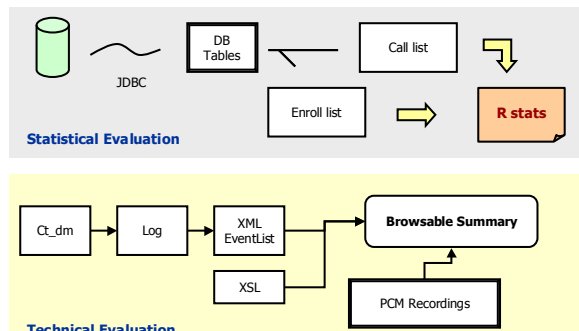


Figure 4: Data collection sources

ple, misrecognitions, repetitions and other phenomena, which are dealt with within the dialogue system.

The availability of detailed records of numeric quantities such as those in figure 3 is mostly useful for the assessment of system quality. To start, one could monitor the fraction of confirmation questions which received negative answers, discovering possible relation to changes made to the system or even demographic data. The most exiting prospective use, related to this one, is to predict the behavior of the system with particular users and change dialogue strategies accordingly. Given the availability of call data, one could think of applying knowledge discovery techniques to this task. Such techniques would be most useful if an independent rating of the dialogue quality is available, such as a measure of "successfulness" obtained subjectively by a human rater, which reviewed the recorded conversation. The authors are not aware of current standard for quantifying such ratings; therefore, further

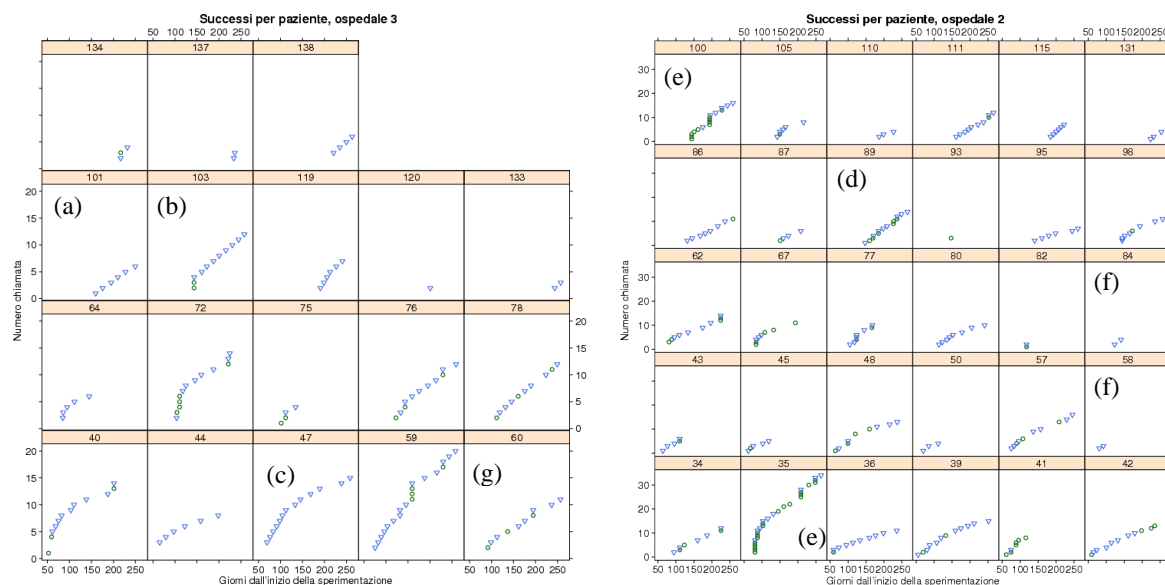


Figure 5: call patterns for two hospitals. Patient ID is indicated at the top of each box. Day of the call (zero being start of the clinical trial) is on the x-axis. On the y-axis the cumulative number of calls is shown for that patient. Dialogue sessions reaching the final salutation message are plotted as triangles, otherwise as circles.

investigation on this prospective technique is ongoing.

Clinical trial: call diagrams

After the internal evaluation was completed, the dialogue description was finalized and deployed for the clinical trial, which started on August 2003.

Figure 5 provides one view on the patterns of use of the telephone system over time for a subset of the patients enrolled into the system in the treatment group. Each patient is shown in its box. A dot is placed on the plane for each call; the x coordinate of the point relates to the day in which the call was received (day zero being the begin of the trial). The vertical axis displays the cumulative number of calls. Points have a different shape to indicate whether the dialogue reached the final salutation message (triangle) or not (circle). This representation was chosen as a first attempt to distinguish conversations that ended properly from those which did so prematurely. Patient enrollment is still in progress, therefore not all “first calls” appear close to the date at the beginning of the trial.

Patterns in these plots can be recognized by inspection. Call diagrams like (a) correspond to people that regularly call the system and interact successfully with it. They show dominantly “good-end” calls. The slope of curves (a) and (b) is approximately one call every two weeks. Plot (b) shows that there has been a “training period”, during which patient 103 was unable to successfully complete two dialogues. The patient called back again the same day, and at the third attempt succeeded in entering the values; subsequent calls were performed smoothly and on schedule. Curve (c) shows that the respective patient followed the physician’s direction to call more frequently, once a week, during the first two months of the system’s use, then to reduce the number of calls to every other week.

People which call the system regularly, but experience usage problems, may show up as in diagrams (d) and (e). In the former, one finds a large fraction of “circle” symbols, but the curve does not have remarkable jumps: this means that on the average the call frequency was kept as required, despite of the outcome of specific calls. As a remark, the fact that a call is shown as a circle does not mean that the call was unsuccessful at acquiring values, but merely that the bye-message was not played: the more important clinical data were nevertheless recorded in the database. This is seen, for example, in (g), whose user from time to time hangs up the conversation after the data acquisition phase. Plots like (e) show multiple attempts done in a row in the same day, or close dates: they appear as rising “steps” of terminated calls.

Patients that regrettably stop using the system altogether are also quite easily distinguished, as in (f).

It is remarkable to note that not all quitters have a record of unsuccessful previous calls.

Conclusions

This paper described an infrastructure for monitoring hypertensive patients. The system keeps an EHR of the managed patients. Homey has been built for two main purposes: first, to develop a system to acquire data from chronic patients and to assist physicians in charge in doing informed therapeutic decision; second, to study the dialogue adaptation techniques that it is possible to deploy in this domain. The dialogue adaptation task needs to be performed based on the data that can be gathered during telephone calls. Post-hoc data review is also important because allows one to detect per-patient usage patterns and take appropriate action per time.

This research has been partially funded by EU Fifth Framework IST project number IST-2001-32434.

References

- (1) Friedman RH, Stollerman JE, Mahoney DM, Rozenblyum L. The virtual visit: using telecommunications technology to take care of patients. *J Am Med Inform Assoc* 1997; 4(6):413-425.
- (2) 1999 World Health Organization-International Society of Hypertension Guidelines for the Management of Hypertension. Guidelines Subcommittee. *J Hypertens* 1999; 17(2):151-183.
- (3) Mulrow CD. Evidence-based Hypertension. BMJ Books, 2001.
- (4) Rogers MA, Small D, Buchan DA, Butch CA, Stewart CM, Krenzer BE et al. Home monitoring service improves mean arterial pressure in patients with essential hypertension. A randomized, controlled trial. *Ann Intern Med* 2001; 134(11):1024-1032.
- (5) Falavigna D, Gretter R, Orlandi M. A mixed language model for a dialogue system over the telephone. Proceedings of ICSLP 2000; Beijing, China: 16 October 2000.
- (6) Azzini I, Orlandi M, Lanzola G, Gretter R, Falavigna D. First steps toward an adaptive spoken dialogue system in medical domain.: EURO-SPEECH 2001, Aalborg, Denmark, September 2001., September 2001.
- (7) Rabiner L. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc IEEE* 1989; 77(2):257-286.

