Automated Spoken Dialog System for Hypertensive Patient Home Management

Toni Giorgino\textsuperscript{a}, Ivano Azzini\textsuperscript{a}, Carla Rognoni\textsuperscript{a}, Silvana Quaglini\textsuperscript{a}, Mario Stefanelli\textsuperscript{b}, Roberto Gretter\textsuperscript{b} and Daniele Falavigna\textsuperscript{b}

\textsuperscript{a}Laboratory of Medical Informatics, Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy
\textsuperscript{b}Istituto per la Ricerca Scientifica e Tecnologica ITC – irst, Panté di Povo, Trento, Italy

Abstract

Recent advances in automatic speech recognition and related technologies allow computers to carry on conversations by telephone. We developed an intelligent dialog system that interacts with hypertensive patients to collect data about their health status. Patients thus avoid the inconvenience of traveling for frequent face to face visits to monitor the clinical variables they can easily measure at home; the physician is facilitated in acquiring patient information and cardiovascular risk, which is evaluated from the data according to noted guidelines. Controlled trials to assess the clinical efficacy are under way.

Key words: Data Collection; Hypertension; Medical Records Systems, Computerized; Patient Education; Telemedicine

1 Introduction

Hypertension is a chronic disease whose management requires a careful monitoring of blood pressure values. It is estimated that prevalence of hypertension in the general population is as high as 44\% in Europe and 28\% in North America [1]. Uncontrolled high blood pressure does not have specific clinical manifestations until organ damage develops [2]; it is therefore important that patients follow the prescribed therapy, unless they are affected by side effects, which need to be reported as soon as possible [3].

On the other hand, recent evidence [4] suggests that careful hypertension care and control produces behavioral changes and physiologic improvements...
in the prospective outcome. It is therefore not preposterous to conjecture that new technology, making data collection easier and faster, may increase the effectiveness of hypertensive patients’ care [5].

Recently, Dual Tone Multi Frequency (DTMF) based systems have been used as a supplement to patient-physician traditional face to face encounters. When a user dials a DTMF-based system, she listens a series of pre-recorded audio messages, which, at each step, prompt her to type some data on the telephone’s keypad. The data that can be entered are necessarily limited to numeric quantities or codes and navigation is usually restricted to a tree-like structure. Despite this somewhat cumbersome usage, controlled studies showed such DTMF systems to be successful in home monitoring patients with chronic diseases, like hypertension [6,7].

Within the EU-funded research project Homey we built an intelligent dialog system to monitor and manage patients with essential hypertension. We tried to keep the dialogue as close as possible to the usual patient-physician interaction; it also gives advice on recommended health behaviour and next scheduled visits, by issuing alerts and prompts. The domain medical knowledge is represented according to world-wide accepted guidelines for the management of hypertension and dyslipidemia [8,9]. The data acquisition process may develop both during the traditional encounter with the physician, and via an automatic dialog engine that talks to the patient through the phone. Information acquired is stored in a purposely-designed Electronic Health Record (EHR), whence it can be queried via a web interface.

We have involved into the project some Italian hospitals willing to cooperate in the design and development of the system. The monitoring service they provide normally requires each patient to travel to the outpatient clinic and meet her physician approximately every 2–4 months. Each encounter is used to monitor blood pressure values, habits and other clinical variables, and to make therapy adjustments. The physician takes into account these results in order to prescribe or modify the pharmacological therapy and to possibly prompt the patient to make changes to her lifestyle.

2 Architecture

A simplified data flow inside the dialog system application is shown in Figure 1. Two actors are allowed to enter data into the EHR. In the first place, the physician uses a conventional (graphics, keyboard and mouse) interface to enter patient information. On the other hand, the patient is also allowed to enter via the telephone the data that she can acquire by herself at home and send them to the care-providing centre.
To accomplish this task, the patient is requested to dial a toll free number once every one or two weeks. She is connected to a call center and interacts with the automated spoken dialogue system for the management of hypertensive patients. The system interprets a *dialog description* written in a high level language, which specifies the spoken interaction: which steps to take, what questions are to be asked to the user and what possible utterances are allowed in reply. The user is first authenticated with a numeric password. Then, the dialog engine interprets a state vector which represents her health state and other associated information. The role of the state vector as information container for each patient will be explained in more detail in section 2.4.

The system presents the caller with a series of questions (also known as *prompts*), dealing with specific aspects of her health state and behaviour. The questions are organized by topic in “contexts”. The contexts implemented in the current version, are listed in Table 1. It is worthwhile noting that not all the contexts and questions are activated at each call. In fact, the adaptivity features, which will be described later, modify the flow of the call depending on the specific patient, tailoring it to the answers to previous questions and the data recorded in the EHR.

Speech recognition and dialogue interpretation technologies we used in our application are provided by the SPINET system (Speech Into Enriched Text). The system, developed by the ITC-irst research institute, provides both the engine which performs the actual recognition on the sampled audio signal, and the dialog manager [10]. The system speaks to the caller via a synthetic voice (which does not suffer the limitation of prerecorded prompts, restricted to fixed sentences) which is created at run time by a commercial text-to-speech software.

### 2.1 Dialogue initiative

The dialog system we developed leverages two advances in dialog technology, as it is *mixed initiative* and *adaptive*.
The most simple variety of spoken interaction with a system is the *form filling*, in which the speaker is only allowed to answer questions posed by the system, one at a time, as if the user is filling a form by dictating fields’ content on the telephone, as in the following example:

**System:** Please tell me the heart rate you have measured today.  
**Patient:** 90.  
**System:** Now, please, say your systolic arterial pressure.  
**Patient:** My pressure is 120.

*Mixed-initiative* dialog systems instead try to mimic the behavior of a telephone answerer, whose counterpart may anticipate answers to questions that have not been formulated yet. Mixed initiative dialogues are thus different from the simpler form filling interactions. When talking to a mixed initiative system, the user has a greater degree of control in the progression of the dialogue. At certain points, she may answer to a question only partially, e.g., she is requested two values, but chooses to give only one, or she may want to anticipate some answer, providing data that was not yet asked, as in:

**System:** Please tell me your heart rate.  
**Patient:** I have not measured it, but today my weight is almost 85 kilograms.  
**System:** So, your heart rate is unreported, and your weight is 85 kilograms?  
**Patient:** Yes.

It is apparent that in a mixed initiative system, parts of users’ utterances are to be interpreted textually as well as *semantically*: in the example above the system needs to understand that the spoken number “85” has the meaning of a weight and not a heart rate.

The mixed initiative approach may be pushed to an extreme and leave the speaker free, in principle, to give the information she desires in any order [11,12]. This approach, although exciting, has drawbacks for a data collection application. For instance, such an open prompt may give a naive caller the

![Figure 1. Architecture overview: data flow (left) and block diagram (right)](image-url)
false impression that the system will understand anything she would say, which
is not true. The user would then lose focus from to the task to be completed,
i.e. the acquisition of a limited number of important, self-measured clinical
parameters.

2.2 Automatic Speech Recognition and Grammars

At the core of the natural dialog system stays an Automatic Speech Recogni-
tion (ASR) module. An ASR system (please refer to diagram in figure 2)
is a software designed to convert a given sound signal (1), digitally sampled
(2), into an equivalent textual representation (3). More precisely, the ASR
tries to find the closest match between the sentence said and a set of alter-
natives, described in a purposely-designed grammar [13]. Various grammars
are activated during different stages of the dialogue, as required by the dia-
logue context. The activation and deactivation of grammars happens under
the control of the dialog engine (4), which elaborates the recognition results
and maintains the state of the dialogue.

Given a spoken input, grammars restrict the hypothesis space in which the
search for matches is carried out. Choosing the right search space is essential,
not only for limiting the computational effort required to find the match (which
has to take place in real time), but also – more importantly – because one wants
to reduce the chance of a misrecognition. Grammars are somewhat analogous
to regular expression patterns – flexible specifications which the computer tries
to match to an input. It should be noted, however, that a regular expression
match is deterministic, while the match found by ASR gives a probabilistic
measure of similarity, which is usually evaluated by complex Hidden Markov
Models [14].

Grammars are usually represented by means of directed graphs whose arcs
 correspond either to terminal symbols (i.e. words) or to non-terminal symbols
(i.e. grammars themselves) [15]. In practice, they are stored inside the com-
puter in a textual representation similar to the example shown in figure 3.
While ASRs will in principle match an utterance only if the word sequence may be expressed by the grammar, a less restrictive approach, especially useful for complex sentences, can be often employed: it foresees the use of bigram models with a statistical description of the allowed sentences.

Grammars and bigrams are developed to drive an ASR in each particular application, and are called its language model. Figure 4 summarizes the relationships between the input, dialog engine, grammars and speech recognition components.

2.3 The Electronic Health Record

A centralized, comprehensive EHR is the backbone of the system. It has the function of storing both the data entered by physicians during visits and those spoken on the phone by patients themselves; it will then generate reports that can be reviewed by physicians. Reports include cardiovascular risk indicators, evaluated according to clinical guidelines [8,9], and alerts, which could be issued graphically or possibly via other means such as Short Message Service (SMS) delivered on mobile phones or pagers.

The user interface to the database was designed with the collaboration of physicians (Figure 5). The interface is web-based, therefore health professionals and patients can access it with standard web browsers from any convenient location. Authentication, access roles and cryptographic mechanisms for web-based applications have been applied to the security and privacy of the data as mandated by standards on health informatics [16]. The interface is made up by forms which are in turn grouped in sections as shown in Table 2.

2.4 Adaptability

Choosing the right initiative strategy and designing appropriate grammars is an essential part of the effort of building an effective dialog system. Such choices are usually made on a trial-and-error basis, resorting to field tests. Sometimes, however, the ability of users varies with time as they learn to use the system more effectively, for example starting to take advantage of mixed initiative interaction. We shall now discuss how the problem of choosing the correct dialogue strategy was tackled, by taking into account a possible increase, from call to call, of a user’s ability to be recognized [17].

A requirement for our dialogue application was to allow for adaptive dialogues. To be perceived as intelligent, in fact, a dialog system should be able to modify its behavior to match the context of the call and the ability of the human
Table 2
EHR sections

<table>
<thead>
<tr>
<th>Section</th>
<th>Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Anamnesis</em> – stores personal data and familial and pathological anamnesis.</td>
<td>50</td>
</tr>
<tr>
<td><em>Life style and risk factors</em> – lists patients’ life habits and displays three standard risk indicators, e.g. the global ten-years Framingham risk.</td>
<td>30</td>
</tr>
<tr>
<td><em>Drugs and tests prescribed</em> – seven web screens are used to record prescribed drugs, side effects possibly observed, and outcomes of examinations and tests.</td>
<td>130</td>
</tr>
<tr>
<td><em>Physical examination</em> – several subsections allow keeping track of physical examination outcomes and doctors’ annotations</td>
<td>85</td>
</tr>
<tr>
<td><em>Queries and Report</em> – perform queries on the database, e.g. to obtain the list of patients that should visit the outpatient clinic within the week, tests prescribed but not carried out yet, and printing a per-patient summary of personal and clinical data.</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 3. Role of terminal and non-terminal elements in a sample grammar.

**(a)** *Today* is a non-terminal token, i.e. a subgrammar or shortcut, which is expanded into further, more complex expressions, such as “Today”, “I feel like”, or even silence, or noise. *(b)** Question mark allows the previous token to be matched or not, and still the match of the other elements may succeed. *(c,d)** *Numbers* is another reference to a subgrammar, which is bound to capture numeric expressions in a suitable range. Subgrammars can be reused: contextual information, given by the surrounding words, allows one to recover which concept a particular number is referring to. *(e,f)** Matched words, pauses, etc. are not passed on to the output, unless they are marked as “important” with the double-slash notation. The label behind the double-slash (semantic tag) is attached to the matched string in order to carry semantic information.
Figure 4. Analog signals and language models are input into a Speech Recognition Engine, which produces a recognition hypothesis in textual form, according to the degree of match between the two.

Figure 5. Homey’s web-based graphical user interface, as shown to the physician. The telephone icon, which appears in the form, reminds the reviewer that the data were entered by the patient via the spoken dialog system.
caller. The system we have designed changes its behavior depending on both the progress of the current call, and the clinical history of the caller. This is achieved by preserving a representation of call data, called the state vector, which is made up of a set of variables; some of them affect the dialogue flow (e.g. if the user is not a smoker, she will not be further asked about the number of cigarettes she smokes). Others are output variables, i.e. strings, integers, or Boolean values, which are filled with the recognition results during the progress of the call. When the telephone is hung up, the values of the output variables reflect what the user has uttered during the call.

Conversely, the database holds other values that should affect the dialogue (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 certain patient, these values are extracted from the database and the corresponding state vector is prepared by an adaptivity agent. The name stems from the fact that this agent will also perform some of the adaptations to make the interaction more effective and friendly.

The dilemma of open (mixed initiative) versus closed (system-directed) questions is then solved if one lets beginner users talk their way through a pedantic but error-proof sequence of questions. After completing successfully a preset amount of calls, the prompts may become more concise, possibly giving hints on how to further shorten the dialogue providing multiple data per utterance.

Once a user becomes “expert” according to a suitable criterion, she would plausibly know in advance what data the system is expecting her to provide, and an open question like “Please say the values you have measured today” would not be really confusing. A similar source of adaptation implemented provides learning of user’s answers, to avoid repeatedly asking selected questions whose answers have been constant in some time span (e.g. “Do you still do swimming”?).

3 Evaluation

The dialog application should be at the same time easy to use and understand; it should also be tolerant of recognition errors, i.e. should not go in states from which callers are unable to recover easily. A number of issues in the design of our application were therefore analyzed through 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
dialogue.

- **Extension of grammars and lexicon** – Grammars should be sophisticated enough to capture most of users’ answer schemes. Open-ended questions (e.g., asking the reason of a skipped call) are particularly worrisome; the allowed lexicon may consist of a large number of words which should be explicitly listed.

- **Reformulation of questions’ wording** – Users’ choice of words used in their answers is influenced by the way questions are asked. Psychological factors are to be kept into account and questions be written so to elicit concise and direct responses.

- **Extraction of patient’s learning curve** – To address adaptability issues, one desires to put each patient into an ability class. One may 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 goals of the project are both to raise patient compliance with the guideline and to expand the availability of data for the use and judgment of the clinician. The collection of quantitative information involving real patients may assess whether the system helps to achieve a better quality of life.

To debug our dialog application according to the goals listed above, we designed a thorough evaluation study divided into two phases. The first phase (internal trial) involved a group of volunteers, which were assigned a realistic disease profile. The profile included therapy, distribution of blood pressure values and other individual information; the profile, in turn, affected the inquiry of side effects during the dialogue. Most of the phone calls were recorded and, later, they were transcribed and analysed. The first phase, therefore, mainly addressed technical and usability issues. It involved about 15 people and collected 400 dialogues, amounting to approximately 1200 minutes of conversation, of which 400 were human speech.

The second part of the evaluation is a controlled clinical trial, with real patients, which will be discussed in section 4.

### 3.1 Recognition quality

The data collected about grammars resemble that shown in Table 3, which displays statistics for grammars associated to two questions. Grammar *Cough* was associated to a question, which asked what kind of cough one experienced as a side effect (when, and how strong). The grammar *Compliance* was associated to a more open-ended question, which inquired the reason why a prescription was not taken (presence of side effects, missing drugs, etc. were among the allowed answers). Both grammars capture information that is di-
rectly useful for the clinician.

The utterances collected during the first evaluation phase were divided by grammar and analyzed. The percentages given in the table show, for the two grammars cited, the fraction of correct recognitions (OK), incorrect (FAIL) and out-of-vocabulary (OOV), the latter meaning the user’s utterance was not foreseen in the lexicon. Not surprisingly, OOV incidence is higher for the open question; this is an indication that the lexicon for that grammar needs to be extended with more words or word combinations. Similar data was collected for all of the 127 grammars used by our application.

Given the clinical importance of some of the values acquired, at certain dialogue steps the user is asked confirmation questions: the systems gives a feedback on the data as understood by the ASR. If the value recognized by the system was not correct, the user has a chance to modify it.

3.2 Dialogue quality

Continuously collecting dialogue parameters during test and use phases allows assessing quantitatively the effects yielded by changes. This surveillance applies not only to grammars, as shown in the previous section, but also to the structure of the dialogue itself. Table 4 summarizes the extent of the dialogue-related data collected during the second half of the internal trial, when the application was in its final version. Data collection was performed automatically and routinely from the transcripts of the telephone calls. Per-call and per-day breakups can be obtained, for data such as how many questions were asked, how much time was spent in each, how many times confirmation questions received negative answer, and so on. Currently, about 24 quantitative indicators are automatically extracted per each call; new indicators can be introduced at any time and evaluated on subsequent as well as past calls.

Figure 6 was obtained from two such indicators. It shows the fraction of confirmation questions per call per day. A marked decrease in this fraction is apparent between days 14 and 15, when a more natural dialogue strategy for confirming multiple values at once was tested and deployed. Fewer confirmation questions make the dialogue shorter and more natural. Dialogue data collected is structured in a way that makes it particularly suited for applying data mining and pattern discovery techniques.
Figure 6. Evaluation of dialogue strategy performance. Quality of the user interaction, correlated to the inverse of confirmation questions, increased after the introduction of multiple-confirmation, on day 15 (x axis has arbitrary origin).

Table 3
Evaluation of grammars performance (example). Incidence of correct recognitions (OK), incorrect recognition (FAIL) and out-of-vocabulary (OOV) are quoted for two grammars.

<table>
<thead>
<tr>
<th>Grammar</th>
<th>OK</th>
<th>FAIL</th>
<th>OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough</td>
<td>81%</td>
<td>10%</td>
<td>9%</td>
</tr>
<tr>
<td>Compliance</td>
<td>41%</td>
<td>48%</td>
<td>11%</td>
</tr>
</tbody>
</table>

Table 4
Evaluation of dialogue strategy performance. First quartile, median and third quartile are quoted for distributions.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Naive</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogues collected</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>Cumulative dialogue time (min)</td>
<td>345</td>
<td>381</td>
</tr>
<tr>
<td>Fraction achieving conclusion (%)</td>
<td>79</td>
<td>80</td>
</tr>
<tr>
<td>Questions per dialogue</td>
<td>23 (29)</td>
<td>34</td>
</tr>
<tr>
<td>% of confirmation questions</td>
<td>32.5(34.8)</td>
<td>36.3</td>
</tr>
<tr>
<td>Average time per question (s)</td>
<td>9.8 (10.3)</td>
<td>11.3</td>
</tr>
<tr>
<td>Average time per call (min)</td>
<td>4.1 (5.1)</td>
<td>5.9</td>
</tr>
</tbody>
</table>

12
4 Controlled clinical trial

A motivation for the adoption of the system is that its use will prove economically viable and beneficial to the collaborating parties. Clinical evidence suggests that careful monitoring and self-monitoring of patients can have a beneficial impact not only on the allocation of resources in hospitals [7], but also on the actual health condition of subjects [6], and their motivation to change behaviours in favour of healthier habits [18,19].

To verify this hypothesis, we collaborated with the caregivers to design a randomized controlled clinical trial. Patients, after being informed of the scope of the study, are enrolled in the EHR. They are randomly assigned to either a control group or a treatment group (population ratio 1:1, stratified by sex). Physicians have encounters with patients of both groups with the same frequency; subjects in the treatment group, in addition to the face to face visit, are asked to periodically dial a toll free number, which connects them to the Homey service. The frequency of calls, which they are asked to make, is once a week for the first two months of use of the system, and once every two weeks later on.

The hypothesis under test is whether, within six months of study, systolic or diastolic blood pressure will decrease in the treatment group by at least 5 mmHg, with respect to the baseline of 140 ± 15 over 90 ± 9 (mm Hg ± SD). The decrease is sought with respect to conventional treatment, i.e. additional to the decrease already attained due to pharmacological or behavioral therapy. Transient effects are in fact avoided by enrolling patients which are in a steady therapy regimen. Study power and significance levels are taken as $p = 0.80$ and $\alpha = 0.05$; this yielded a study size of 304 patients, equally distributed among the collaborating medical institutions. Important, albeit qualitative, user satisfaction surveys will be administered at the middle and the end of the study period, as well. The clinical trial started on August 2003.

5 Future developments

The goal of the project described is to explore the capabilities of spoken dialog systems in the medical domain. While the current prototype showed satisfying performance, its development was not as straightforward as one would expect for such techniques to be adopted in mainstream practice. The system described here is a prototype of future configurable and component-based dialog systems, which may allow a novel modality for users and physicians to access EHRs. We are currently designing a framework which will leverage reuse of pre-built components for recurrent tasks (patient authentication, database ac-
cess) and common questions (asking weight, compliance to the therapy, etc.). Such a framework would enable a rapid prototyping and deployment of dialog applications in domains similar to the one described here, such as diabetes, for which we also have developed a prototype.

Further techniques can be studied in the process of making the system further more user-friendly; for example, the callers’ ease of use could be increased by putting more “intelligence” into the adaptive agents, to shorten the dialogue duration. One way could be to generate shorter prompts spoken to users considered “expert”. Besides the learning of answers cited in section 2.4, it is conceivable to perform more complex statistical tests to classify the level of experience of the users; even the likelihood of her encountering problems on specific prompts could be estimated at runtime, as in [20,21]. In this way, only users that are more reliably understood by the system could have confirmation questions waived in favour of a more general summary confirmation, such as “is all of the above information correct?”.

6 Conclusions

This paper described a prototype home monitoring system for hypertensive patients. The system includes an EHR of the managed subjects; the data may be updated either directly by the physician, or by the patients themselves, via a dialogue on the telephone with an intelligent system. This system allows them to avoid the inconvenience of going for the visit to record those values that can easily measured at home. The physician is able to review all the information entered, along with the derived risk indicators, to make informed decisions. The system is currently being used in three major Italian hospitals. Physicians in specialistic centers for the care of hypertension are enrolling their patients in the EHR system described and using it to store the outcomes of encounters and laboratory tests. A controlled clinical trial to verify the clinical effectiveness is in progress.

Acknowledgements

This research has been partially funded by EU under the Fifth Framework IST project HOMEY, number IST-2001-32434.
References


URL citeseer.nj.nec.com/gorin97how.html


URL citeseer.nj.nec.com/walker98learning.html


URL citeseer.nj.nec.com/article/litman00predicting.html

URL citeseer.nj.nec.com/541013.html