Interacting with Computers 21 (2009) 276–287

Contents lists available at ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Reducing working memory load in spoken dialogue systems

Maria Wolters [a,*], Kallirroi Georgila [b], Johanna D. Moore [b], Robert H. Logie [c], Sarah E. MacPherson [c], Matthew Watson [d]

[a] Centre for Speech Technology Research, School of Informatics, University of Edinburgh, United Kingdom
[b] Human Communication Research Centre, School of Informatics, University of Edinburgh, United Kingdom
[c] Centre for Cognitive Ageing and Cognitive Epidemiology and Human Cognitive Neurosciences Unit, Department of Psychology, University of Edinburgh, United Kingdom
[d] Department of Psychology, University of Sunderland, United Kingdom

ARTICLE INFO

Article history:
Received 14 June 2008
Received in revised form 14 May 2009
Accepted 29 May 2009
Available online 21 June 2009

Keywords:
Spoken dialogue systems
Cognitive ageing
Working memory
Processing speed
Usability
Universal design

ABSTRACT

We evaluated two strategies for alleviating working memory load for users of voice interfaces: presenting fewer options per turn and providing confirmations. Forty-eight users booked appointments using nine different dialogue systems, which varied in the number of options presented and the confirmation strategy used. Participants also performed four cognitive tests and rated the usability of each dialogue system on a standardised questionnaire. When systems presented more options per turn and avoided explicit confirmation subdialogues, both older and younger users booked appointments more quickly without compromising task success. Users with lower information processing speed were less likely to remember all relevant aspects of the appointment. Working memory span did not affect appointment recall. Older users were slightly less satisfied with the dialogue systems than younger users. We conclude that the number of options is less important than an accurate assessment of the actual cognitive demands of the task at hand.

1. Introduction

Spoken dialogue interfaces can make phone-based services far more intuitive to use. If users can simply say the required option, there is no need to remember which option is mapped to which key on the telephone keypad.

Menu-driven systems can suggest commands that describe the option, such as "Say 'listen' if you want to listen to your voice mail" (Perugini et al., 2007). Systems with more advanced natural language understanding can even dispense with commands altogether, simply asking users "How can I help you?" (Suhm et al., 2002; Lai et al., 2008).

Spoken dialogue systems (SDS) consist of five main components. Automatic speech recognition (ASR) converts audio signals of human speech into text strings, natural language understanding (NLU) determines the meanings and intentions of the recognised utterances, dialogue management (DM) controls the interaction, natural language generation (NLG) generates system responses, and text-to-speech synthesis (TTS) converts the system utterances into actual speech output. In this context, we encounter many problems that do not arise in the touch-tone world. For example, how can systems recover gracefully from speech recognition errors (Hockey et al., 2003)? How can users be encouraged to use phrases that the NLU component can interpret correctly (Zoltan-Ford, 1991; Sheeder and Balogh, 2003)? How can we ensure that users understand and retain the key information presented in a dialogue (Walker et al., 2005)?

In this study, we investigate how phone-based services can be made more usable for a particular group of users, older people. Designing interfaces for older users is notoriously challenging (Gregor et al., 2002; Zajicek, 2006). Not only do cognitive and perceptual abilities decline with age (Baeckman et al., 2001; Fozard and Gordon-Salant, 2001), but the spread of abilities in older people is far larger than in any other segment of the population (Rabbitt and Anderson, 2006). Although many strategies have been suggested in the speech technology and computational linguistics literature for making spoken dialogue systems more reliable (e.g. Walker et al., 2001; Bohus and Rudnicky, 2005), almost all of the solutions that have been tested experimentally were evaluated with younger or middle-aged users, not with older adults.

Existing work on adapting SDS to older people tends to fall into two main categories: experimental assessments of full end-to-end systems (Zajicek et al., 2004; Black et al., 2005a; Giorgino et al., 2005) and guidelines that are largely based on laboratory studies

* Corresponding author. Present address: School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, Scotland, United Kingdom. Tel.: +44 131 650 6542.
E-mail addresses: mwolters@inf.ed.ac.uk (M. Wolters), kgeorgil@inf.ed.ac.uk (K. Georgila), j.moore@ed.ac.uk (J.D. Moore), rlogie@staffmail.ed.ac.uk (R.H. Logie), sarah.macpherson@ed.ac.uk (S.E. MacPherson), matthew.watson@sunderland.ac.uk (M. Watson).
URL: http://homepages.inf.ed.ac.uk/mwolters (M. Wolters).

0953-5438/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2009.05.009

(Hawthorn, 2000; Petrie, 2001). Very few studies have systematically compared the impact of different design choices on users' performance, because such experiments take a long time to set up properly. But such systematic comparisons are key to translating theoretical results from cognitive psychology into workable design guidelines. In fact, guidelines for accommodating working memory capacity are notorious for being based on misconceptions (Bailey, 2000; Sharp et al., 2007). For example, the seminal work of Miller (1956) on the magic number 7 ± 2 is often interpreted as an upper limit on the number of options that can be presented at any one time. Instead, the optimal number of options depends on many other factors such as the modality in which they are presented or their complexity.

In this paper, we assess whether two approaches for reducing working memory load in voice interfaces will benefit older users and/or users with lower working memory capacity:

1. reducing the number of options,
2. repeating information in confirmations.

Limiting the number of options makes it easier to remember all of the options in order to select the right one, while providing confirmations reinforces information provided during the dialogue.

The rest of this paper is structured as follows. First, we summarise previous work on SDS for older users with particular attention to the effects of cognitive ageing (Section 2). We then describe the rationale, design, and methodology of our study (Sections 3 and 4). Our results, which are reported in Section 5, are organised according to the three main dimensions of usability according to ISO 9241 (ISO, 1998), effectiveness (Section 5.1), efficiency (Section 5.2), and user satisfaction (Section 5.3). Surprisingly, the two approaches we tested did not differ significantly in either effectiveness or user satisfaction, only in efficiency. We discuss the implications of this result for further work on spoken dialogue interfaces in Section 6 and conclude with a plan of further work in Section 7.

2. Background

2.1. Adapting SDS to older users

There are three relevant strands of research on making SDS more usable for older people: (1) work on older users' attitudes to voice interfaces, (2) work on SDS for applications that predominantly target older users, and (3) work on adapting general SDS to older people.

There is relatively little data on older users' attitudes to SDS. However, in general, older users tend to be more critical judges of usability than younger ones (Stephens et al., 2006). While the five older users who evaluated the smart home interface described in Möller et al. (2007) were more positive than younger users, Möller et al. (2008) report no significant differences in user satisfaction between older and younger users in a larger trial of the same system. Research on interactive voice response (IVR) systems has found that older users tend to be more critical (Rogers et al., 1998).

A prime example of SDS that indirectly target older users are telecare systems, such as phone-based systems for delivering health care interventions (Pollack, 2005; Bickmore and Giorgino, 2006), appointment scheduling systems (Zajicek et al., 2004), smart home systems (Alexandersson, 2008; Gödde et al., 2008), and phone-based symptom management systems (Black et al., 2005a). Many of the conditions covered by symptom management systems mainly affect older people, such as diabetes (DI@L-LOG, Black et al., 2005a,b) and hypertension (HOMEY, Giorgino et al., 2005).

In a randomised controlled trial, patients using HOMEY managed their blood pressure better than patients in the control group, who received traditional ambulatory treatment (Giorgino et al., 2005). DI@L-LOG was evaluated in a field study with diabetes patients aged 55 and over (Black et al., 2005b). Users liked being able to call the system at their own convenience. The biggest problem, speech recognition, was dealt with by constraining the inputs that the system requested. Despite the need for clarification in the speech system, speech interactions did not take longer than touch-tone ones, because touch-tone input requires a short break in the conversation to attend to and operate the key pad. The DI@L-LOG experience shows that spoken dialogue interfaces for older users can be successful, if the systems are sufficiently restrictive.

Shaping user input is particularly important for older users, as the results of Gödde et al. (2008) show. They compared the interactions of older and younger users with a voice interface to a smart home system. Older users were less likely to speak to the system in a way that was easy for the system to understand, and achieved lower task success.

SDS have also been used successfully to deliver task and medication reminders to older people (NURSEBOT, Roy et al., 2000; Pollack et al., 2003). While dialogue management in DI@L-LOG and HOMEY is largely deterministic, NURSEBOT uses complex statistical models called Partially Observable Markov Decision Processes to manage user interaction (Roy et al., 2000).

Zajicek et al. (2004) tested a purpose-built Voice-XML based appointment scheduling system with six older adults, one from the US and five from the UK. Four users successfully arranged an appointment on their own; a fifth user succeeded when walked through the system by an experimenter. Explicit confirmations were used both to verify information provided by the user and to reassure the user that their input had been processed successfully. In order to accommodate memory limitations, messages were kept as short as possible, eliciting or confirming one piece of information at a time. Lists of options were replaced by open questions prefaced with when, where, etc. (wh-questions). The system also provided context-dependent help messages. In line with the findings of (Black et al., 2005b and Dulude, 2002), the main problems older users reported were speech recognition errors and unhelpful error recovery dialogues. The speech recognition, which was tailored to US English, worked best for the US user.

Sharit et al.'s (2003) study of interactive voice response (IVR) systems highlights a further important issue, the transparency of the system's structure. Both the menu structure itself and possible paths through the menu structure should be as clear as possible.

When older users were provided with a graphical aid that explained the menu structure of the IVR systems they were operating, performance improved. Sharit et al. suggest that both touch-tone and voice-based systems should be designed so that it is easy for older users to find and retrieve information. They also recommend that older users be given additional support such as graphical aids to help them plan the conversation with the system. This is particularly important given the demands of using ASR. Users not only need to plan their interaction of the system, they also need to phrase their responses so that the ASR and NLU components can process their input. This may increase cognitive load (Baber et al., 1996).

2.2. Effects of cognitive ageing on usability

Having reviewed some of the previous work on SDS for older people, we now summarise relevant work on cognitive ageing that has shaped the design of the experiment reported in this paper.

Cognitive ageing is a complex phenomenon. Its effects can be seen as early as middle age (Garden et al., 2001). Age-related changes affect many interrelated aspects of cognition, such as information processing speed, mental flexibility, fluid intelligence and memory (Salthouse, 2004).

As with other aspects of ageing, such as visual or auditory ageing (Arking, 2005), we see considerable inter-individual variation in the speed and the extent to which abilities decline (Rabbitt and Anderson, 2006). This variation is not only related to cognitive abilities when young, but also affected by events during the individual's lifetime (Deary et al., 2004). To complicate matters even further, crystallised cognitive abilities, such as knowledge of vocabulary, may even improve with age. Such abilities are attained over a lifetime and maintained by practice in old age (Horn, 1982; Salthouse, 2003).

Age-related changes in cognitive abilities are strongly interrelated. For example, age-related decline in working memory is mediated by processing speed, but potentially also other cognitive abilities (Levitt et al., 2006). Some cognitive psychologists have proposed a single common cause of age-related cognitive changes, a decline in the speed with which incoming information is processed (*cognitive slowing*). Structural equation models of cognitive ageing show that age has a direct effect on information processing speed (Charlton et al., 2007). This single factor in turn accounts for a significant proportion of the variance found in mental abilities such as fluid intelligence and working memory capacity (Salthouse, 1993; Schaie, 1989). However, it has been argued that cognitive slowing cannot fully account for age-related cognitive decline (Bugg et al., 2006; Salthouse, 1991; Schaie, 1989). The rate of age-related decline varies across tasks, which is not necessarily consistent with a single-factor model (Rabbitt et al., 2004). The results of Rabbitt et al. (2006) point to fluid intelligence as an additional, separate factor that is needed to account for cognitive ageing.

Therefore, we require a comprehensive battery of tests that measure fluid intelligence, crystallised intelligence, working memory and information processing speed, in order to get a reasonably complete picture of the relevant abilities of a participant.

Such a detailed battery is also of practical relevance, since the aspects of cognitive ageing that predict users' performance on a particular task depend on the nature of the task itself (Czaja and Lee, 2007).

For example, in their study of older people's use of IVR systems, Sharit et al. (2003) found that hearing, working memory, and selective attention were good predictors of the variance in users' performance. Once these factors had been included in the model, age became irrelevant. This illustrates that it is not so much chronological age which affects performance, but the changes in ability that occur as a result of ageing.

In order to tease out how voice interfaces should be modified to accommodate age-related decline on working memory, several authors have looked to the cognitive ageing literature (Bond and Camack, 1999; Hawthorn, 2000; Petrie, 2001; Zajicek, 2004). A general recommendation echoed again and again is that limited working memory capacity should be accommodated by restricting the number of options that are presented at any one time. But determining the right number of options is far from straightforward (Bailey, 2000): should it be seven (Miller, 1956), two (Zajicek and Khin Kyaw, 2005), or four (Bond and Camack, 1999)? Two important confounders need to be addressed before we can begin to answer this question: modality and task.

2.2.1. Confounder 1: modality

It is well known that users can cope with long lists of options if the list is presented visually (Sharp et al., 2007). Since all options can be scanned as often as the user wishes, they do not need to be remembered. Visual interfaces for older users in particular can accommodate a large number of options if the options are easy to scan, recognise, and digest (Kurniawan et al., 2002; Zaphiris et al., 2007).

Repeated, quick scanning of information is far more difficult in auditory than in visual interfaces. Typically, lists of options can only be re-scanned by requesting the relevant parts of the original message to be played again in full. This is particularly problematic if the user is not only required to remember the complete list as in Zajicek and Morrissey (2001), but to reason about the options.

In multiple component models of working memory, the way in which incoming information is presented also determines how it is stored during processing (Alan Baddeley, 2007; Logie and vander Meulen, 2009). While visuo-spatial information is stored in the visuo-spatial sketch pad, verbal information is stored in the phonological loop. These slave systems are in turn coordinated by a central executive system, which controls both storage and processing of information in working memory.

When users reason about a list of auditorily presented options, they need to subvocally rehearse these options in their phonological loop. Meanwhile, the central executive not only coordinates this storage process, but also manages reasoning. The model predicts that such a high load on the central executive will cause performance to drop (Duff and Logie, 2001). The less reasoning required, the lighter the load on the central executive, and the better the performance will be.

The picture is further complicated by age-related problems with hearing. If messages are not clear enough or not loud enough, users' hearing loss can affect their speech comprehension and problem solving performance (Rabbitt, 1990; Schneider et al., 2005).

2.2.2. Confounder 2: task demands

Czaja and Lee (2007) demonstrated that we need to assess the cognitive demands of specific tasks before we can begin to adapt to cognitive ageing. Working memory has been implicated in the ability to use graphical aids (Lohse, 1997; vander Meulen et al., 2009), remember to attend appointments (Morrow et al., 1998), search for and retrieve information (Sharit et al., 2004), and, crucially, use touch-tone IVR systems (Sharit et al., 2003).

But it is not just the nature of the task that needs to be considered, it is also the complexity of the task. Since working memory is the short-term store for processing information, tasks where users need to process more information or process information more thoroughly will tax it more than tasks that require no complex cognitive operations (Sjölinder et al., 2003).

If we adopt a strategy for reducing the load on working memory, such as reducing the number of options, we also need to carefully consider the impact this will have on other aspects of the dialogue, more specifically its length. Reducing the number of options that can be presented in one message often results in deeper task hierarchies and hence longer dialogues. The longer the complete dialogue, the more difficult it is to remember information that has been gathered and decisions that have been made earlier in the dialogue. Huguenard et al. (1997) found that the short-term advantage conferred by reducing the number of options was cancelled out by the long-term disadvantage of forgetting information that had been presented earlier in the interaction.

Finally, as we have discussed earlier, it is not just the amount of material that matters, but also the depth at which it is processed. Commarford et al. (2008) showed that if users do not have to remember all options, but only need to select the appropriate choice from a list, users with short working memory spans profit from longer lists of options. When the number of options was shortened, users with shorter spans actually performed worse than their counterparts with longer spans. This might be due to the fact that fewer options result in longer dialogues with more complex branching structures.

3. Aims and design

Our study was designed around a task that not only requires users to choose between a number of options, but is also highly relevant to telecare applications: scheduling health care appoint-

ments over the phone. We assessed two approaches to accommodating users with low working memory span (WMS):

Reduce number of options: If users are presented with a large number of options at each step in the appointment scheduling dialogue, they are less likely to schedule correct appointments and to remember scheduled appointments.
Provide confirmations: If the system confirms each aspect of the appointment, users will find it easier to remember the appointment, since relevant information is repeated.

We expected that users with lower WMS would benefit more from reduced numbers of options and repeated confirmations than users with higher WMS. Apart from (Commarford et al., 2008), we are not aware of any formal experimental study that assesses the impact of systematically varying the number of options on the usability of a SDS for users with low WMS. Neither are we aware of any such studies that look at the effect of varying confirmation strategies on the usability of SDS for this particular user group. Our study is intended to address this gap in the evidence base. Since cognitive ageing is multifaceted, we assessed not only working memory, but designed a battery of tests that yield a comprehensive picture of overall cognitive abilities.

We measured two aspects of task success: *Completion* and *Recall*. *Completion* measures whether users successfully arranged an appointment with the correct health professional that fits their schedule, while *Recall* assesses whether users were able to remember the appointment they scheduled. Both aspects are equally important: if a user manages to arrange an appointment, but fails to attend because crucial details have been forgotten, the appointment has essentially not been scheduled successfully.

Task success results are presented in Section 5.1, data on efficiency is presented in Section 5.2, and results on user preferences and user satisfaction are reported in Section 5.3.

4. Method

4.1. Cognitive tests

Our battery of tests covered the two main dimensions of intelligence, fluid intelligence, which is linked to abstract reasoning, and crystallised intelligence, which is linked to acquired knowledge, as well as working memory and information processing speed. All tests were presented visually, to avoid problems due to age-related hearing loss (Rabbitt, 1990). The full battery took 60–90 min to administer.

Fluid intelligence was assessed using Ravens' Progressive Matrices (*Ravens*, Raven et al., 1998). Participants were not timed. *Crystallised intelligence* was measured using the Mill Hill Vocabulary test (*MillHill*, Raven et al., 1998). *WMS* was assessed with a widely used sentence reading span test, administered through ePrime (*SentSpan*, Unsworth and Engle, 2005). We chose reading span over digit span tests since reading span requires participants to process the stimuli instead of simply remembering them. Thus, it taps into the key function of working memory as a short-term store for information processing. Participants' responses on the WMS task were timed. Before attempting the main task, participants completed a range of practice items. In this paper, we report the absolute score, which aggregates participants' scores for all test items. [1] *Information processing speed* was assessed using the Digit Symbol Substitution subtest of the Wechsler Adult Intelligence Scale-Revised (*DSST*, Wechsler, 1981).

4.2. Wizard-of-Oz simulation

We implemented the dialogue strategies for our appointment scheduling system using a Wizard-of-Oz (WoZ) simulation (Dahlbäck et al., 1993). In a WoZ setup, while the users are led to believe that they are interacting with a fully automated system, all or part of the system is simulated by a human.

WoZ experiments are an invaluable tool for investigating different design options for SDS, because end-to-end systems are very time-consuming to build and test. In addition, WoZ studies allow experimenters to isolate the effects of high-level information presentation and dialogue management from the problems introduced by the limitations of current automatic speech recognition (ASR) and natural language understanding (NLU) systems. The user data gathered in WoZ experiments can be used to improve both these components. Thus, WoZ studies of dialogue system prototypes lay the groundwork for building full, end-to-end systems (Fraser and Gilbert, 1991; Dahlbäck et al., 1993).

WoZ systems differ in the degree to which the system is simulated. In this study, the wizard simulated ASR, NLU, and dialogue management (DM), while natural language generation (NLG) and text-to-speech synthesis (TTS) were fully automated. The wizard processed user utterances, selected the appropriate next dialogue act, and updated the health professional's calendar with information provided by the user. Although the WoZ system suggested health professionals, half-days, and half-hour time slots automatically, the wizard was able to override the automatic selections manually if necessary. The wizard also manually marked the time slot that represented the final booking. All speech output was generated automatically. Output sentences were generated using a simple template-based natural language generation system and then fed to the unit selection text-to-speech synthesiser Cerevoice (Aylett et al., 2006). Pilot studies have shown that older people can understand the synthetic speech generated by Cerevoice as well as human speech if messages contain familiar material such as times and dates (Wolters et al., 2007).

Each dialogue proceeded through three main stages. First, the system asked which health professional the user wanted to see, then, a half-day was arranged, and finally, a half-hour time slot within that half-day was agreed on. In all three steps, the system initially presented the user with a fixed number of options: one (yes/no answer), two, or four (Fig. 1). The user's choice was either confirmed explicitly through a confirmation dialogue, implicitly by mentioning the user's choice again in the next stage of the dialogue, or not confirmed at all (Fig. 2). The wizard was unable to skip any of these stages. In a final step, the wizard confirmed the appointment, giving four pieces of information: the health professional, the day of the appointment, the time of the appointment, and the location of the appointment. All of these items, except for location, had been discussed earlier.

When varying the number of options, the most basic unit is one option, which then needs to be confirmed or disconfirmed by the

1 Option (Yes/No):
System: Would you like to see the occupational therapist?
2 Options:
System: Would you like to see the occupational therapist or the community nurse?
4 Options:
System: Would you like to see the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse?

Fig. 1. Dialogue strategies used to examine number of options.

[1] We also computed a timed score, for which we discarded all items where participants took more than the mean plus 2.5 standard deviations of the time taken for the practice items. The cut-off was computed individually for each participant. Results for the timed score are similar to results for the absolute score; both are highly correlated ($\rho = 0.89$).

Explicit Confirmation Subdialogue (Explicit):
User: I would like to see the occupational therapist, please.
System: You would like to see the occupational therapist. Is that correct?
User: Yes.
Implicit Confirmation in Answer (Implicit):
User: I would like to see the occupational therapist, please.
System: When would you like to see the occupational therapist, on Monday afternoon or on Friday morning?
User: Monday afternoon would be best.
No Confirmation (None):
User: I would like to see the occupational therapist, please.
System: When would you like to come, on Monday afternoon or on Friday morning?
User: Monday afternoon would be best.

Fig. 2. Dialogue strategies used to examine confirmation provision.

user. This is a popular design choice if speech recognition is required, because it greatly constrains the user's answers. The next alternative is presenting two options at a time, as suggested by Zajicek (2004) in her design pattern *Partition Message*. Restricting the possible input to two options constrains the space of potential answers to one of the two possible options plus a generic pattern for rejections. If both options are rejected, the system proceeds to offer an alternative pair. As we will see below, such a strategy is far more efficient than simple yes/no answers. Thus, presenting two options is potentially a win–win situation: not only is it more efficient than presenting one option at a time, but it also allows speech recognition performance to remain relatively high. The largest number of options that were presented was four at a time. This represents the upper limit suggested in some of the available guidelines (Bond and Camack, 1999; Sharp et al., 2007).

For our confirmation strategies, the baseline was not to provide any confirmation. Explicit confirmation subdialogues are often used for key pieces of information that need to be jointly agreed or "grounded" (Clark and Brennan, 1991; Traum, 1994; Brennan, 1998) before the dialogue can proceed. If information is agreed in a special subdialogue, the user is highly likely to attend to the repeated information. If the system has made a mistake, error recovery at this point is relatively straightforward.

Instead of engaging the user in a subdialogue, systems that use implicit confirmations merely repeat the information that the system assumes to have been agreed, such as the health professional whom the user wishes to see, or the half-day on which the user wishes to attend the clinic. The lack of the confirmation subdialogue can potentially make the dialogue significantly shorter. On the other hand, the user might not attend to the repeated information and thus fail to notice that the system has interpreted the previous dialogue incorrectly. Even if the user notices the incorrect interpretation, the mistake can be difficult to repair, because users may react with long utterances such as "No, I don't want to see the physiotherapist, I said that I wanted to see the occupational therapist." Users may also articulate the names of the health professionals particularly carefully. This paradoxically makes it more difficult for the ASR to understand what users are saying, because ASR systems are typically trained on normal speech, not on hyperarticulated speech (Stent et al., 2008).

4.3. Questionnaire

The user questionnaire was based on the ITU-T recommendation P.851 (ITU-T Rec. P.851, 2003) as implemented in (Möller et al., 2007), which is one of the de-facto standards in the field and builds on relevant previous work such as PARADISE (Walker et al., 1998) and SASSI (Hone and Graham, 2000). The items were adapted to the task and some were slightly reworded. Each system was evaluated separately. We did not ask users for a final summary evaluation, since they had been exposed to nine very different systems. The questionnaire consisted of 39 items, including perceived task completion, overall impression, and 37 items that were rated on a five-point Likert scale. These items, which are listed in Appendix A, contain an explicit item to measure user satisfaction. Completion took around 5 min. This time was long enough to distract the user from the original task, but short enough to allow users to complete the questionnaire after each of the nine interactions. The comparatively large number of items allows the questionnaire data to be subjected to factor analysis to uncover underlying dimensions of user judgements. The items were presented in meaningful groups and were matched closely to the implementation specified in Möller et al. (2007). We believe that grouping the items did not substantially affect the outcome of the factor analysis (see Section 5.3 below).

4.4. Tasks

Participants were asked to book nine appointments in total. Each task consisted of scheduling an appointment with one of four different health care professionals: a community nurse, a diabetes nurse, an occupational therapist, and a physiotherapist (Example: "Please book an appointment with the occupational therapist."). Participants were told that after scheduling each appointment, they would be asked to rate the system they had just interacted with and recall the appointment they had scheduled. Each appointment was booked using a different SDS. Four lists with nine tasks each were created in which each health professional appeared at least twice. Each user was randomly assigned to one of the lists. The order of health professionals on each list was randomised.

Each task was presented to participants on a screen in a large font. After memorising the task, participants pressed a key to start the interaction. Participants were able to recall the task at any time by pressing the space bar.

In addition to the task, participants were also given a schedule that showed the days and times on which they were free. Each schedule spanned a working week from Monday to Friday. For each half-day, at least two half-hour slots were marked as "unavailable". These slots were selected randomly. The users' schedules were designed to overlap with the schedules of each of the four health professionals by at least two half-days, so that it would always be possible to book an appointment. Error recovery was limited: both wizard and user could ask each other to repeat the last utterance, and the wizard was able to backtrack if users signalled that a mistake had been made. During the interaction with a particular system, the number of options and the confirmation strategy were never changed.

For each appointment booking task, participants interacted with a different prototype dialogue system. The order of systems was randomised separately for each participant. As a result, no tasks were associated with any one of the dialogue systems. Moreover, over the whole data set, all dialogue strategies are equally likely to occur in the first three dialogues, the mid three dialogues, or the last three dialogues. This minimises any potential confounding efffects of system order. After participants had booked an appointment, the experimenter removed the schedule and asked them to evaluate the system they had just interacted with using the 39-item questionnaire described in Section 4.3. Upon completion of the questionnaire, participants were asked to recall the appointment. They were asked about all four items of information presented in the final system confirmation: health professional, day, time, and location. This delayed recall is crucial to our design. Since the questionnaire took around 5 min to complete, the experimental setup simulated a momentary distraction between the user hanging up the phone and noting the appointment down in their diary. Users were not told whether the details they had

Table 1
Participant demographics (mean ± std. dev. with range in brackets).

	Younger	Older
Age (years)	22 ± 2.5 (18–29)	65 ± 8.5 (52–84)
Education (years)	17 ± 2.5 (12–22)	15 ± 5 (9–30)
% Female	70.8% (n = 17)	62.5% (n = 15)

Table 2
Differences between age groups across cognitive measures (t-test).

Test	Younger	Older	Sig.	95% CI
DSST	74.8	52.1	$p < 0.000$	[−28.1, −17.4]
MillHill	42.1	53.7	$p < 0.000$	[7.6, 15.4]
Ravens	54.3	49.8	$p < 0.001$	[−7.0, −2.1]
SentSpan	58.5	47.6	$p < 0.01$	[−18.4, −3.5]

recalled matched the actual appointment details. Since we were testing the systems, and not the users, we did not want to put users under any pressure to "perform".

4.5. Participants

We recruited 26 older and 24 younger participants. The younger participants were undergraduates recruited through an advertisement on the student jobs website. The older participants were recruited through two user panels and a local community centre. All participants signed informed consent forms and received full and timely information before each cognitive assessment as well as before the dialogue experiment. Two older subjects, a 75-year-old female and an 81-year-old male, were excluded from the analysis, because they were unable to complete the reading span test. Three participants did not provide information about education. Table 1 summarises information about our participants. The differences in gender and years of education between the two groups are not significant (χ^2 test for gender, $p = 0.76$, $\chi^2 = 0.0938$, df = 1; Kruskal–Wallis test for education, $p = 0.13$, Kruskal–Wallis $\chi^2 = 2.2936$, df = 1). [2]

Although in many studies, older participants are typically aged 60+, we decided to adopt a lower age limit of 50. Since SDS still need substantial basic research before they can be deployed successfully with older users, we anticipated that people who are 50 now are more likely to use voice interfaces in their old age than people who are 70 or older. We were confident that we would see some effects of cognitive ageing in users as young as 50 years as cognitive abilities such as working memory start to decline as early as middle age (Garden et al., 2001). The number of participants is sufficient for establishing large or medium sized effects (Cohen, 1988) with satisfactory power. This effect size is adequate for motivating guidelines—if effects are any smaller, it is doubtful whether they would affect everyday usability.

The battery of cognitive tests revealed significant differences between younger and older users (Table 2). These results show the expected pattern: crystallised intelligence increases with age, and information processing speed, fluid intelligence and WMS decrease with age. (cf. Section 2.2). As expected, *Ravens* and *MillHill* are not correlated ($\rho = -0.1$, $n = 48$, $p = 0.5$), *SentSpan* is correlated with performance on *Ravens* ($\rho = 0.33$, $p = 0.02$, 95% CI [0.06, 0.57]), and *DSST*, the speed of processing task, is highly correlated with all three other measures ($p < 0.005$ or better, *Ravens*: $\rho = 0.57$, 95% CI [0.34, 0.74], *MillHill*: $\rho = -0.4$, 95% CI [−0.62, −0.13], *SentSpan*: $\rho = 0.5$, 95% CI [0.25, 0.69]).

[2] Following established practice in cognitive psychology, we used years of education in order to match older and younger participants for level of educational achievement. We are aware that this measure is but a crude proxy. Not only has the education system changed significantly over the decades, but older adults were also encouraged to leave school earlier, regardless of whether they were "clever" enough for a degree course.

Table 3
Scoring system.

Aspect	Points	Criterium
Day	1	Day of the week correct
	0	Day of the week incorrect
Time	2	Hour and minutes correct
	1	Only hour correct
	0	Hour incorrect
Health prof.	2	Complete name correct
	1	Type correct (nurse/therapist)
	0	Type incorrect
Location	2	Correct
	1	Correct synonym
	0	Incorrect

5. Results

5.1. Effectiveness

In order to assess effectiveness, we examined whether users scheduled an appointment with the correct health professional at a time that was labelled as available in their schedule (*Correctness*) and whether users were able to remember appointment details correctly (*Recall*).

Correctness was uniformly high. Younger users scheduled 94% of their appointments with the correct health professional, older users scheduled 91.2% of their appointments correctly. This difference is not statistically significant (χ^2-test, $p = 1$). There was no effect of dialogue strategy or any of the cognitive measures. The schedule was always used correctly: no user scheduled an appointment at a time they could not make.

We measured *Recall* by determining how many aspects of the appointment were remembered correctly. The scoring system is summarised in Table 3. The maximum score per task was seven (=everything correct), the minimum score was 0 (=nothing correct). Table 4 shows average *Recall* for both younger and older users. Overall, performance is near ceiling. As a consequence, *Recall* is not normally distributed, and standard analysis through ANOVAs is not feasible. Instead, we use Kruskal–Wallis tests and Spearman rank correlation coefficients to assess each of our hypotheses in turn.

We tested the influence of dialogue strategy on the performance of all users, older users, who may benefit more than younger users, and users with low WMS, who are not necessarily older. We defined users with low WMS as those whose *SentSpan* results fell into the first quartile of the sample. The results of our significance tests show that our two hypotheses about the effect of dialogue strategy on user performance must be rejected (cf. Table 5). Users neither benefit from fewer options nor from explicit confirmations. None of the dialogue strategies we tested helps users with lower WMS.

One might suspect that this finding is mainly due to the ceiling effect we observed in our users' performance, which obscures any performance differences due to cognitive ability. If we look at the correlations between cognitive measures and user performance, though, we see clear effects of cognitive ability on *Recall*—but not the effects we anticipated (Table 6). Whereas WMS does not affect *Recall*, *Recall* correlates with information processing speed as measured by *DSST*. [3]

[3] For these by-participant correlations, we used the total scores of each participant.

Table 4
Recall by dialogue strategy and age group (mean ± std. dev.).

# Opt.	Confirmation			*Total*
	Explicit	Implicit	None	
Older Users (n = 24)				
1	6.0 ± 1.3	6.5 ± 0.6	6.0 ± 2.5	6.2 ± 1.5
2	6.3 ± 0.8	6.3 ± 1.3	6.3 ± 0.9	6.3 ± 1.0
4	6.3 ± 1.3	6.5 ± 0.7	6.6 ± 0.5	6.5 ± 0.8
Total	6.2 ± 1.2	6.4 ± 0.8	6.3 ± 1.3	6.3 ± 1.1
Younger Users (n = 24)				
1	6.1 ± 0.7	6.5 ± 0.6	6.8 ± 0.9	6.4 ± 0.8
2	6.1 ± 0.7	6.3 ± 0.8	6.3 ± 0.8	6.2 ± 0.7
4	6.4 ± 0.7	6.7 ± 0.7	6.4 ± 0.6	6.5 ± 0.6
Total	6.2 ± 0.7	6.5 ± 0.7	6.5 ± 0.8	6.4 ± 0.7

Table 5
Effect of dialogue strategy on *recall* (Kruskal–Wallis tests).

Group	Variable	
	# Options	Confirmation
All users	$\chi^2 = 2.14$, $p < 0.35$	$\chi^2 = 3.77$, $p < 0.2$
Older users	$\chi^2 = 2.29$, $p < 0.35$	$\chi^2 = 1.15$, $p < 0.6$
Low WM users	$\chi^2 = 0.82$, $p < 0.7$	$\chi^2 = 2.06$, $p < 0.4$

Table 6
Correlation between *recall* and cognitive measures (Spearman's ρ).

Test	ρ	Sig.
DSST	0.328	$p < 0.05$
Ravens	0.161	n.s.
MillHill	0.011	n.s.
SentSpan	0.184	n.s.

5.2. Efficiency

Efficiency was measured as the total number of turns per dialogue. System turns consist of a complete system message. User turns are coherent sequences of one or more utterances produced by the user. The beginning of a user turn was delimited either by the start of the dialogue or the end of a system message, while the end of a user turn was delimited by the beginning of a system message or the end of the dialogue. User turns sometimes partially overlapped with the preceding and/or following system messages. The optimal number of turns was eight, if there were no explicit confirmation subdialogues, with two turns each for agreeing health professional, day, and time and two turns for confirming the appointment. ANOVA results show clear effects of dialogue strategy (confirmation strategy: $p < 0.000$, df = 2; number of options: $p < 0.000$, df = 2; confirmation strategy × number of options: $p < 0.005$, df = 4) and age group ($p < 0.000$). [4] We also find significant interactions between confirmation strategy and age group ($p < 0.01$). The relevant medians and marginal medians are summarised in Table 7.

Avoiding explicit confirmation subdialogues reduces the median number of turns in a dialogue from 17 to 9 for younger users and from 22 to below 13 for older users. Likewise, presenting more than one option reduces the median number of turns from 17 to 14 for older users and 11 for younger users (two options). A further two turns can be saved if the number of options is increased to four.

Older users' dialogues are on average two turns longer than those of younger users. This is partly due to older users' inclination to greet the system and say "good-bye" after the final confirmation. However, not all extra turns reflect increased chattiness and sociability, since extra material that is not directly relevant to the task often forms part of the same turn as directly task-relevant utterances. The longest dialogues by far are those where the wizard presents one option at a time and where each choice is confirmed in a separate subdialogue (median length 27 for older, 23 for younger users). It appears that using explicit confirmations makes older users less efficient. When the system presents only implicit confirmations or no confirmation at all, older users' dialogues are only around two turns longer than younger users' dialogues. When the system uses explicit confirmations, on the other hand, the gap widens to around five turns.

Table 7
Median dialogue lengths by dialogue strategy and user group. All = All Dialogues.

# Options	Confirmation			All
	Explicit	Implicit	None	
Younger users				
1	23	17	15	17
2	15	9	9	11
4	14	9	9	9
All	17	9	9	13
Older users				
1	27	13	15	17
2	18.5	11	12	14
4	19	10	10	12
All	22	11	12.5	15

5.3. Satisfaction

We extracted three main outcome variables from our questionnaire:

Impression: users' overall impression of the nine systems. Overall impression of the conversation was rated using a continuous ruler which was marked with five anchor points ("very poor", "poor", "neutral", "good", and "very good"), where "very poor" corresponded to 1, "very good" to 5. Ticks on the ruler were converted to an interval-scaled variable ranging from 1 to 5, with the value rounded to the first decimal point.
Satisfaction: users' satisfaction with each of the nine systems. This was taken from the questionnaire item "Overall, I am satisfied with the booking system" which was rated on a discrete 5-point Likert scale, with 1 corresponding to "strongly disagree" and 5 corresponding to "strongly agree".
Perceived Completion: perceived task completion. This was a binary variable, with "yes" indicating that the user thought the task had been successfully completed, and "no" indicating that the task could not be completed.

We also explored the underlying factor structure of the questionnaire using factor analysis.

For 40 appointments, which were in fact scheduled successfully, users forgot to rate *Perceived Completion*. For all remaining appointments, users reported that they were scheduled successfully. Older and younger users do not differ significantly in their *Impression* ratings (Younger users: 3.7 ± 1, older users: 3.6 ± 1, $t = -0.4565$, df = 423.622, $p = 0.65$, 95% CI [−0.3,0.2]). Dialogue strategy does not affect users' overall impression of the system, either (*t*-tests). Only 14 out of 48 users (4 older, 10 younger) expressed clear preferences, with *Impression* scores distributed over a range of three and more. A detailed analysis reveals, however, that these users tended to prefer one or two prototypes over others. We found no trends in individual judgements that could be attributed to dialogue strategy.

[4] $p < 0.000$ indicates a significance level of $p < 0.0005$ or better.

Table 8
Mean *satisfaction* ratings.

# Options	Confirmation			*Total*
	Explicit	Implicit	None	
Younger users				
1	3.9	3.9	3.8	3.8
2	3.8	3.5	3.7	3.7
4	3.8	3.7	3.8	3.8
Total	3.8	3.7	3.8	3.8
Older users				
1	3.4	3.3	3.4	3.4
2	3.1	3.4	3.3	3.3
4	3.4	3.3	3.1	3.3
Total	3.3	3.3	3.3	3.3

Reliability: The information provided by the booking system was clear; the information provided by the booking system was incomplete; the booking system is unreliable; the booking system made a lot of errors.

User Satisfaction: Overall, I am satisfied with the booking system; using the booking system to book healthcare appointments was comfortable; I would use the booking system again in the future; using the booking system was worthwhile.

Look and Feel: The voice of the booking system sounded natural; the booking system was friendly; the booking system reacted like a human; the conversation with the booking system was pleasant.

Efficiency: The dialogue led quickly to the desired aim; the dialogue was too long.

Perceived Cognitive Load: I had to concentrate in order to hear the booking system correctly; I had to concentrate hard when making the appointment.

Unpredictability: I was not always sure what the booking system expected of me.

Fig. 3. Factor structure of user questionnaire.

For *Satisfaction*, we did not see an effect of dialogue strategy on scores, but there was a clear age effect: older users are less satisfied than younger users (Wilcoxon rank sum test, $W = 18001$, $p < 0.000$, 95% CI $[-1, -0.0005]$). Table 8 summarises mean scores.

We determined the underlying dimensions of users' responses to the 38 non-binary questionnaire items [5] using maximum-likelihood estimation factor analysis (R method `factanal` (R Development Core Team, 2006)). Missing values were replaced by the mean of the corresponding questionnaire item (Kaiser normalisation). Following (Darlington, n.d.), we chose the largest factor model for which all factors were interpretable. Factors were classed as interpretable if there was at least one questionnaire item which had a loading of 0.4 or higher on that factor (Bortz, 1993). The higher the loading of an item on a factor, the stronger the item's association with it. The total number of input vectors was 432 (one questionnaire per system, nine questionnaires per user).

The resulting solution had six factors, which explain 66.5% of the variance in the data. Of these factors, the first four represented positive judgements, the last two negative judgements about the data. Fig. 3 lists each factor together with the key questionnaire items. Following (Bortz, 1993), key items were characterised by having a loading of 0.6 or higher on that factor.

Our six-factor structure compares well with the eight factors found by Möller et al. (2007), which are summarised in Fig. 4. The items on four factors are very similar: "User Satisfaction", "Reliability", "Perceived Cognitive Load", and "Efficiency" (Möller et al.: "Interaction Speed"). Since there were very few user errors, we found no separate "Error" factor. Our sixth factor, "Unpredictability", consists of a single item, which is included in Möller et al.'s "Reliability" factor. Since that item also clearly loads on our "Reliability" factor (load on factor >0.4), the two can potentially be collapsed. The largest difference between the two factor structures is our third factor, "Look and Feel". Our users liked systems that responded like a "good human": responsive, friendly, and natural. Möller et al.'s "Cooperativity" item also contributed to this factor, but its load was only medium (0.544). Although items were not presented in a random sequence, there is evidence that this did not unduly affect the factor structure. Four out of six factors contain items from different sections of the questionnaire; the only exceptions are the two-item factor "Efficiency" and the one-item factor "Unpredictability". The close match between our factor structure and the previously published factor analysis of Möller et al. (2007) also suggests that the differences in age and cognitive ability between our participants did not overly affect the factor structure we obtained.

[5] We excluded *Perceived Completion*, which was always "yes".

C1 (User Satisfaction): I prefer to operate domestic devices in a different way; I would use the system again in the future; I could direct the dialogue as I wanted; overall, I am satisfied with the system; the system is helpful for operating domestic devices; the interaction with the system was pleasant; domestic devices can be operated e ciently with the system; operating domestic devices via speech was comfortable.

C2 (Perceived Cognitive Load): A high level of concentration is required when using the system; I got easily lost in the dialogue flow; I felt relaxed.

C3 (Efficiency): The system did not always react as expected; the information provided by the system was clear; the system did not always do what I wanted.

C4 (Reliability): The system made many errors; the system in unreliable.

C5 (Ease of Use): I had to concentrate to acoustically understand the system; it is easy to learn to use the system.

C6 (Cooperativity): The system behaved in a cooperative way.

C7 (Naturalness): The dialogue was balanced between me and the system; the system voice sounded natural.

C8 (Speed of Interaction): The system reacted too slowly.

Fig. 4. Factor structure from Möller et al. (2007). Factor names added by the present authors.

6. Discussion

In this study, we attempted to accommodate users with low WMS by combining two approaches: using confirmations to reinforce aspects of the appointment that had already been agreed upon, and reducing the number of options presented in a single system turn. Neither strategy had a measurable effect on users' performance, even though our participants had a wide range of WMS (cf. Table 2). If anything, older users benefited from being presented with more options and fewer explicit confirmations—particularly in terms of efficiency. Despite overall high performance levels, we found clear correlations between measures of cognitive ageing and performance: Users with a slower information processing speed found it more difficult to remember all relevant aspects of the appointment. Contrary to expectations, we found no effect of WMS.

Our study clearly benefited from a rigorous, comprehensive cognitive assessment battery. It not only highlighted the area where older people have an advantage over their younger counterparts, crystallised intelligence, but also pointed to a possible explanation of our findings in the shape of the clear correlation between speed of information processing and appointment recall.

In fact, this correlation suggests a possible reason why our strategies did not affect the performance of users with low WMS. The central challenge in our appointment scheduling task was not to remember all available appointments, but to monitor one's own schedule and detect the option that fits. Since slots were only

labelled as free or blocked, users could easily solve the task by scanning, with no need for additional planning. Thus, our results complement Commarford et al.'s (2008) finding that users with a lower WMS benefit from being presented with more options at a time, because at each step in the interaction, they are more likely to be presented with the correct choice. As a result, the overall interaction becomes shorter and less complex.

Even though dialogue strategy and user age do not appear to affect *effectiveness*, they clearly affect *efficiency*. Our results suggest that systems should present more than one option at a time unless the intended users are unable to monitor a list of options for the one that suits them best. Likewise, explicit confirmation dialogues should be avoided unless the user requests them as a feature or a low speech recognition confidence score requires them.

Our older users may also have benefited from their experience with appointment scheduling. Appointments have a clear schema (who, when, where) that the design of our system strictly adhered to. Such schemata can be exploited in many eHealth applications, because often, there are standard questionnaires and procedures that users will be familiar with. If computer-based systems follow these standards closely, users may well find the computer versions easier to navigate, because the implementation follows the users' mental model of the task.

In our corpus analysis, we have found that some older users like to take the initiative, suggesting person, day, and time in the first utterance, while some prefer to take the backseat and passively respond to the system's prompts (Georgila et al., 2008a). These differences in interaction styles and the highly idiosyncratic user preference results suggest that it may be less important to adapt to different user groups than to adapt to individual users. This requires not just robust speech recognition and flexible language models that adapt to the user's voice and vocabulary. We also need to develop adequate recovery strategies for communication problems. All of these challenges are open research questions in speech technology and computational linguistics.

Much recent work in spoken dialogue systems has explored statistical approaches to dialogue management (Lemon and Pietquin, 2007). Since data from interactions with real users is typically not sufficient for exploring the large space of potential dialogue policies, systems are typically trained with simulated users (Schatzmann et al., 2006). Using the data gathered in this experiment, we have successfully built simulated users that reproduce the behaviour of older and younger users in order to learn dialogue policies (Georgila et al., 2008b). So although the system-initiative design of our nine SDS does not reflect the current state of the art, the resulting interaction data can be used to improve cutting-edge systems.

More generally, we conclude that low-level design decisions such as the number of options to present in a single turn or the confirmation strategy to be used can only be taken after the cognitive demands of the task at hand have been analysed in detail. Even with a task as simple as appointment scheduling, the mental operations required depend on the particular goal. In our experiment, users merely had to find a free slot in their calendar and scan the system's messages for a good-enough fit. In the real world, users may plan their appointments in more detail, especially if they need to consider travel time or factor in expected waiting times. As discussed in Section 2.2.2, we would expect that the more complicated those plans are, the higher the cognitive load becomes. The more planning required, the more important it becomes that all relevant information is available to the user, and that key attributes are highlighted properly (Moore et al., 2004; Polifroni and Walker, 2006). These design parameters can be assessed through user modelling techniques (Walker et al., 2005; Carenini and Moore, 2006) and cognitive walkthroughs (Nielsen and Mack, 1994).

A further constraint on system design is the quality of the spoken input processing. Automatic speech recognition may be so unreliable that the disadvantages of explicit confirmations and yes/no questions are far outweighed by the increased reliability. This can only be assessed in experiments that use recognisers with appropriate acoustic and language models. Acoustic models should be tailored to the participants' dialect (Zajicek et al., 2004) and to the characteristics of older voices (Vipperla et al., 2009). Language models should be adapted to older people's speaking styles (Vipperla et al., 2009; Wolters et al., 2009) and to the application domain (Clarke et al., 2005).

Finally, the more systems deviate from normal procedures, the more important navigational aids become. Sharit et al. (2003) found that graphical representations of the menu structure helped users plan their traversal of the IVR systems' menu structure. Likewise, in telephone-based symptom management systems that rely on standardised questionnaires, users may be given small cards that summarise the scale used for the questionnaire (C. Hibberd, pers. comm.). In a small way, this even applies to the dialogue structure adopted in our experimental system. Although it covers the main components of an appointment ("who", "when", "where") in a very transparent way, it does not reflect the usual structure of such dialogues, where the receptionist suggests a time, and the user either accepts or requests another time and/or date. Such mixed-initiative dialogues can become very difficult to handle, especially if the user specifies several alternatives with various restrictions or if the user initially accepts the option offered by the system, only to reject it after further deliberation.

7. Conclusions and future work

Our results show that both older and younger users became more efficient in their interactions with our appointment scheduling system when they were presented with the maximum number of options (four) at a time and when choices were not confirmed explicitly. Notably, this increased efficiency did not come at a price: task success remained at a similar, high level when the number of options was increased. Our detailed questionnaire revealed that users judged the appointment scheduling system along six main dimensions: system reliability, satisfaction, look and feel, perceived efficiency, perceived cognitive load, and unpredictability. These criteria were important to both older and younger users.

A more detailed analysis of the interactions between system and users shows significant difference in interaction styles between older and younger users (Wolters et al., 2009). In future, we plan to annotate the interactions in more detail using the schema outlined by Möller et al. (2007) to uncover usability issues related to misunderstandings, user errors, and system errors.

We also intend to investigate how performance of different modules of end-to-end spoken dialogue systems, such as natural language generation, dialogue management, or automatic speech recognition, affects system ratings on these six high-level factors, using a statistical regression analysis along the lines of, e.g., Walker et al. (1998) and Möller et al., 2007. In particular, we expect to see significant effects of speech recognition performance, since recognising the speech of the older users in our corpus is more difficult than recognising younger users' speech, even when an appropriate language model is used (Vipperla et al., 2009). In our work on improving the design of spoken dialogue systems for older people, we plan to investigate the benefits of user modelling (Moore et al., 2004; Polifroni and Walker, 2006) and strategies for accommodating user initiative (Cohen et al., 1998).

For these experiments, we intend to work with different segments of the older population. Although the users we recruited

for this study match the typical population of undergraduates that is often used in HCI experiments, neither undergraduates nor their older equivalents are particularly good representatives of their generation, since they are generally healthy and well-educated. Moreover, the kind of older person who volunteers for experiments with new technologies may well be likely to be more open to experience and more active than other older people—and hence more likely to age better.

Another aspect we need to account for is previous exposure to technology, in particular speech technology, users' experiences with these systems, and users' attitude to technology in general (Gödde et al., 2008). We plan to address these issues through a mix of questionnaires and semi-structured interviews.

Acknowledgements

We would like to thank our six reviewers and the members of the Cognitive Ageing seminar, Department of Psychology, Edinburgh, for their insightful comments, Neil Mayo and Joe Eddy for coding the Wizard-of-Oz interface, Neil Mayo for technical help with the experiment, Vasilis Karaiskos for administering the spoken dialogue experiment, and Melissa Kronenthal for transcribing the dialogues. This research was supported by the MATCH project (SHEFC-HR04016) and a Wellcome Trust VIP grant to Kallirroi Georgila.

Appendix A. The Questionnaire

The first item, *perceived task completion*, was a yes/no item. The second item, *overall impression*, was measured on a continuous, five point scale. In addition to our detailed factor analysis, we used a single item (bolded below) to provide a global assessment of user satisfaction.

The remaining 37 items were rated on a five-point Likert scale (1 – strongly disagree, 2 – disagree, 3 – neutral, 4 – agree, 5 – strongly agree).

A.1. Achieving your goal

1. The appointment booking system did not always do what I wanted.
2. The information provided by the booking system was clear.
3. The information provided by the booking system was incomplete.
4. Appointments can be booked efficiently with the system.
5. The booking system is unreliable.

A.2. Communication with the system

1. I felt the booking system understood me well.
2. I always knew what to say to the booking system.
3. I had to concentrate in order to hear the booking system correctly.
4. The voice of the booking system sounded natural.

A.3. System behaviour

1. The booking system reacted too slowly.
2. The booking system was friendly.
3. The booking system did not always react as expected.
4. I was not always sure what the booking system expected of me.
5. The booking system made a lot of errors.
6. I was able to easily recover from errors.
7. The booking system reacted like a human.
8. The booking system behaved in a cooperative way.

A.4. Dialogue

1. It was easy for me to lose my way during the conversation.
2. The dialogue was clumsy and unnatural.
3. I could direct the dialogue in the way I wanted.
4. The dialogue was too long.
5. The dialogue led quickly to the desired aim.
6. The dialogue was balanced between myself and the booking system.

A.5. Personal assessment

1. The conversation with the booking system was pleasant.
2. I felt relaxed during the conversation with the booking system.
3. I had to concentrate hard when making the appointment.
4. The conversation with the booking system was fun.
5. *Overall, I am satisfied with the booking system.* (outcome measure; used for global assessment of user satisfaction)

A.6. Usability of the system

1. The booking system was difficult to use.
2. It was easy to learn to use the booking system.
3. Using the booking system to book healthcare appointments was comfortable.
4. The booking system was too inflexible.
5. The booking system was not helpful for making healthcare appointments.
6. I would prefer to make healthcare appointments in a different way.
7. I would use the booking system again in the future.
8. Booking an appointment via the booking system was as easy as booking an appointment via a receptionist.
9. Using the booking system was worthwhile.

References

Baddeley, Alan, 2007. Working Memory in Thought and Action. Oxford University Press, Oxford, UK.

Alexandersson, J., 2008. i2home – towards a universal home environment for the elderly and disabled. K ünstliche Intelligenz 08 (3), 66–68.

Arking, R., 2005. Biology of Aging. third ed. Oxford University Press, New York, NY.

Aylett, M., Pidcock, C., Fraser, M., 2006. The Cerevoice Blizzard Entry 2006: a prototype database unit selection engine. In: Proceedings of the 2nd BLIZZARD Challenge.

Baber, C., Mellor, B., Graham, R., Noyes, J.M., Tunley, C., 1996. Workload and the use of automatic speech recognition: the effects of time and resource demands. Speech Communication 20, 37–53.

Baeckman, L., Small, B.J., Wahlin, A., 2001. Aging and memory: cognitive and biological perspectives. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging. Academic Press, San Diego, CA, etc., pp. 349–377.

Bailey, B., 2000. How to improve design decisions by reducing reliance to superstition. Let's start with Miller's Magic 7 ± 2. <http://www.humanfactors.com> (retrieved 08.02.08.).

Bickmore, T., Giorgino, T., 2006. Health dialog systems for patients and consumers. Journal of Biomedical Informatics 39, 556–571.

Black, L.-A., McMeel, C., McTear, M., Black, N., Harper, R., Lemon, M., 2005a. Implementing autonomy in a diabetes management system. Journal of Telemedicine and Telecare 11 (Suppl. 1), 6–8.

Black, L.-A., McTear, M., Black, N., Harper, R., Lemon, M., 2005b. Evaluating the DI@L-LOG system on a cohort of elderly, diabetic patients: results from a preliminary study. In: Proceedings of Interspeech, Lisbon, Portugal, pp. 821–824.

Bohus, D., Rudnicky, A., 2005. Sorry, I didn't catch that!—an investigation of non-understanding errors and recovery strategies. In: Proceedings of the 6th SIGDial Workshop on Discourse and Dialogue, Lisbon, Portugal, pp. 128–143.

Bond, C., Camack, M., 1999. Your call is important to us, please hold. Ergonomics in Design 7, 9–15.

Bortz, J., 1993. Statistik. sixth ed.. Springer, Berlin/New York.

Brennan, S., 1998. The grounding problem in conversations with and through computers. In: Kreuz, S. (Ed.), Social and Cognitive Psychological Applaches to Interpersonal Communication. Lawrence Erlbaum, Hillsdale, NJ, pp. 201–225.

Bugg, J.M., Zook, N.A., DeLosh, E.L., Davalos, D.B., Davis, H.P., 2006. Age differences in fluid intelligence: contributions of general slowing and frontal decline. Brain and Cognition 62, 9–16.

Carenini, G., Moore, J.D., 2006. Generating and evaluating evaluative arguments. Artificial Intelligence 170 (11), 925–952.
Charlton, R.A., Landau, S., Schiavone, F., Barrick, T.R., Clark, C.A., Markus, H.S., Morris, R.G., 2007. Astructural equation modeling investigation of age-related variance in executive function and DTI measured white matter damage. Neurobiology of Aging 29, 1547–1555.
Clark, H.H., Brennan, S.E., 1991. Grounding in communication. In: Resnick, L.B., Levine, J., Behrend, S.D. (Eds.), Perspectives on Socially Shared Cognition. American Psychological Association, Washington, DC, pp. 127–149.
Clarke, K., Lewin, M., Atkins, D., Kalawsky, R., 2005. Testing a framework for multimodal control in the home environment. In: Proceedings of the IEEE Workshop on Perspectives in Pervasive Computing, pp. 87–95.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. second ed.. Lawrence Erlbaum, Hillsdale, NJ.
Cohen, R., Allaby, C., Cumbaa, C., Fitzgerald, M., Ho, K., Hui, B., Latulipe, C., Lu, F., Moussa, N., Pooley, D., Qian, A., Siddiqi, S., 1998. What is initiative? User Modeling and User-Adaptive Interaction 8, 171–214.
Commarford, P.M., Lewis, J.R., Smither, J.A.-A., Gentzler, M.D., 2008. A comparison of broad versus deep auditory menu structures. Human Factors 50 (1), 77–89. <http://hfs.sagepub.com/cgi/content/abstract/50/1/77>.
Czaja, S., Lee, C., 2007. The impact of aging on access to technology. Universal Access in the Information Society 5, 341–349.
Dahlbäck, N., Jönsson, A., Ahrenberg, L., 1993. Wizard of Oz studies – Why and How. Knowledge-Based Systems 6, 258–266.
Darlington, R., n.d. Factor analysis. http://www.psych.cornell.edu/Darlington/factor.htm (last retrieved 15.09.07.).
Deary, I.J., Whiteman, M.C., Starr, J.M., Whalley, L.J., Fox, H.C., 2004. The impact of childhood intelligence on later life: following up the Scottish Mental Surveys of 1932 and 1947. Journal of Personality and Social Psychology 86, 130–147.
Duff, S.C., Logie, R.H., 2001. Processing and storage in working memory span. Quarterly Journal of Experimental Psychology A 54, 31–48.
Dulude, L., 2002. Automated telephone answering systems and aging. Behaviour & Information Technology 21, 171–184.
Fozard, J.L., Gordon-Salant, S., 2001. Changes in vision and hearing with aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging. Academic Press, San Diego, CA, pp. 241–266.
Fraser, N., Gilbert, G., 1991. Simulating speech systems. Computer Speech and Language 5 (1), 81–99.
Garden, S., Phillips, L., MacPherson, S., 2001. Mid-life aging, open-ended planning, and laboratory measures of executive function. Neuropsychology 15, 472–482.
Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., Moore, J., Watson, M., 2008a. A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In: Proceedings of the 6th International Conference on Language Resources and Evaluation.
Georgila, K., Wolters, M., Moore, J., 2008b. Simulating the behaviour of older versus younger users. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Human Language Technologies (ACL/HLT), pp. 49–52.
Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R., Falavigna, D., 2005. Automated spoken dialogue system for hypertensive patient home management. International Journal of Medical Informatics 74, 159–167.
Gödde, F., Möller, S., Engelbrecht, K.-P., Kühnel, C., Schleicher, R., Naumann, A., Wolters, M., 2008. Study of a speech-based smart home system with older users. In: International Workshop on Intelligent User Interfaces for Ambient Assisted Living, pp. 17–22.
Gregor, P., Newell, A.F., Zajicek, M., 2002. Designing for dynamic diversity – interfaces for older people. In: Proceedings of ASSETS 2002, The 5th International ACM Conference on Assistive Technologies, Edinburgh UK, pp. 151–156.
Hawthorn, D., 2000. Possible implications of aging for interface designers. Interacting with Computers 12, 507–528.
Hockey, B.A., Lemon, O., Campana, E., Hiatt, L.M., Aist, G., Hieronymus, J., Dowding, J., Gruenstein, A., 2003. Targeted help for spoken dialogue systems. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 147–154.
Hone, K.S., Graham, R., 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering 6, 287–305.
Horn, J.L., 1982. The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: Craik, F., Trehub, S. (Eds.), Advances in the Study of Communication and Affect. Aging and Cognitive Processes. Plenum Press, New York, NY, pp. 237–278 (Chapter 8).
Huguenard, B.W., Lerch, F.J., Junker, B.W., Patz, R.J., Kass, R.E., 1997. Working memory failure in phone-based interaction. ACM Transactions on Computer–Human Interaction 4, 67–102.
ISO, 1998. ISO Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs) – Part 11: Guidance on Usability.
ITU-T Rec. P.851, 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.
Kurniawan, S.H., Ellis, R.D., Zaphiris, P., 2002. Comparing older and younger adults' traversal time in expandable and non-expandable hierarchical structures. In: Proceedings of the Human Factors and Ergonomics Society Conference, pp. 185–188.
Lai, J., Karat, C.-M., Yankelovich, N., 2008. Conversational speech interfaces and technologies. In: Sears, A., Jacko, J. (Eds.), The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, second ed. Lawrence Erlbaum, New York, NY, pp. 381–391.
Lemon, O., Pietquin, O., 2007. Machine learning for spoken dialogue systems. In: Proceedings of Interspeech, Antwerp, Belgium.
Levitt, T., Fugelsang, J., Crossley, M., 2006. Processing speed, attentional capacity, and age-related memory change. Experimental Aging Research 32, 263–295.
Logie, R., vander Meulen, M., 2009. Fragmenting and integrating visuo-spatial working memory. In: Brockmole, J. (Ed.), Representing the Visual World in Memory. Psychology Press, Hove, UK, pp. 1–32 (Chapter 1).
Lohse, G., 1997. The role of working memory in the graphical information design. Behaviour and Information Technology 21, 273–280.
Miller, G.A., 1956. The magical number seven, plus or minus two. Psychological Review 63 (2), 81–97.
Möller, S., Engelbrecht, K.-P., Oulasvirta, A., 2007. Analysis of communication failures for spoken dialogue systems. In: Proceedings of Interspeech.
Möller, S., Gödde, F., Wolters, M., 2008. A corpus analysis of spoken smart-home interactions with older users. In: Proceedings of the 6th International Conference on Language Resources and Evaluation.
Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: a case study. Computer Speech and Language 21 (1), 26–53.
Moore, J.D., Foster, M.-E., Lemon, O., White, M., 2004. Generating tailored, comparative descriptions in spoken dialogue. In: Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, pp. 917–922.
Morrow, D.G., Hier, C.M., Menard, W.E., Leirer, V.O., 1998. Icons improve older and younger adults' comprehension of medication information. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 53, 240–254.
Nielsen, J., Mack, R.L. (Eds.), 1994. Usability Inspection Methods. John Wiley & Sons, New York, NY.
Perugini, S., Anderson, T.J., Moroney, W.F., 2007. A study of out-of-turn interaction in menu-based, IVR, voicemail systems. In: CHI'07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM Press, New York, NY, USA, pp. 961–970.
Petrie, H., 2001. Accessibility and usability requirements for ICTs for disabled and elderly people: a functional classification approach. In: Nicolle, C., Abascal, J. (Eds.), Inclusive Design Guidelines for HCI. Taylor and Francis, London, UK, pp. 29–60.
Polifroni, J., Walker, M., 2006. An analysis of automatic content selection algorithms for spoken dialogue system summaries. In: Proceedings of IEEE/ACL Spoken Language Technology Workshop, pp. 186–189.
Pollack, M., 2005. Intelligent technology for an aging population: the use of AI to assist elders with cognitive impairment. AI Magazine 26, 9–24.
Pollack, M., Brown, L., Colbry, D., McCarthy, C.E., Orosz, C., Peintner, B., Ramakrishman, S., Tsamardinos, I., 2003. Autominder: an intelligent cognitive orthotic system for people with memory impairment. Robotics and Autonomous Systems 44, 273–282.
R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rabbitt, P., 1990. Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. Acta Oto-Laryngologica Supplement 476, 167–175.
Rabbitt, P., Anderson, M., 2006. The lacunae of loss? Aging and the differentiation of cognitive abilities. In: Bialystok, E., Craik, F.I. (Eds.), Lifespan Cognition: Mechanisms of Change. Oxford University Press, New York, NY. Ch.23.
Rabbitt, P., Diggle, P., Holland, F., McInnes, L., 2004. Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. The Journals of Gerontology Series B, Psychological Sciences and Social Sciences 59, 84–97.
Rabbitt, P., Scott, M., Thacker, N., Lowe, C., Jackson, A., Horan, M., Pendleton, N., 2006. Losses in gross brain volume and cerebral blood flow account for age-related differences in speed but not in fluid intelligence. Neuropsychology 20, 549–557.
Raven, J., Raven, J., Court, J., 1998. Manual for Raven's Progressive Matrices and Vocabulary Scales. Harcourt Assessment, San Antonio, TX.
Rogers, W.A., Meyer, B., Walker, N., Fisk, A.D., 1998. Functional limitations to daily living tasks in the aged: a focus group analysis. Human Factors 40, 111–125.
Roy, N., Pineau, J., Thrun, S., 2000. Spoken dialogue management using probabilistic reasoning. In: Proceedings of the 38th Annual Meeting of the Association for Computation Linguistics (ACL), pp. 93–100.
Salthouse, T.A., 1991. Mediation of adult age differences in cognition by reductions in working memory and speed of processing. Psychological Science 2, 179–183.
Salthouse, T.A., 1993. Speed mediation of adult age differences in cognition. Developmental Psychology 29, 722–738.
Salthouse, T.A., 2003. Interrelations of aging, knowledge, and cognitive performance. In: Understanding Human Development: Lifespan Psychology in Exchange With Other Disciplines. Kluwer Academic, Berlin, Germany, pp. 265–287.
Salthouse, T.A., 2004. What and when of cognitive aging. Current Directions in Psychological Science 13, 140–144.
Schaie, K.W., 1989. Perceptual speed in adulthood: cross-sectional and longitudinal studies. Psychology and Aging 4, 443–453.
Schatzmann, J., Weilhammer, K., Stuttle, M., Young, S., 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowledge Engineering Review 21, 97–126.
Schneider, B.A., Daneman, M., Murphy, D.R., 2005. Speech comprehension difficulties in older adults: cognitive slowing or age-related changes in hearing? Psychology and Aging 20, 261–271.

Sharit, J., Czaja, S.J., Hernandez, M., Yang, Y., Perdomo, D., Lewis, J., Lee, C.C., Nair, S., 2004. An evaluation of performance by older persons on a simulated telecommuting task. Journals of Gerontolology, Series B: Psychological Sciences and Social Sciences 59, 306–315.

Sharit, J., Czaja, S.J., Nair, S., Lee, C.C., 2003. Effects of age, speech rate, and environmental support in using telephone voice menu systems. Human Factors 45, 234–251.

Sharp, H., Rogers, Y., Preece, J., 2007. Interaction Design, second ed. John Wiley, London, UK.

Sheeder, T., Balogh, J., 2003. Say it like you mean it: priming for structure in caller responses to a spoken dialog system. International Journal of Speech Technology 6, 103–111.

Sjölinder, M., Höök, K., Nilsson, L.-G., 2003. The effect of age-related cognitive differences, task complexity and prior internet experience in the use of an online grocery shop. Spatial Cognition and Computation 3, 61–84.

Stent, A.J., Huffman, M.K., Brennan, S.E., 2008. Adapting speaking after evidence of misrecognition: local and global hyperarticulation. Speech Communication 50 (3), 163–178.

Stephens, E.C., Carswell, C.M., Schumacher, M.M., 2006. Evidence for an elders' advantage in the naive product usability judgments of older and younger adults. Human Factors 48, 422–433.

Suhm, B., Bers, J., McCarthy, D., Freeman, B., Getty, D., Godfrey, K., Peterson, P., 2002. A comparative study of speech in the call center: natural language call routing vs. touch-tone menus. In: CHI'02: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM Press, New York, NY, USA, pp. 283–290.

Traum, D., 1994. A Computational Theory of Grounding in Natural Language Conversation. Ph.D. Thesis, University of Rochester.

Unsworth, N., Engle, R.W., 2005. Individual differences in working memory capacity and learning: evidence from the serial reaction time task. Memory and Cognition 33, 213–220.

vander Meulen, M., Logie, R., Freer, Y., Sykes, C., McIntosh, N., Hunter, J., 2009. When a graph is poorer than 100 words: a comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. Applied Cognitive Psychology.

Vipperla, R., Wolters, M., Georgila, K., Renals, S., 2009. Speech input from older users in smart environments: challenges and perspectives. In: Proceedings of HCI International, San Diego, CA.

Walker, M., Litman, D.J., Kamm, C.A., Abella, A., 1998. Evaluating spoken dialogue agents with PARADISE: two case studies. Computer Speech and Language 12, 26–53.

Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, G., 2005. Generation and evaluation of user tailored responses in multimodal dialogue. Cognitive Science 28 (5), 811–840.

Walker, M.A., Passonneau, R.J., Boland, J., 2001. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, pp. 515–522.

Wechsler, D., 1981. Manual for the Wechsler Adult Intelligence Scale-Revised. The Psychological Corporation, New York.

Wolters, M., Campbell, P., DePlacido, C., Liddell, A., Owens, D., 2007. Making synthetic speech accessible to older people. In: Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 288–293.

Wolters, M., Georgila, K., MacPherson, S., Moore, J., 2009. Being old doesn't mean acting old: older users' interaction with spoken dialogue systems. ACM Transactions on Accessible Computing 2 (1). Available from:<http:/doi.acm.org/10.1145/1525840.1525845>..

Zajicek, M., 2004. Successful and available: interface design exemplars for older users. Interacting with Computers 16, 411–430.

Zajicek, M., 2006. Aspects of HCI research for older people. Universal Access in the Information Society 5 (3), 279–286.

Zajicek, M., Khin Kyaw, Z., 2005. The speech dialogue design for a PDA/Web based reminder system. In: Proceedings of the 9th IASTED International Conference on Internet and Multimedia Systems and applications, Hawaii, pp. 394–399.

Zajicek, M., Morrissey, W., 2001. Speech output for older visually impaired adults. In: People and Computers XV – Interaction without Frontiers. Joint Proceedings of HCI 2001 and IHM 2001. Springer, London, UK, pp. 503–513.

Zajicek, M., Wales, R., Lee, A., 2004. Speech interaction for older adults. Universal Access in the Information Society 3 (2), 122–130.

Zaphiris, P., Kurniawan, S., Ghiawadwala, M., 2007. A systematic approach to the development of web design guidelines for older people. Universal Access in the Information Society 6, 59–75.

Zoltan-Ford, E., 1991. How to get people to say and type what computers can understand. International Journal of Man–Machine Studies 34, 527–547.