A Platform for Multimodal Wizard of Oz User Interaction Studies

Michael CODY
Adaptive Speech Interfaces,
Media Lab Europe
Dublin, Ireland
codym@mle.media.mit.edu

David REITTER
Adaptive Speech Interfaces,
Media Lab Europe
Dublin, Ireland
reitter@mle.media.mit.edu

Fred CUMMINS
Department of Computer Science,
University College Dublin
Dublin, Ireland
fred.cummins@ucd.ie

Abstract

This paper describes WOzOS (Wizard of Oz Operating System), a Java-based platform that can be used in multimodal Wizard of Oz (WOz) experiments. In addition to the platform design, a study using WOzOS is described. Human-machine interactions in a multimodal environment were the focus of this study. WOzOS was used to collect multilingual (English, Portuguese and Swedish) data for research purposes, including for training of a multimodal fusion/fission service.

1 Introduction

In the Wizard of Oz (WOz) paradigm, a user interacts with what appears to be a fully functioning automatic system. Unbeknownst to the user, however, the system responses are generated in real time by human operators who remain unseen (Fig. 1). In this manner, a variety of user interfaces can be prototyped, and user responses to a range of system behaviours can be studied. Such simulation-based research is useful in the design of multimodal interfaces (Oviatt et al., 1992) as well as for the collection of multimodal corpora (Yang et al., 2000). Multimodal data gathered in this manner can guide system design and help in optimizing human-computer interaction.

In many cases, WOz systems are purpose-built to investigate specific systems (Oviatt et al., 1992, McInnes et al., 1997, Wyard and Churcher, 1998). Two research groups have tried to provide more generic platforms, which would allow simulation-based research in a wider variety of contexts. In NEIMO (Coutaz et al., 1996), Apple's *HyperCard* was used to assemble an UI in real time. The system allowed both GUI elements and video, but not sound, to be transmitted and logged. In SUEDE (Klemmer et al., 2000), a speech interface could be generated from a relatively simple toolkit. WOzOS represents a further contribution to the set of tools that can support simulation-based research across a variety of situations. It combines standard GUI elements with spoken input and synthesized voice output.

In this paper, we describe the basic WOzOS architecture, and report on a study we conducted, in which multilingual, multimodal data was collected for research and development purposes.

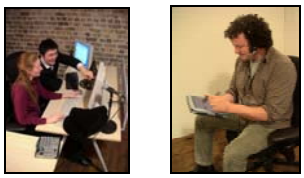

Figure 1: Wizards (behind the scenes) and Client (oblivious to deception)

2 The WOzOS Platform

2.1 Hardware

WOzOS runs on three networked workstations and comprises three separate applications: Operator, Session Manager and Client. Two of the workstations are "Wizard" stations (Session Manager & Operator), the other is the Client interface. Deception is the principle underlying the experiments, so the Wizard stations should be hidden and, preferably, should not be in the same room as the subject. The platform includes a commercial text-to-speech (TTS) module (ScanSoft's *RealSpeak*) for generation of synthetic system spoken output, but this could readily be replaced by an open source TTS module (e.g. *Festival*). In addition to a workstation, the Client also uses a head-mounted microphone and headphones.

In the following description, the application used in our corpus collection

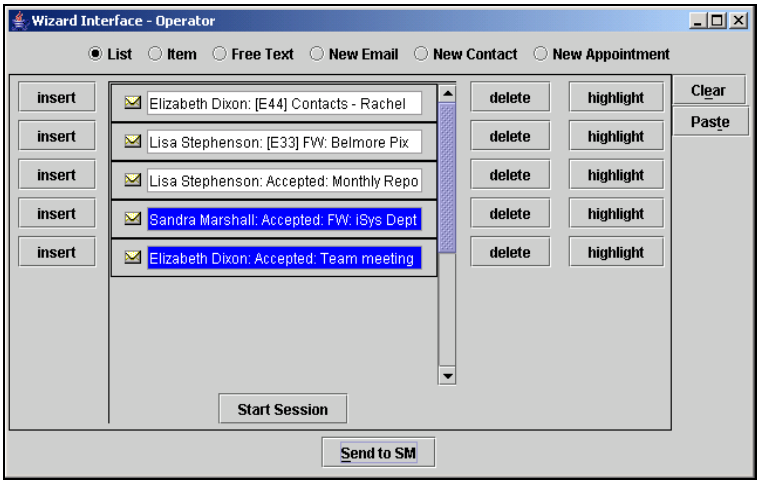

Figure 2: WOzOS Operator application

will be used as an illustrative example. In it, the Client interacts with a Personal Information Manager (PIM), which provides a subset of Microsoft's *Outlook* functionality, augmented by speech output.

2.2 Wizards

The example domain requires the rapid construction of appropriate system responses (graphical and TTS) in real time based on Client requests to the system. A credible system response must be fast; excessive delays could influence subjects' impression of whether they are interacting with a computer system (Oviatt et al., 1992). In WOzOS the general target was a response time of a few seconds at most, this target was achieved; average response time[1] of a typical, sample session was 5.5 seconds, with standard deviation 4.7. To facilitate a quick response the Wizards compose their GUI responses using templates and widgets that can be included or excluded. WOzOS can simulate adaptive systems, where human performance is assumed in all decisions related to how to adapt. In our case, this means that Wizards choose from a constrained set of user interface elements (widgets on the screen, words and phrases by voice), in order to compose the output. The choice Wizards have in reacting makes them a study subject also, and allows for studies of adaptivity, provided the situations triggering adaptation (noise, physical activity) are simulated.

The Wizards assume two roles: the **Operator** retrieves data from an application such as *Outlook*, and prepares the visual representation. In parallel, the **Session Manager** assembles the voice output and ensures that all experimental tasks are carried out. On-line chunking of the data in task sets has proven helpful in navigating the data created.

2.2.1 Operator

The Operator's interface is used to assemble screen content appropriate to the experimental context and Client requests. The Operator can copy content from other application sources and paste it into a screen area. This can then be further edited before sending to the Session Manager.

In the PIM context, within which we have experimented using WOzOS, content like e-mails or contact data may be copied from *Outlook*. The Operator's application automatically arranges the content either as a list (Figure 2) or as a single item. The data from this initial study was used for design of a multimodal PIM application (FASiL project), which is to run on a mobile device e.g. PDA. For this reason the output screen area was kept to a size consistent with such a platform. There is no reason why screen size has to be limited. Indeed, following further development we hope to demonstrate

[1] Response time being defined as time difference between end of Client input and start of system (Wizard generated) response.

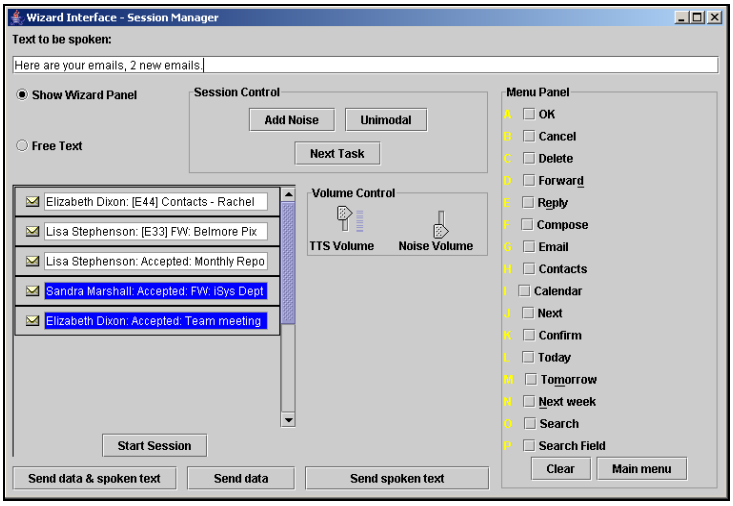

Figure 3: WOzOS Session Manager application

WOzOS in different contexts, using differing sources for content and output screens. To aid and speed up the Wizards' work all application functions have hotkeys associated with them.

2.2.2 Session Manager

The Session Manager application controls the sequence of the experimental session and the output presented to the Client. The Session Manager can set the experimental environment (output modalities presented, TTS/noise volume, background noise), compose TTS utterances and add additional screen elements (buttons) as appropriate to output.

The Session Manager receives screen content (Figure 3) from the Operator and has the option of editing and/or augmenting this as desired. In parallel, the Session Manager can compose TTS utterances to send to the client. TTS utterances can be sent with screen content or independently. TTS utterances are often used as system 'pacifiers' to mask delays in Wizard responses. This TTS-only output can also be used where a unimodal (voice only) context is appropriate. We can add background noise if required by the experiment.

2.3 Client

The Client Application receives the graphical user interface dialogs from the Session Manager (Figure 4). Running on a Tablet PC, the subject can use the GUI on the touchscreen. Text input aids (handwriting recognition, on-screen keyboard) could be used, though they were not appropriate for our experiment.

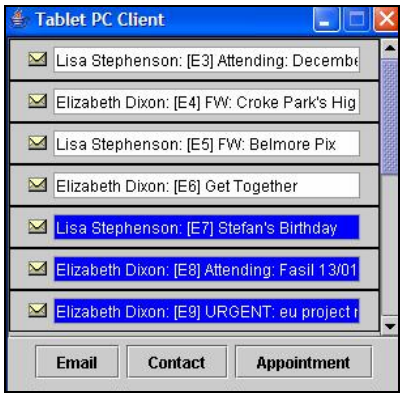

Figure 4: WOzOS Client application

All Client interactions with this application screen (mouse clicks/drags, scrolling, text input) are displayed immediately on the Operator (Figure 5) and Session Manager application screens to allow the Wizards to monitor the Client interaction.

In addition, screenshots showing Client interaction are generated and sent to Session Manager for storage.

2.4 Text-to-Speech

The TTS utterances are output via the Session Manager workstation. To speed up response times and to maintain consistent system utterances, we found it useful to define a set of utterances and to encode these in an abbreviated form, e.g. "pw" is expanded and uttered as "Please wait, I'm checking this".

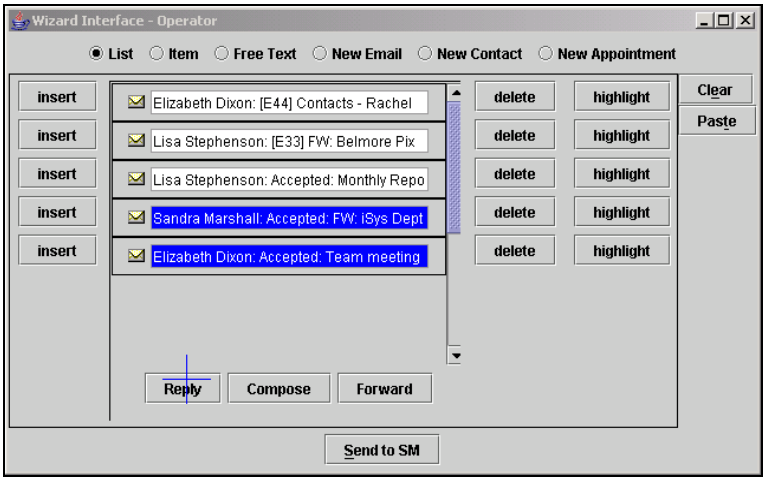

Figure 5: Client interaction (click on "Reply") is displayed to Wizards

2.5 WOzOS session data

All session data are logged on the Session Manager workstation. Time-stamped events of Client and Wizard actions are logged, along with an audio recording of both TTS utterances and Client speech. Screenshots of the screens sent to the Client, and their interaction with the screens are saved and referenced in the log. We provide a tool chain to generate XML data in the *TASX*[2] format from the raw log files generated by WOzOS.

3 Multimodal Study using WOzOS

WOzOS was developed as part of the FASiL[3] project. The R&D vehicle of this project is a multimodal, multilingual PIM application, suitable for use on a portable platform, which is developed by a consortium of European partners.

3.1 Study aims

In general a particular focus of our research has been the coordination of the graphical and voice outputs, which are tailored to meet the demands of a specific situation and user (Reitter et al., 2004). In line with this and the particular objectives of the FASiL project, our study had two separate goals. Firstly, we wished to gather a reasonably large corpus of user interaction with a Virtual Personal Assistant (VPA), which gave access to email, calendar and contact information. This corpus was to be used in training machine learning modules responsible for dialogue management (grammar induction, etc.). Our second goal was to supply an empirical basis for the evaluation of algorithms for multimodal fusion, as well as fission, i.e. the generation of multimodal system output, where models of coherence and pronominalization in multimodal human-computer dialogue play a role. This second goal required the design of a formal experiment in which both available modes and user situation were controlled variables. The SmartKom (Türk, 2001) project had similar motivations[4] for conducting a large multimodal WOz study, however this was only done in one language: German. Our study was carried out in three languages: English, Swedish and Portuguese.

As a first step, we had to devise a suitably constrained set of tasks which novice subjects could be required to complete. To this end, we ran a unimodal (voice only) pilot study (70 subjects, three languages), which helped to refine our subsequent task design.

[2] http://tasxforce.lili.uni-bielefeld.de

[3] Flexible and Adaptive Spoken Language and Multi-Modal Interfaces

[4] SmartKom also uses gestural input, so video of interaction was annonated as part of the corpus. WOzOS presently restricts itself to screen interactions.

3.2 Study description

Three sites (Ireland, Sweden, Portugal) each conducted the study using 30 subjects recruited locally. Subjects were professionals and students (average age: 34.6, standard deviation: 10.6), had prior experience with personal information management software such as *Outlook*, but no specific training in spoken user interfaces. Subjects completed a fixed list of tasks such as "arrange a meeting with Veronica for tomorrow at three to discuss the new recruitment procedures". Each subject completed tasks in a unimodal (voice only) and multimodal condition and in the presence/absence of a noise distractor. Wizards received training; training materials as well as subject tasks were standardized across experimental sites. The detail of the experimental design and results will not be further discussed here.

Subjects found the system easy to use, and were not, in general, aware that they were interacting with people, rather than an automaton. This is somewhat surprising as the system offered near-perfect speech recognition and understanding, which is beyond the abilities of any state-of-the art system.

4 Conclusion

We have described a new Wizard of Oz system that can facilitate the collection of user responses to a simulated system. WOzOS supports both screen-based and voice input and output. Two Wizards collaborate to assemble content and system responses on the fly, allowing for a rich set of behavioural data to be collected. We have used WOzOS to collect interaction data in three languages, and will shortly release the resulting annotated, multilingual corpus through the FASiL consortium (www.fasil.co.uk). One of our FASiL research partners, The Royal National Institute for Deaf People (UK) also ran a parallel study using WOzOS on a large number of hearing impaired subjects of mixed backgrounds and ages. The study design was similar to that outlined above, but obviously had to be adapted for subjects with significant hearing loss. This corpora and WOzOS itself will also be made publicly available

Acknowledgements

We would like to acknowledge the hard work done in putting together the platform and conducting the study. In particular Nathalie Richardet, Erin Panttaja, Steffi Richter & Wei Zhu (Media Lab Europe), Roman Zielinski, Sara Holm & Per Idoff (Cap Gemini, Sweden), Nuno Beires, Luis Almeida; & Rui Gomes (PT Inovação, Portugal), Guido Gybels & Jamie Buchanan (RNID, UK) and the team at ScanSoft.

The authors also gratefully acknowledge the support of the European Commission, grant IST 2001 38685.

References

Joëlle Coutaz, Daniel Salber and Eric Carraux. 1996. *NEIMO, a Multimodal Usability Lab for Observing and Analyzing Multimodal Interaction*. In "CHI'96 Conference Proceedings Companion".

Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker and Annie Wang. 2000. *SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces*, In "CHI Letters: Proceedings ACM Symposium on User Interface Software and Technology", Vol. 2, no. 2, 2000, pp. 1-10.

F.R. McInnes, M.A. Jack, F. Carraro and J.C. Forster. 1997. *User responses to prompt wording styles in a banking service with a Wizard of Oz simulation of word-spotting*. In "Proceedings of IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunications Services", IEE Digest No. 1997/147, pp.7/1-6, June 1997.

Sharon Oviatt, Philip Cohen, Martin Fong and Michael Frank. 1992. *A Rapid Semi-Automatic Simulation Technique for Investigating Interactive Speech and Handwriting*. In "Proceedings of the International Conference on Spoken Language Processing 2", pp.1351-1354.

David Reitter, Erin Panttaja and Fred Cummins. 2004. *UI on the fly: Generating a multimodal user interface*. In "Proceedings of Human Language Technology conference 2004 / North American chapter of the Association for Computational Linguistics", (HLT/NAACL-04).

Ulrich Türk, 2001. *The Technical Processing in SmartKom Data Collection: a Case Study*. In "Proceedings of Eurospeech2001".

Peter Wyard and Gavin Churcher. 1998. *A Realistic Wizard of Oz Simulation of a Multimodal Language System*. In "Proceedings of 5th. International Conference on Spoken Language Processing", (ICSCP98).

Yeonsoo Yang, Masayuki Okamoto and Toru Ishida. 2000. *Applying Wizard of Oz Method to Learning Interface Agent*. In "IECE Transactions Fundamentals", Vol. E00-A, no.1, January 2000.