A REALISTIC WIZARD OF OZ SIMULATION OF A MULTIMODAL SPOKEN LANGUAGE SYSTEM

Peter Wyard and Gavin Churcher

BT Labs, Martlesham Heath, Ipswich IP5 3RE
[peter.wyard, gavin.churcher]@bt-sys.bt.co.uk

ABSTRACT

This paper describes a Wizard of Oz (WOZ) system that allows the realistic simulation of a multimodal spoken language system. A Wizard protocol has been drawn up which means that the WOZ system will simulate the limitations of an automatic system rather than allow the user to engage in the full range of human-human dialogue. In support of this protocol is a sophisticated Wizard response panel and underlying response generation functionality. This enables the Wizard to respond to complex multimodal inputs in near real-time. The chosen application is a 3D retail service, in which users can select furnishings from a database according to colour, pattern, fabric type, etc., transfer furnishings to objects in a virtual showroom, ask about prices and matching of fabrics, etc. The system includes a "virtual assistant", i.e. a synthetic persona which speaks the verbal system output. Users make their input by a combination of fluent speech and touchscreen input. The paper describes a formal trial carried out with the WOZ system, and discusses the results.

1. INTRODUCTION

1.1 The Aims of a Strict WOZ Simulation

The Wizard of Oz (WOZ) technique is a well-known approach to the simulation of fully automatic speech dialogue systems [1,2]. The purpose of performing WOZ simulations is generally to gather information about how users will be likely to interact with a proposed system before that system is fully constructed. This information may be used to help design and build individual system components, to design the user interface, or to design the overall system functionality and style of interaction. Many types of data may be collected, including speech data for training recognisers, language data for language models and grammars, "multimodal data" to design a modality integration module, multimodal dialogue data to design a dialogue manager, subjective user reactions, and objective user performance measures. The last two may be used to design the user interface, the overall system functionality and the style of interaction.

Although there is general agreement about the aims of WOZ experiments, there are differences over how strictly a WOZ system should seek to simulate accurately a real system. Some researchers do not make a serious attempt to simulate an actual automatic system (the "loose approach"). Either the simulated system properties may be indistinguishable from those of a human, or there may be an attempt made to limit the simulation in some respects, but not according to a well-defined protocol.

We believe that, for our purposes at least, it is important that a genuine attempt is made to simulate the limitations of a target automatic system (the "strict approach"). This is because if the system appears to have fully human capabilities, the user's inputs will be more complex, use wider vocabulary, be more idiomatic, etc. than if clear limitations become evident. The language (and other data) collected will then not be as useful as it might be for building real system components, since it is likely to be well beyond their current capability.

Although the loose approach may be favoured simply because it is much easier to implement than the strict approach, it does have some other benefits. It may give information about how users would *like* to interact with a system, both in terms of dialogue strategy and language, which is not available from the strict approach, because the latter is very constraining. In fact, the strict approach also gives this information, since the user is not aware of the system limitations initially and only gradually tends to restrict her input to get the system to work. Thus the early part of the data from a particular user often gives an idea of how they would like to interact.

The loose approach to WOZ simulation may be followed up with rapid prototyping of real systems to gain further data, and this is a common and workable approach. However, we wished to explore the possibility of giving the real system components a significant "head-start" by strict WOZ simulation. A further benefit of the strict approach is that it is possible to gain useful information about user responses to error conditions (i.e. where the system is unable to process the user's input for whatever reason).

1.2 Difficulties of a Strict WOZ Simulation

Prompt and response systems, which have a finite state dialogue and often a fairly small set of input phrases, are in principle not hard to simulate using a WOZ system [3]. In the case of an advanced spoken language system, which also has a flexible dialogue management system, the problems of accurate WOZ simulation are much greater. Oviatt et al [4] describes previous work in this area, but no thorough attempt was made to simulate limitations of speech understanding by the WOZ system, and the scope of their system and the tasks given to the subject were more limited than in the work described in this paper.

A realistic WOZ simulation of an advanced system requires two key things. Firstly, a sophisticated Wizard response panel and infrastructure (Section 2). Secondly, an appreciation that the simulation of components like the grammar and the dialogue manager will only be approximate, and a working assumption that the approximation will be close enough to make the exercise worthwhile in gathering realistic data.

2. EXPERIMENTAL METHODOLOGY

2.1 System Description

The Wizard of Oz system comprises three components: a user interface, a wizard response panel and a Natural Language (NL) server.

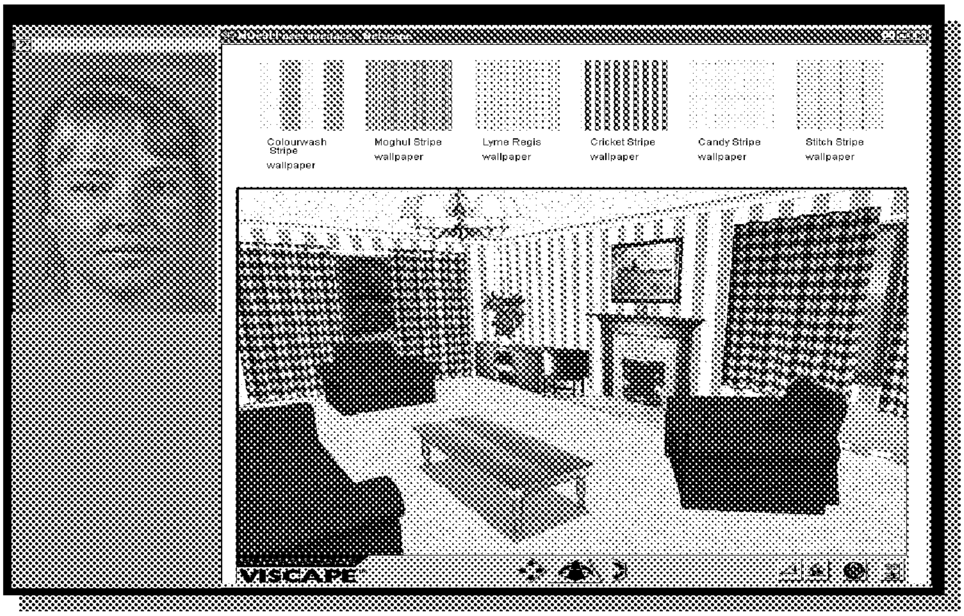


Figure 1: User Interface

The user interface (Figure 1) is derived from an existing non-speech 3D retail system developed at BT Labs which allows a user to browse fabrics and then place them onto objects in the virtual world. It consists of two windows; the first shows the Virtual Shop Assistant, which is a 2D talking head developed by BT Laboratories and is able to lip-sync and make small facial movements whilst playing Text-to-Speech from the Laureate TTS system [5], The second window is a World Wide Web browser containing a Java applet which displays a palette containing up to six fabrics along with the fabric name and the class of objects the fabric is applicable to, and a Superscape Viscape 3D world. The 3D world depicts a scene from a typical sitting room. The sofa and chairs, the curtains and the walls are active and may be touched by the user. The NL server receives touch information from the user interface, determines which objects or fabrics have been touched, and then sends this information to the wizard response panel.

The wizard response panel (Figure 2) comprises four main sub-panels: error response, fabric descriptor, history and function panel. The error response panel contains seven buttons that the wizard can select in accordance with the wizard protocol filter described in detail below. Six fabric descriptors can be used: fabric name, colour, pattern, material, class of object the fabric can be applied to, and pricing constraints. The wizard is able to type the first letter or two into a descriptor's text box, and a list automatically shows the nearest match.

The history panel assists the wizard in two ways: it records the touches the user makes in the virtual room or on the palette, and also keeps a record of previous interactions. The wizard can keep thus track of the local discourse context and can interpret the user's requests correctly. For example, if the user touches the sofa in the virtual room, the history displays a message showing that the sofa has been touched; it also shows which fabric is currently on the sofa. With a click of a button the wizard can quickly transfer attributes such as the fabric name and colour to the descriptor panel. For example, imagine the user said "put that on the chair" <touches the sofa>. The history panel will display the touch event as "11:23:09 User touches sofa (Cartouche Stripe)". The wizard simply has to click the 'Fabric' button on the history panel, which transfers the arguments for Cartouche Stripe to the descriptor panel. He or she then selects the destination object (chair) and finally triggers the 'transfer fabric' on the function panel. All events are timestamped to assist the wizard in determining the local context.

The wizard's actions send commands to the NL server, which in turn instructs the user interface to change the fabrics displayed on these objects. Future implementations will also allow the server to move objects around in the room and to change the user's viewpoint

The NL server consists of a number of components: an internal representation of the 3D world, a response generation module, and various database lookup modules. The server is able to return a number of fabrics which meet certain specified criteria, for example within a colour range, of a particular material and so on, but is also able to relax these criteria in a controlled manner so that at least one fabric is returned for each query. These relaxed constraints are passed onto the response generation module. The server also logs all communications between the user interface and wizard panel (such as individual touches by the user on the screen, and functions triggered by the wizard) for transcription and later analysis.

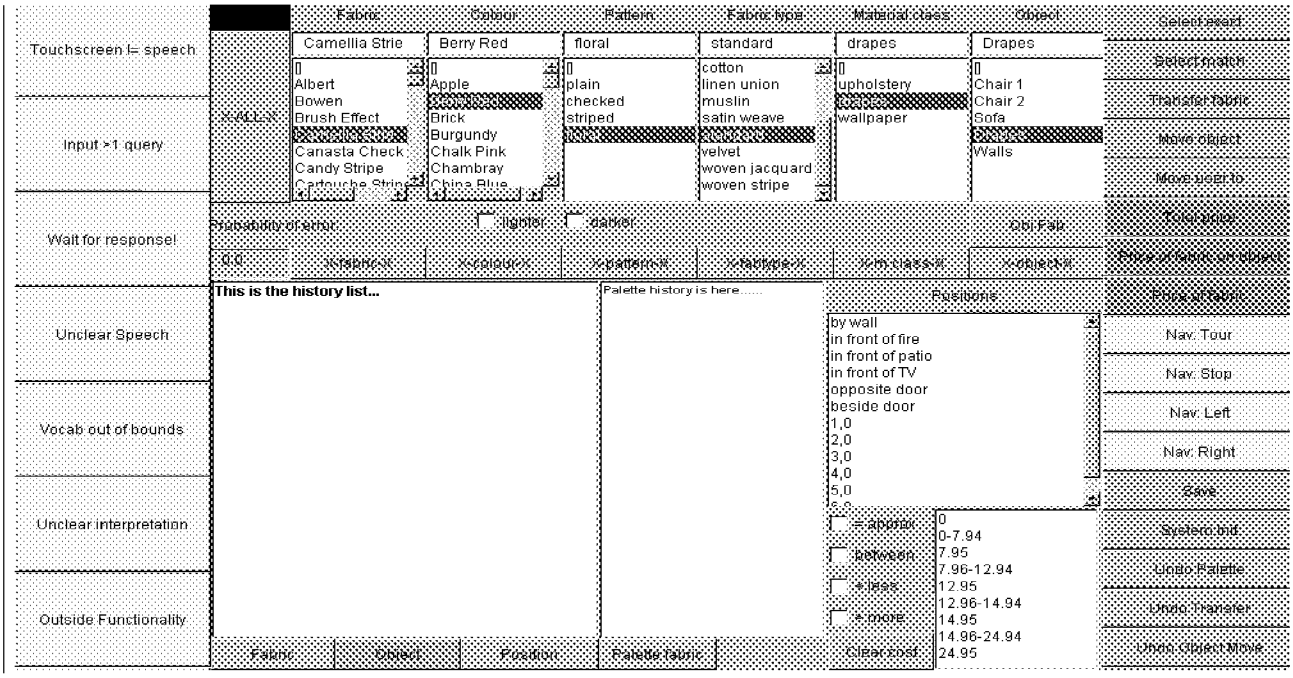
2.2 Wizard Protocol

The wizard employs a seven-stage filter which is applied sequentially when assessing whether user input could be processed by the simulated system. The wizard response panel is designed to assist in this mental process by showing a sequence of error buttons corresponding to these stages. If the filter is passed through without triggering an error then the wizard processes the input. The questions the wizard must ask him/herself are:

1. Can the user's touch and speech be interpreted as a combined event?
2. Has the user specified a single query in their request?
3. Has the user waited for a response from the system before offering more input?
4. Is the user's speech clear, with no false starts, hesitations etc?
5. Is the user's vocabulary and grammar within that of the target system as specified by legal/illegal examples?
6. Can the user's request be interpreted through context or is it ambiguous/unclear?
7. Is the user's request outside of the scope of the system's functionality? (as per paper specification).

The filter illustrates the tight restrictions that are placed on the user input to pass only those which would be interpreted by the target system. In a fully automatic system it might sometimes be difficult to detect which stage has been violated.

Each stage has an associated error response that is automatically generated by the NL server. Currently the response does not take context into account so each generates a single spoken response. Future implementations will make use of more sophisticated response generation techniques, involving context to give more helpful advice. Although the filter is sequential, as the wizard becomes familiar with the protocol, he or she is able to make a quick assessment without having to go through the filter one error at a time, hence the application of this filter protocol is rapid, allowing a near real-time response from the system.



Fabric: Camellia Strie
Colour: Berry Red
Pattern: floral
Fabric type: standard
Material class: drapes
Object: Drapes

This is the history list...

Palette history is here......

Figure 2: Wizard Panel

2.3 Physical Experiment Set-up

A lab comprising two adjacent rooms divided by a one-way mirror was used for the trials. The wizard sat in the control room and was able to see into the subject room through the mirror. A video camera was erected in the subject room to provide a live 'over the shoulder' view of the user interface and the user's hands as they touched the screen. The video feed was displayed to the wizard on a monitor and was simultaneously recorded directly onto Super VHS. At the same time, a high quality microphone placed above and to the right of the user's head recorded sound onto DAT. The wizard monitored the sound via headphones. The wizard is assisted by an experimenter in the subject room who outlines the capabilities of the system, acts as timekeeper and conducts a debriefing interview after the trial. As discussed below, the experimenter can provide help to the user, but must avoid priming them. There were 11 subjects, 5 of which were women, and all were recruited from BT Laboratories. Hence, the subjects were generally fairly familiar with computers.

To assist in the later transcription of the user's interactions with the system, the NL server logs are synchronised at the beginning of each user trial by the experimenter who triggers a function which aligns the inputs with the DAT.

2.4 User tasks

The users were asked to complete a series of tasks which involved selecting and applying fabrics to items in the virtual room. Before the trials were started, the experimenter outlined the system's capabilities, but deliberately did not instruct users on how to interact with the system. This was to avoid priming to one particular mode of operation. Priming is a key issue in

Wizard of Oz experiments; they are designed to collect 'typical' behaviour of the user, and so must avoid pre-conditioning this behaviour. The experimenter provides assistance only when absolutely necessary, and even then must avoid directly instructing the user on how to interact with the system, but give more indirect help such as rephrasing an error message.

The users then began a simple 'warm-up' task, followed by up to nine main tasks. There was no pressure to complete them all; working at their own pace users completed the tasks sequentially until they ran out of time. We chose task-based trials rather than allow the user to randomly browse as we wanted to elicit a wide range of user behaviour. The tasks required the user to find and apply fabrics and wallpapers according to a number of criteria, including colour, pattern and fabric name. A number of the tasks required the user to select their own fabrics and wallpaper constrained only by a given colour scheme. Selection of materials according to a budget also played a role; when the user overspent, he or she had to select an alternative, cheaper fabric. Again, to avoid priming, the tasks were described using minimal text and mostly consisted of a colour screen shot of the room in the fabrics that we wanted the user to select.

3. RESULTS AND DISCUSSION

The Wizard of Oz trials produced eleven sets of interactions, each lasting approximately 40 minutes. Only one of the users suspected that the system might not be fully automatic, although other workers have reported that people are easily fooled! The interactions were recorded on a number of media as described in Section 2.3. The experimenter conducted a post-trial debriefing interview and questionnaire which assessed the user's reaction to the system. Orthographic transcription of the audio has yielded a database of about 1200 user utterances, which is currently being used for robust parsing and grammar learning experiments. Further work is required to produce a unified database with transcribed speech and timed touch events.

The interviews and questionnaires gave much information about the users' perception of the system. In brief summary: the majority of subjects liked being able to use speech combined with touch; only one thought that using existing 'computer' technology (i.e. mouse, keyboard) would be preferable. Some subjects did not fully grasp the metaphor used – a shop assistant that you could talk to naturally – but were constrained by expectations of using desktop computers. In terms of using touch and/or speech modalities, two subjects consistently used pointing only, three subjects did not point at all to items in the virtual world and instead used speech, and three extensively used implied reference, for example, "Show me some striped wallpapers" followed by "try this one" <user touches palette>.

Analysis of the videos highlighted the limitations of the fabric descriptors and comparatives understood by the system; users wanted to ask for "wider stripes" or "paler material" and subsequently had difficulty selecting the fabric or wallpaper that they wanted. Users' reaction to the virtual assistant was useful in that it showed that prompts, in particular error messages, need to be carefully crafted in order to be understood and accepted. It appears that maxims such as only giving enough information and not being overly verbose are applicable in this domain.

4. CONCLUSIONS AND FUTURE WORK

The main conclusion is that the aim of a realistic WOZ simulation of an advanced multimodal spoken language system was successfully achieved. Two separate wizards (with spoken language expertise) were able to learn how to operate the system and follow the protocol with a few hours of training. The wizards were able to respond to the user in near real time, thanks to the response panel, and did not suffer from cognitive overload. The wizard responses were both fairly self-consistent and fairly consistent between each other, although no quantitative measures of this are available. We believe that the wizard protocol (Section 2.2) was followed sufficiently accurately that the data will be useful for real system. Since the protocol itself is imprecise, it is not possible to measure accurately how well it was followed. However, the data shows many examples of each stage of the protocol being violated by the user, an error response being generated, and subsequent modification of user behaviour and language. This produces the double benefit of indicating where we should consider extending the target real system, and producing data which, while fairly expressive and fluent, is also tractable by current parsers.

The next step in our programme is to produce a modified WOZ system with the obvious defects removed, such as the excessive delay in the response time of the talking head, and the lack of functionality in being able to display "the others" or "some more" after the system has said "There are more than twenty red checked fabrics. Here is a selection." We will then run a second WOZ trial and proceed to build a fully automatic system using data from both WOZ trials.

5. REFERENCES

[1] Fraser, N.M. and Gilbert, G.N. 1991. "Simulating Speech Systems", Computer Speech and Language 5, pp 81-99.

[2] Dybkjær, L. and Dybkjær, H. 1993. "Wizard of Oz Experiments in the development of the dialogue model for P1", Technical Report 3, Center for Cognitive Informatics, Roskilde University.

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

[4] Oviatt, S., Cohen, P., Fong, M., and Frank, M. 1992. "A rapid semi-automatic simulation technique for investigating interactive speech and handwriting", Proceedings of the International Conference on Spoken Language Processing, 2, 1351-54

[5] Page J.H. & Breen A.P. 1996. "The Laureate Text-to-Speech system: Architecture and Applications." BT Technology Journal, Vol. 14, No 1, January.