DEVELOPING ROBUST, USER-CENTRED MULTIMODAL SPOKEN LANGUAGE SYSTEMS: THE MUESLI PROJECT

Gavin Churcher and Peter Wyard

BTexaCT, Adastral Park, Martlesham Heath, Ipswich, England

ABSTRACT

The Multimodal Spoken Language Interfaces (MUESLI) project at BTexaCT aims to conduct practical research into developing advanced multimodal spoken dialogue systems with a distinctive user focus. In this paper we intend to convey the main aims and motivation behind the project and to describe the particular research application we have built. We recently conducted a user trial where a fully automatic version of the system was used, allowing us to investigate how users make use of the input modalities and to reveal the current limitations of the system.

1. INTRODUCTION

This paper describes the Multimodal Spoken Language Interfaces (MUESLI) project at BT Adastral Park. The project aims to conduct research into multimodal spoken dialogue systems which has a particular focus on usability. To this end, the project has involved users in the evaluation of successive stages of development, from storyboarding through to a fully automated system.

In order to support our aims, we have built a multimodal advanced spoken language interface to a useful multimedia application, in order to act as a research vehicle. (Note that in addition we have a programme of more formal research experiments to address particular issues which arise, such as user preferences for different synthetic personae, or the effect of multiple modalities on the nature of the dialogue). We are focusing on human-machine dialogue systems, in distinction to other multimodal work on human-human communication in shared spaces.

This paper will describe the design methodology and motivation behind the MUESLI project culminating in a description of the first user trial of the automatic system and its outcome. The system provides a rich, interactive environment from which guidance can be drawn on the use of robust, co-operative multimodal systems in practical applications.

It has long been our assumption that natural spoken language will add greatly to the effectiveness of human-computer interfaces, although in a different way. Multimodality should yield both synergy, where the combined effect of the modalities is greater than their sum, and also error recovery, where for example speech recognition errors can be recovered from using information from the other input modalities.

2. THE MUESLI SYSTEM

The MUESLI system [1,2] allows a user to decorate the furniture, walls and curtains of a ‘virtual’ sitting room. The user can select fabrics using a number of criteria including the colour, pattern, material and price, and apply them to the objects in the room. The complexity of the criteria means that the use of traditional interfaces based around pull-down menus would be cumbersome. The MUESLI system enhances traditional windows-style interfaces with speech recognition and touch input through a touch screen. System intelligence and a dialogue component enables a mixed-initiative conversation between the user and an on-screen agent about the fabrics they choose. The user is also able to navigate through the virtual room using speech, and is able to refer to objects by speech or touch. The primary interaction metaphor has been changed from controlling a screen to conversing with the Virtual Assistant. Figure 1 shows the multimodal system interface with its 2D fabric palette, 3D virtual world (VW) and virtual assistant. The BT talking head used incorporated the BT Laureate Text-to-Speech system [3].

Fabrics may be selected according to a wide range of criteria, and placed on furniture, curtains and walls in the virtual living room or displayed on the palette. The user can refer to objects in the virtual room and to fabrics on the palette by speech or touch, and talk about them with the assistant. The user can combine speech and touch inputs in a single dialogue turn. Users can ask for prices of fabrics or complete selections, for cheaper fabrics, etc. and can ask for advice on which fabrics would suit with a given fabric. The user can also move around the room and inspect the contents from different distances and angles and can move furniture within the room. A natural dialogue is possible with the assistant, catering for various forms of reference (anaphor, deixis, etc.) and ellipsis. The assistant is able to take the initiative at certain points in the dialogue, currently to offer fabric matching advice and to suggest movements within the VW.

A sample of typical dialogue will highlight some of the features of the system:
User: Show me some plain green fabrics in linen.
System: There are more than ten fabrics. Here is a selection. [displays a selection of 6 fabrics on palette].
User: Put this one on the curtains [points to Plain Slub Velvet fabric on palette].
System: Plain Slub Velvet is not a drapes fabric.
User: Put it on the sofa. (anaphor).
System: [transfers Plain Slub Velvet to the sofa, pause].
Would you like to view the room from the door? (system initiative)
User: Move this chair over here. [points to chair then position]. (user ignores system)
System: [moves chair to new position]
User: How much is this selection?
System: This design costs three thousand, five hundred and fifty pounds.
User: What cheaper upholstery fabrics do you have?
System: Here is a selection of cheaper upholstery fabrics.

3. THE MUESLI PROGRAMME

From the onset of the project we have taken a user-centred approach to investigating how input modalities are combined in a way to produce a new powerful synergistic interface.

Neither the technology nor the human factors issues associated with multimodal interfaces are very well understood. We therefore considered it best to proceed in a staged approach, from paper exercises to a fully automatic system, as follows:

1. Storyboarding possible system interactions with a focus group, using pictures of the system's output screen.
2. Exploring different system designs with a focus group, using a set of pre-canned (but real) output screens.
3. Performing Wizard of Oz simulations of successive versions of the system with a user.
4. Constructing a complete trial system.
5. Performing an incremental series of user trials.

This iterative design methodology has two main benefits. Firstly, it enables us to maintain a user focus, and adapt the system we build to what users find helpful, rather than being purely technology-driven. Secondly, the system remains more flexible longer into the programme, and is therefore more useful as a research vehicle for answering our basic questions about multimodality.

Storyboarding is a low-cost, quick and flexible approach which allows a wide range of ideas to be explored, but its disadvantage is that it is very unlike the real system, and may not give accurate user predictions as to what they would like. The next stage was a simulated system that the focus groups could comment upon. By mocking up the output screen using Macromind Director, the focus group could watch a set of canned interactions between the demonstrator and the system. The main advantage of using focus groups in this way is that it provided us with a wide range of user opinion on both the interface and the system’s proposed functionality. One drawback of this approach is that the users’ interactions are passive and may therefore be different from when using the system themselves.

3.1 Wizard of Oz

Stage three of our programme used the Wizard of Oz (WoZ) methodology. This allows users to interact with what they believe is a real system, without having to create the entire system first. It is a half-way house between the focus groups and the fully automatic trial system. The WoZ simulation is less flexible and more expensive than the scripted scenario, but users’ reactions may now be expected to be very similar to those in response to a real system.

In the context of the overall MUESLI programme, it is worth noting that the WoZ system is not just a ‘smoke and mirrors’ mock-up of the system. It is a carefully crafted simulation of a fully automatic system, and it contains a number of components which are reused in the automatic system, such as the response generation module, all the components of the user interface and the communication infrastructure between the user interface and the interaction server.

The eleven users who took part in the Wizard of Oz trials provided us with data that could be used for both language and dialogue modelling. The sessions were recorded on video and audio, and later transcribed. Post-session questionnaires and interviews highlighted the users’ opinions of the system. This information proved valuable in the design of the fully automatic system. A detailed description of the Wizard of Oz trials, results and analysis is given in [4].

3.2 Fully Automatic System

The full system reused many of the components from the Wizard of Oz system. The human wizard was replaced by the spoken language understanding and dialogue components; the remainder of the system could be reused in a fully automatic version with the addition of a speech recognition component, the multimodal interpreter and an appropriate dialogue manager.
Understanding speech and gesture and reasoning over its possible interpretation is a key functionality. At the full system’s disposal is the same input information that the wizard had: touch and speech, domain knowledge and a model of the world. Speech, once recognised, is passed onto a semantics module that can create multiple interpretations of what was said; touch is assigned semantics, too. An event handler decides when the user’s turn has finished and processes the interaction; it passes the multiple semantic hypotheses to the modality integrator. Using a unification-style approach, the hypotheses are combined, and in the case of conflicting information, discarded. The set of unified hypotheses are then passed to the dialogue manager which can determine the most likely hypothesis and react accordingly, depending on the dialogue context. The effective integration of different modalities, their so-called synergy, is one of the more interesting and complex parts of creating the fully automatic system. For a discussion of the approaches to multimodal integration and understanding, see [5]. Additional discussion and a description of the approach taken by the MUESLI system can be found in [2].

4. TRIAL OF FULLY AUTOMATIC SYSTEM

The user trials for the fully automatic system followed a similar line to those of the Wizard of Oz trials. Our intention was to run successive trials with a range of users who had varying degrees of experience in using computer interfaces. Our first trial, described here, used people from BTexaCT who had varying experience of computer interfaces, including using speech recognition products, but none had any experience of advanced multimodal interfaces such as the MUESLI system. There were 7 users: 4 male, 3 female.

The seven users managed to complete, to their satisfaction, all of the tasks that were given to them, taking approximately 40 minutes per user. This gave us 1,100 input sentence recordings on DAT. Users in fact completed 70-80% of the tasks correctly when in fact they had not; for example, the fabric they had chosen was a plain blue but not the correct plain blue as given graphically in the task description.

4.1 Results

The seven users managed to complete, to their satisfaction, all of the tasks that were given to them, taking approximately 40 minutes per user. This gave us 1,100 input sentence recordings on DAT. Users in fact completed 70-80% of the tasks correctly since a number of them ran out of time before completing all of the tasks. Some users believed that they had completed a task, when in fact they had not; for example, the fabric they had chosen was a plain blue but not the correct plain blue as given graphically in the task description.

Speech Recognition Results

Overall, the Nuance speech recogniser gave as a word accuracy that varied between 71% to 85%. Our results showed that word accuracy broadly follows a distribution where the accuracy is low for the warm-up task, but then improves and typically
levels off from the first main task. This can be attributed to users’ language adjustment during the warm-up task. Analysis of the transcriptions shows that a number of the users oversimplified their language after initial recognition difficulties. To the users, it was not clear why their utterances were not accepted: the utterance may not be in the system’s language models, or the speech recogniser did not decode the utterance with high enough confidence. One particular user tried an utterance which was in the language model and would have been successful except that it was rejected by the speech recogniser. The user proceed by grossly simplifying his language and trying again. Since this was successful, the user consistently used this simplified form for the same task.

We found that switching over speech recognisers from Nuance to the human wizard brought about a drastic change in the language of some of the users. The users, at this point, were told that the second recogniser was significantly slower, but more accurate than the previous. Since the recogniser client interfaces were identical, no-one suspected that a human was being used. Not only did we find that the users’ language changed, but also their expectations of the capabilities of the system. For instance, one user went from explicitly telling the system to put particular fabrics on items of furniture to asking the system to choose one automatically from a selection.

Whilst overall speech recognition accuracy with the Nuance recogniser was moderate, we found that the addition of the touch modality allowed the users to successfully complete the current sub-tasks they were attempting. This was shown either by recovering the next turn by modifying their language to include a gestural reference, or by using the best combination of speech and gesture that worked. This flexibility of choice over the modalities contributed to the users’ view of the usability of the system (see below).

Users’ Opinions

The basic concept of multimodal speech and touch input was very favourably received, and people were happy with the idea of talking to a system like this, although in previous focus groups people had expressed concern about speaking to a computer system in a public place. Most users felt that the balance of speech and touch was about right.

We asked users about the amount of spoken feedback from the Virtual Assistant, because in the Wizard of Oz trial quite a number had thought that there was too much. In this trial one person strongly agreed that there was too much feedback, but the others disagreed. The feedback was limited to help towards error recovery (i.e. the system was not proactive in suggesting fabrics or making other initiatives). We feel that it would be useful if users could select the degree of error recovery help that the system provides.

A central metaphor of the MUeSLI system is that the user is engaged in a conversation with the Virtual Assistant. Two users agreed that speaking to the Assistant felt easy and natural, but the others disagreed, some strongly. This is an important area for system improvement. We feel that people would be quite happy to talk to the system, but many do not appear to engage in an easy, natural conversation. We need to discover how much this is due to limitations in the agent used (its appearance, voice, etc.) and how much to limitations in the recognition and language technology.

Regarding alternative input technologies that might have been provided instead of speech and touch, one user would have preferred a pointer device instead of the touch screen, but all the others disagreed. Again, one user would have strongly preferred a mouse and keyboard, but all the others disagreed.

Users were almost equally split as to whether they had to do too much speaking in order to use the system. This result suggests that the capability to customise the interface into multiple versions, e.g. where conversation was stressed, and another which was as functional and efficient as possible, making heavy use of non-verbal actions and shortcuts, could be beneficial.

Comments regarding the usability reflect the fact that the weakest part of the system is still the speech recognition. This highlights the importance of robust natural language understanding and the benefits of multiple input channels. It should be remembered, however, that the MUeSLI system deliberately put a heavy emphasis on spoken language, as this was our dominant research interest.

5. CONCLUSION

This paper has outlined the MUeSLI programme and illustrated the benefits of a user-centred approach to advanced multimodal interface design from storyboarding through to Wizard of Oz and finally trials of the fully automatic system. Using iterative trials permits researchers of multimodal systems to develop systems that are increasingly robust and usable, whilst also providing valuable insight into actual use of such advanced systems.

6. REFERENCES