WOZ experiments in Multimodal Dialogue Systems

Pilar Manchón Portillo
University of Seville
p.manchon@indisys.es

Guillermo Pérez García
University of Seville
g.perez@indisys.es

Gabriel de Amores Carredano
University of Seville
jgabriel@us.es

Abstract

This poster describes a new implementation of a multimodal dialogue system in the Home Machine Environment and the platform developed to conduct the set of experiments designed to model the system's behaviour in this scenario. The research carried out in this paper has been partially funded by EU Project Talk (Contract No 507802) and the Spanish Ministry of Science and Technology under Project TIC2002-00526.

1 Introduction

The objective of these experiments is to extend an existing spoken dialogue system integrating new input and output modalities. In order to achieve this goal, we have designed a WOZ platform where several experiments will be conducted. The experiments' design will be discussed and justified.

2 System Description

The original system is based on the Information State Update approach and has been especially designed to deal with Natural Command Languages. It consists of a number of OAA (Open Agent Architecture) agents which share information and perform different tasks according to a predefined overall dialogue strategy.

The dialogue history is recorded and taken into account in order to disambiguate subsequent utterances and achieve a more natural Human-Computer Interaction. The system can deal with multiple coordinated commands, spontaneous corrections, anaphoric reference resolution and several additional spoken-dialogue-related phenomena.

With regard to the chosen scenario, this particular system application can control several types of devices: lights, music, fan, dimmers, blinds, telephone (…).

The final objective is to integrate additional modalities in the system whereas at the same time allowing for greater flexibility, efficiency and naturalness in the overall interaction. This presents a great deal of additional complexity, since not only all modality-dependent issues have to be addressed independently but also the new issues arisen with the integration of modalities must be taken into account.

The system will deal with both graphical and spoken input, as well as a combination of the two:

- "Turn the lights on" (spoken input)
- Click (graphical input)
- "Turn **this** on" + Click (multimodal input)

3 Experiment Description

The objective of the experiments to be conducted is to record the interactions between human users and the wizard from different perspectives, in order to gather information to configure the basic system.

The experiments will take place in our labs and the special setting is described below.

Completely naïve subjects will provide reliable data about the first reaction of an untrained user before becoming familiar with the system. At the same time, as the subjects become more familiar with the system, we will learn about efficiency and learnability. The analysis will include among other issues:

- possible obstacles or difficulties to communicate
- biases that prevent the interactions from being completely natural
- corpus of natural language in the home domain
- modality preference in relation to task
- modality preference in relation to system familiarity
- task completion time
- combination of modalities for one particular task
- inter-modality timing
- multimodal multitasking

The subjects will initially be given just enough information to perform the tasks, but will not be given precise instructions as to how to proceed with the system. They will be given very general information such as "you may talk to the system", "you may select things by touching the screen" or "you may do both things at the same time". In subsequent phases, the subjects will be provided with more and more information as they become familiar with the system.

As far as the subjects are concerned, they will be interacting with an intelligent multimodal dialogue system and no other human will be involved. They will be provided with one task at the time that will appear on a computer screen. They will be alone in a room especially prepared for the experiment. There will be:

- a touch-screen
- a microphone
- speakers
- a camera
- several devices
- a list of tasks
- a general description of the situation

The interaction between subject and system will be recorded from all perspectives. The camera abovementioned will video record the experiment. Special software will be used to record the touch-screen activity and all agents in the experiment set-up will log all their actions

The wizard will be out of sight but will be able to hear what the subject says and see their touch-screen. Although the subject's input will also be processed and logged by a speech recognition engine, the wizard will pretend to understand everything (within a predefined set of guidelines), excepting a few artificially introduced recognition errors. In response to the subject's actions, the wizard will produce speech, display a written message or image, execute an action, or any combination of the former. When producing speech, the wizard will use synthetic speech or pre-recorded prompts.

4 Platform Description

4.1 Hardware:

a. Wizard computer
The wizard has two main roles: interaction with the user simulating the real system, and control of the physical devices.

b. User Tablet PC
The user will be requested to perform tasks within the home machine environment and this tablet PC will allow her to access the graphical display, speak or both.

c. WiFi router
The user's Tablet PC and the wizard computer will communicate by means of a WiFi router that will allow the user to move freely around the room.

d. Home devices
Our lab setting includes a number of lights and a blind connected to X10 modules. A security camera is simulated with a pre-recorded video. A telephone terminal is also simulated on the screen.

4.2 Software

4.2.1 Wizard Agents

a. Wizard Helper
This is basically a control panel that enables the wizard to:

- Talk to the user. The panel is connected to a TTS running on the user's computer. The wizard can either choose among a number of possible sentences (previously determined according to the possible actions of the user) or type an alternative answer if the user's behaviour differs from what had been foreseen.

- **Remotely play audio and video files** (to simulate the camera and telephone).

b. Device Manager
This agent connects the wizard computer with the physical devices and with the user's Home Setup. When the Wizard clicks on the "kitchen light – on" button, this agent sends an X10 message to the kitchen light and also updates the user's Home Setup.

4.2.2 User Agents

a. Home Setup
This is a modified version of the actual system agent that displays the current setting of the house and its devices. The user may use the mouse or pen to click on the devices. When the device is clicked, it blinks (so that the wizard can see it with his remote screen) and sends a log message to the Log Manager. The Home Setup is linked to the Device Manager, so as to ensure its immediate update.

b. Telephone Simulator
This is the telephone terminal access icon on which the user can click. It blinks when clicked on and sends a message to the Log Manager as the rest of the Home Setup devices.

c. ASR Manager
Although the wizard will just listen to the user and will not take into account the recognizer output, the ASR will be activated in parallel (word+word grammar) to provide additional data. We have implemented several wrappers for different commercial ASRs.

d. TTS Manager
This agent synthesizes the wizard text messages and allows the Log Manager to keep record of the utterance.

e. Log Manager
This agent keeps record of all the user-wizard interactions during the experiment. It includes the information sent by the GUI Agents (Home Setup and Telephone Simulator) and the voice Agents (TTS and ASR Manager)

f. Video Client
This specific agent is used to simulate the security camera.

4.3 Inter-agent communication.

The platform must obviously be distributed and suitable for real-time applications. Although several options were available (Corba, Darpa's Communicator and Stanford's OAA), our final choice was OAA.

4.4 Inter-agent synchronization.

One of the goals of our experiment is to determine how the user interacts with the system when he uses a mixed mode (e.g. two inputs at the same time: Voice: “switch on this light”: Pen: Click on the kitchen lamp icon). In order to expand on Oviatt´s results on multimodal synchronization [1] [2] applied to our environment, a logging system with a precision of less than one second is needed. All our agents are implemented in C or JAVA, programming languages whose libraries allow millisecond precision, and the computers are configured running the NTP protocol.

4.5 Logging

This is the information saved in execution time during the experiment. The logging is therefore focused on low-level information, and especially on the time at which each utterance occurs.
The following table resumes the information logged:

Modality	Information Logged
GUI Input	Icon clicked Time
GUI Output	Message Time
Voice Input	General: Recognizer, Grammar, Language. Hypothesis level: Sentence, Score, Time Init, Time End. Word Level: Word, Score, Time Init, Time End.
Voice Output	Message Time

In order to save all this information, we have chosen the W3C recommendations from Emma (still a working draft), with very few modifications.

4.6 Annotation

This is information saved in post-execution time. Since our goal is mainly focused on the user’s behaviour at dialogue level, much of the important information will be annotated.

The NXT toolkit is our first choice. We will develop our own display and use it to process the information.

5 Conclusion

This experimental platform will allow us to conduct the necessary experiments with the appropriate accuracy and simulation efficiency to ensure the robustness of the final results. In addition to this, different languages can be used and results compared in order to find some potential language-based differences in modality integration in multimodal dialogue systems.

References

[1] Oviatt, S. L., Multimodal interactive maps: Designing for human performance, Human-Computer Interaction, 1997, 93-129 (special issue on "Multimodal interfaces").

[2] Oviatt, S. L., DeAngeli, A. & Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In Proceedings of Conference on Human Factors in Computing Systems: CHI '97.

[3] Hofs et ali. A multimodal interaction system for navigation. In Proceedings of Conference Diabrück 2003, 7th Workshop on the Semantics and Pragmatics of Dialogue: 2003