Simulating Multimodal Applications

C. Tadj†, H. Djenidi†, M. Haouani†, A. R. Cherif‡, N. Levy‡
†École de Technologie Supérieure - Electrical Engineering Department
1100, Notre-Dame Ouest, Montreal, Qc, H3C 1K3 Canada
{ctadj,hdjenidi,mhaouani}@ele.etsmtl.ca
‡PRiSM, Université de Versailles St.-Quentin
45, Avenue des Etats-Unis, 78035 Versailles Cedex, France
{rca, nlevy}@prism.uvsq.fr

ABSTRACT
A design methodology for multimodal-controlled application has been developed using Wizard-of-Oz simulations as the principal mechanism for evaluating and getting input for dialogue design. This methodology may enable multimodal application developers to support dialogues that are optimal with respect to naturalness, especially on a pragmatic level, given the technical restrictions.

Keywords: Multimedia Multimodal Systems, Wizard-of-Oz Simulations, Intelligent Agents, Human-Machine Applications.

1. INTRODUCTION

The term multimodality refers to the ability of a system to make use of several communication channels during user/system interactions.

In a Multimedia Multimodal System (MMS), information like speech, pen strokes and touches, eye gaze, manual gestures and body movements are produced from user input modes. These data are first acquired by the system, then they are analyzed, recognized and interpreted. Only the resulting interpretations are memorized and/or executed. This ability to interpret by combining parallel information inputs constitutes the major distinction between multimodal and multimedia systems.

Multimedia systems are able to obtain, stock and restore different forms of data (text, images, sounds, videos, etc.) in storage/presentation devices (hard drive, CD ROM, screen, speakers, etc.). Modality is an emerging concept combining the two concepts of media and sensory data. The phrase “sensory data” is used here in the context of the definition of perceptions: hearing, touch, sight, etc. [1]. The set of multimedia multimodal systems constitutes a new direction for computing, and provides several possible paradigms which include at least one recognition-based technology (speech, eye gaze, pen strokes and touches, etc.) and leads to applications which are more complex to manage than the conventional windows interfaces, like icons, menus and pointing devices.

There are two types of multimodality: input multimodality and output multimodality. The former concerns interactions initiated by the user, while the latter is employed by the system to return data and available information. The system lets the user combine multimodal inputs at his or her convenience, but decides which output modalities are better suited to the reply, depending on the contextual environment and task conditions.

The literature provides several classifications of modalities. The first type of taxonomy can be credited to Card et al. [2], who focuses on physical devices and equipment. The taxonomy of Foley et al. [8] also classifies devices and equipment, but in terms of their tasks rather than their physical attributes. Frohlich [10] includes input and output interfaces in his classification. Coutaz and Nigay have presented in [4] the care properties that characterize relations of assignment, equivalence, complementarity and redundancy between modalities.

For output multimodal presentations, some systems already have their preprogrammed responses. But
now, research is focusing on more intelligent interfaces which have the ability to dynamically choose the most suitable output modalities depending on the current interaction.

2. MMS DIALOG AND MULTI-AGENTS

The Timed CPN [11] offers a suitable pattern, to design constraints in multimodal dialog fusion. For modeling purpose, each input modality is assimilated to a thread where signal fragments flow. Multimodal inputs are parallel threads corresponding to a changing environment that describes different internal states of the system. Multi-agent systems are multi-threaded: each agent has a control on one or several threads. Intelligent agents observe the states of one or several threads for which it is designed. Then, the agents execute actions that modify the environment.

Petri net modeling is done with the toolkit Design CPN [12]. Its functionalities and its semantics give the possibilities to model concurrency, parallelism, distribution and hierarchy including time and stochastic sensitivity. A Petri net is a diagram flow of interconnected places (or locations) and transitions (or activities) managed by a set of rules. The rules determine when an activity can occur and specify how its occurrence changes the state of the places (by changing their colored marks). The set of colored marks in all places before an occurrence of the net is equivalent to an observation sequence of a multi-agent system. A transition can model an agent. The observation function of an agent is simply modeled by the conditions in each transition guard. The transition activity modifies data and thereby changes the states of the system. A previous work [6] shows how such timed colored Petri nets models the multimodal fusion in a generic way (by a multithreaded dynamic architecture where each modality is associated to a thread of symbolic events). When a fusion is performed, a fusion agent realizes two sorts of fusion processes (see Figure 1). The serial (or syntactic) one, at the signal information level, fuses information from the speech recognizer to make commands. The parallel (or semantic) one involves, for example, the recognized commands or isolated words with the position of the eye-gaze on the screen (as a mouse cursor) to generate a more complex command. The time proximity of two events is the main criterion used to generate a simple or a complex command.

![Figure 1](image-url) 

Figure 1. The fusion agent and the command agent: two intelligent agents of the MMS architecture.

3. METHODOLOGY DEVELOPMENT

MMS are becoming an increasingly attractive option for providing advanced user friendly and robust applications. Most of these MMS integrate speech as the main modality. Unfortunately, most of existing speech-controlled services are based around small vocabularies and isolated word recognition, but as continuous speech recognition technology matures, this will change. Hypothesis in our research study has been that, regardless of the quality of the media used (speech recognition, any other media), all such services may benefit from having dialogues derived from task analysis and Wizard-of-Oz simulation studies. In other words, even though natural syntax and semantics cannot be supported, natural pragmatics may still be of use.

4. WIZARD-OF-OZ SIMULATIONS

Several studies have been undertaken in the area of simulating speech-understanding systems [9], giving suggestions on how to set up such experiments. The WOZ simulating is a method that a user and a person called Wizard (who behaves as if he is a system) are communicating together, given a dialog procedure.
There are differences in utterances used when a user thinks that he is communicating with a human or with a machine. Instinctively, the user behaves differently (talk slowly in speech mode for example). Therefore human-human dialog should not be applied to human-computer dialog interfaces [3]. In the same spirit as most of these experiments, we have used the simulation set-up shown in Figure 2.

Figure 2. Principle of multimodal model, \((IM_i: \text{Input Media } i ; OM_i: \text{Output Media } i)\) and the hardware devices involved in the MMS. Simulation Set-up.

The dialogue design methodology involves at least two separate stages of simulation. The first stage is based on unrestricted, task-oriented, human-human dialogues. For the second round of simulations, a rudimentary dialogue model is used, describing how the target service should behave in different stages of the dialogue.

5. EXPERIMENTAL SETUP AND TASK

The experiments involved in this work use the multimodal system developed using the interactive graphical software environment, based on generic multi-agents architecture [7]. The input modalities of the multimodal software application are speech, eye-gaze and/or touch screen. Displays on monitor screen and voice synthesis are the output modalities. The application is dedicated to paralytics and allows them to navigate in the web and to use windows environment.

5.1. Methodological results

Simulations in the second stage are based on a rudimentary dialogue model. A support tool, the wizard’s device represents this basic organization of different parts of the dialogues as a set of panels each having several groups of messages. This provides the wizard guidance on what answers were appropriate in a given situation.

Wizard’s device defines the communicative space of the simulated system - what the user can do and say. Since the type of simulation we are investigating assumes that the service has limited capabilities of understanding, the dialogue model allows for handling only requests that could be mapped to corresponding functionality in the service in question. When it came to general speech understanding competence, this stage of the simulation assumed rapid speaker-independent continuous speech recognition with a grammatical coverage.

From speech output point of view, the options for making a wizard seem machine-like have consisted of using text-to-speech conversion on one hand, and voice distortion on the other [9]. However, state-of-the-art speech synthesis is generally perceived as less intelligible than human speech, and distorted human speech is by definition less easy to understand than normal speech. It turned out that digitized, undistorted, spoken messages were sufficient to give the impression that the subjects were communicating with a machine (9 out of 10 subjects in our study believed this). Note that messages were spoken in a friendly but formal way.

5.2 General observations

Some observations are made in connection with the principal study (a simulation of a speech-touch-screen-controlled multimodal application) that we believe generalize to other media-controlled services. Most importantly, we verified what has been reported elsewhere [9]. In subsequent interviews, subjects even explicitly expressed the need they had felt to imitate system language in order to compensate for the lack of a clear model of the competence of the dialogue counterpart. The
need for a feedback model reflecting the system's competence also proved important, i.e. making explicit what the system is unable to hear, what it does not understand (vocabulary unrelated to the application) and what it can't do (when the functionality of the application is a limitation).

6. CONCLUSIONS

This paper presents an interactive graphical software environment, based on generic multi-agents architecture. The input modalities of the multimodal software application are speech, eye-gaze and/or touch screen. Display on monitor screen and voice synthesis are the output modalities. The application is dedicated to paralytics and allows them to navigate in the web and to use windows environment.

The WOZ method has been is used for simulations, and the designers and developers need to implement the actual system based on the result of the simulations. Therefore dialogues used for the prototyping are not fully reflected in the final system.

We are currently working on task-oriented conversational agents which meet the tasks and scenarios gradually from example dialogues when we apply the learning technology to the WOZ, and the system absorbs and uses the example dialogues collected by the WOZ well.

This work will be improved by developing a WOZ for experimental evaluation of the user interface in the software environment. Our purpose is to make MMS a fully visual design environment and decrease the architectural model complexity and improve model maintainability, reusability, understandability and reliability.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada. We also wish to thank the Commission Permanente de Coopération Franco-Québécoise 2004-2005, who are in turn supported by Québec’s Ministry for International Relations and France’s Ministry for Foreign Affairs (General Consulate of France in Québec).

8. REFERENCES