UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

ADAPTATION OF MULTIMODAL OUTPUTS

David Filipe Ribeiro da Costa

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Sistemas de Informação

2011

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

ADAPTATION OF MULTIMODAL OUTPUTS

David Filipe Ribeiro da Costa

DISSERTAÇÃO

Projecto orientado pelo Prof. Doutor Carlos Alberto Pacheco dos Anjos Duarte

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Sistemas de Informação

2011

Acknowledgments

I would like to show my gratitude to my family for supporting me throughout all my studies at University. Specially I owe my deepest gratitude to my brother who since day one helped me to surpass the difficulties of this demanding course, without him I couldn't be here writing this work.

This thesis would not have been possible without my advisor Prof. Dr. Carlos Duarte who always supported my work and gave great advices and guidance during this project. I would like to thank you the opportunity to work in this ambitious European project.

It is a honour to me to work in LaSIGE with the group AbstInt. I would like to thank this great group for providing the support and equipment I needed to develop my work.

I am grateful to all my friends made in college specially Bruno Neto, Pedro Feiteira, Joana Neca, João Ludovico, Nádia Fernandes, Tiago Gonçalves and José Coelho. Thank you all for support and the good moments we had together. Finally I would like to thank my good old friends Nelson Dias, Ricardo Constantino, Bruno Ribeiro and Luísa Cabral.

To my mother

Resumo

Este documento centra-se em sistemas multimodais adaptativos mais especificamente nas suas técnicas de adaptação das saídas, ou seja, cisão de diferentes modalidades de saída de forma a permitir uma melhor adaptação ao utilizador.

O primeiro capítulo faz uma pequena introdução às interfaces multimodais e as suas vantagens, tais como ao possibilitarem o uso de modalidades alternativas, e oferecerem aos seus utilizadores opções de interacção naturais. Ao recorrer a modalidades como a voz ou gestos, é possível ter uma interacção mais próxima daquilo a que as pessoas estão habituadas na sua interacção diária com outras pessoas.

Este aspecto é ainda mais relevante quando o grupo de utilizadores alvo é composto por pessoas idosas, o que é o acontece no âmbito do projecto GUIDE, em que o trabalho relatado neste documento se insere. A motivação e os principais objectivos deste projecto estão descritos neste primeiro capítulo e passam por desenvolver uma framework para os programadores de software integrarem facilmente características de acessibilidade nas suas aplicações de TV.

O foco deste projecto é a televisão e as suas mais recentes capacidades de processamento (Set-top boxes). Estas plataformas têm o potencial para se tornarem nos dispositivos de media mais usados devido à sua fácil aceitação e especialmente quando se trata de utilizadores idosos que podem ter à sua disposição aplicações de conferência audiovisual, controlo remoto da casa entre outras aplicações que têm como base simplificar a sua vida quotidiana e afastar da solidão, um problema muito presente nesta faixa etária.

Os utilizadores podem assim empregar modalidades com que já estão familiarizados e optar por aquelas com que são mais eficazes. Utilizadores com limitações de audição podem optar por modalidades visuais, por exemplo. A adaptação envolve assim várias áreas do sistema humano como as capacidades físicas do utilizador, ou seja, a sua capacidade de movimentar os seus braços ou mãos, a sua percepção táctil, as limitações visuais tais como miopia, daltonismo ou visão em túnel, capacidades auditivas e também as cognitivas, ou seja, a capacidade de se concentrarem, perceberem o ambiente ao seu redor ou recordarem. As possíveis soluções face a estes problemas estão também descritas no documento.

Esta flexibilidade proporcionada pelas interfaces multimodais, não significa que estes sistemas não necessitem de operações de selecção e configuração, de natureza técnica,

que não é expectável que os utilizadores realizem devido à sua complexidade. De modo a conseguir realizar estas operações, o recurso a interfaces adaptativas é uma solução a considerar.

Ainda neste capítulo é descrito o papel que a Faculdade de Ciências da Universidade de Lisboa desempenha neste projecto e mais especificamente as minhas responsabilidades e os meus objectivos definidos para este projecto. Ao longo do desenvolvimento deste projecto surgiram várias ideias, estudos e desenvolvimentos que culminaram na escrita de alguns artigos e também aplicações que estão descritas na secção de contribuições. Na secção de planeamento é discutido o que estava inicialmente planeado e as alterações que surgiram.

Com este projecto pretende-se encontrar um mecanismo de adaptação que seja capaz de melhorar o desempenho da cisão multimodal por diferentes saídas. O mecanismo de adaptação de saídas multimodais é responsável por decidir qual a melhor estratégia para, primeiro, seleccionar as melhores modalidades para apresentar conteúdo (baseado no perfil do utilizador, as características do conteúdo e as modalidades disponíveis), segundo, distribuir o conteúdo pelas modalidades seleccionadas (usando estratégias de redundância e/ou complementaridade) e, terceiro, ajustar o conteúdo a cada modalidade. Para o estudo dessas mesmas estratégias a serem usadas foi realizado um trabalho de pesquisa a projectos relacionados com sistemas multimodais e consequentemente cisão multimodal (parte constituinte de uma arquitectura multimodal adaptativa). Descrito ao longo do segundo capítulo estão as arquitecturas usadas e técnicas de cisão e adaptação da informação apresentada.

No terceiro capítulo são apresentados estudos realizados aos utilizadores alvo deste projecto, com o objectivo de conhecer e entender como estes interagem com um sistema capaz de oferecer diferentes modos de interacção e de apresentar conteúdo. Padrões de comportamentos, características e preferências dos utilizadores foram resgistadas de modo a encontrar uma correlação e agrupá-las em diferentes perfis de utilizador. Para este efeito foi concebido uma aplicação multimodal que gera ecrãs a partir de um ficheiro XML de modo a facilmente se criar, modificar ou remover testes. Os utilizadores podiam interagir por gestos (apontando para o ecrâ), usando um controlo remoto ou por voz, podendo combinar estas modalidades diferentes. O conteúdo era apresentado através de elementos visuais (texto, botões, imagens e videos), áudio (sintetizadores de voz) e recorrendo a um avatar. Esta aplicação regista o sucesso ou não na realização dos testes como também o percurso de interaccao do utilizador em cada teste (a ordem em que os elementos foram selecionados e o tempo que demorou a realizar as tarefas). Os resultados e as conclusões retiradas deste estudo estão descritas no final do capítulo.

Depois de definidos os perfis de utilizador concluiu-se que é necessário que o sistema GUIDE consiga ligar novos utilizadores a um perfil. Com essa finalidade foi desenvolvida uma aplicação que serve de inicialização ao sistema. Essa ferramenta, descrita no capítulo

4, introduz as capacidades de interacção ao utilizador e de seguida apresenta diferentes tarefas de modo a avaliar as caracteristicas e preferências do utilizador. Ao concluir as tarefas, a ferramenta é capaz de atribuir um perfil ao utilizador que mais se adequa ao mesmo. Sendo as características do perfil genéricas, o perfil vai sendo moldado e actualizado conforme o utilizador vai interagindo com o sistema GUIDE.

O capítulo 5 começa por apresentar a arquitectura do sistema GUIDE e descreve todos os seus principais componentes. Neste capítulo é demonstrado como funciona o módulo de cisão multimodal começando por definir diferentes níveis de profundidade na adaptação das interfaces das aplicações e cujo nível é selecionado de acordo com as necessidades do utilizador. A cisão é responsavel então por decidir em que modalidades apresentar o conteúdo da apresentação. Depois de atribuídas as modalidades a usar, a informação é enviada aos respectivos dispositivos de saídas para gerar a apresentação. A geração da apresentação é coordenada e gerida pelo módulo de cisão que está em constante comunicação com os dispositivos de saída de modo a garantir uma apresentação coerente. No fim do capítulo é descrito um prototipo do modulo de cisão onde tenta na prática realizar todo o processamento definido nas secções anteriores.

Como forma de conclusão do documento são distinguidas as contribuições desta tese para o projecto bem como o trabalho futuro a realizar na continuação deste trabalho.

Palavras-chave: Sistemas Multimodais, Adaptação,Cisão Multimodal, Saídas

Abstract

This document main focus is on multimodal adaptive systems more specifically in its techniques for adjusting the outputs, i.e., split the information by different output modes to allow the best adaptation to the user. By using modalities such as voice or gestures, it is possible to have interaction closer to what people are used in their interaction with others. This is even more relevant when the target user group consists of elderly people, which is the case with the GUIDE project described in the document. This project aims to develop a framework for software developers to easily integrate accessibility features into their TV based applications. Users can thus use modalities that are more familiar and choose the ones that are most effective when interacting. Users with limited hearing can choose visual modes, for example. Adaptation involves so many areas of the human system as the physical capabilities of the user, i.e., its ability to move their arms or hands, their tactile sense, the visual limitations such as low vision, blindness or tunnel vision, hearing and cognitive capabilities, i.e., the ability to concentrate, remember or understand. Possible solutions that address these issues are also described in the document. This flexibility afforded by multimodal interfaces, does not mean that these systems do not require operations of selection and configuration of a technical nature, which is not expected that users perform due to its complexity. In order to accomplish these operations, the use of adaptive interfaces is a solution to consider. The aim of the work reported in this document is to find an adaptive mechanism that is capable of improving the performance of multimodal fission for different outputs. The mechanism of adaptation of multimodal outputs is responsible for deciding the best strategy to first select the best means to present content (based on user profile, the characteristics of content and modalities available) second, distribute the content by the modalities selected (using strategies for redundancy and / or complementarity) and third, adjust the contents of each modality. To perform the correct adaptation the system needs to know its users, thus user trials were carried out to understand their characteristics, behaviours and interaction patterns and to group different type of users into clusters. This document presents an application developed to assist in those trials. A prototype of an initialisation application to tutor users and match them with a user profile is also described on this document.

Keywords: Multimodal Systems, Multimodal Fission, Outputs, Adaptive

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In our everyday lives, we use multiple ways to communicate with each other using speech, gestures, expressions and vision. The modalities we use for natural conversations are not the same as the ones we use during human-computer interaction and the main reason is that application developers force users to adapt to the computer's form of functioning and not the other way. When interacting with a computer we normally use a keyboard for typing and a pointer device for pointing or clicking, therefore the interfaces are not focused in the use of multimodalities.

Over the last two decades a new form of interfaces arisen, called "multimodal user interfaces", prepared to recognize human language and behaviors. These interfaces integrate recognition technologies such as speech and gestures, relegating the keyboard and mouse to a second plan of input interaction. Of course this will bring some issues as we no longer have the simplicity of graphical user interfaces (GUI) that expect an input as an atomic form and an unequivocal order of events [1].

With multimodal interfaces, users are offered a set of more natural interaction options because it provides alternative modes of input and output other than the usual in human–computer interaction. Resorting to modalities like gesture and/or voice, it is possible to have an interaction closer to what users are used to do in a human to human normal interaction. Even though this sounds interesting for any age group, it turns out to be more compelling to focus on elderly people due to their disabilities or limitations inherent to advanced aging and their lack of experience with graphical user interfaces. Therefore these users can make use of modalities which they are accustomed to or more effective with. For instance a user with visual impairment can choose audio or haptic modalities or even a combination of various modalities to overcome their interaction problems.

The adaptability provided by this type of systems needs indeed some operations of configuration and selection that by any means are not made by the users due to their lack of technical knowledge or familiarity with the system. Adaptation can be very difficult and tedious for the user unless resorting to an adaptive interface solution.

This work is focused on finding that mechanism of adaptation that is able to improve and refine the performance of multimodal fission of different outputs. This mechanism is responsible for the decision making of the best strategy, firstly to bring out the content using the best available modalities suitable to the user's profile and the content features, secondly to distribute that content through the selected modalities (using strategies of redundancy and/or complementarity), finally to adjust that content for each modality chosen [19].

The work was developed under the scope of an European project named Gentle User Interfaces for Elderly people (GUIDE) which is described further.

1.1 Elderly people use of machines

As people grow older, health problems may appear from ageing, which may be physical but also mental. These impairments are an obstacle to a proper interaction with a computer because it is very difficult, for instance, to someone with upper limbs movement limitations to use a mouse and keyboard to give an input.

Using a unimodal system is also restricting the presented information into a single modality excluding persons whom suffer from an impairment of that sensory needed to interact (a blind person cannot see graphical information and a deaf person cannot hear sounds).

The social exclusion and e-exclusion made by these systems is a serious problem as it restricts the actions, the information received and the interaction by users with sensorial impairments that otherwise wouldn't have any problem interacting with them.

Multimodality tries to resolve this issue as it offers the possibility of presenting the same information in different ways (sound, visual, haptic), compensating some sensorial impairments. Presenting information using different modalities isn't new but they are used in most cases to distribute different content in different modes [41].

1.2 GUIDE - Gentle User Interfaces for Elderly People

As stated before this work is being developed in the scope of the European project GUIDE[1] (”Gentle user interfaces for elderly people”) which has the goal of developing a framework for developers to efficiently integrate accessibility features into their applications.

GUIDE puts a dedicated focus on the emerging Hybrid TV platforms and services (connected TVs, Set-Top Boxes, etc.), including application platforms as Hybrid Broadcast Broadband TV (HbbTV) as well as proprietary middleware solutions of TV manufacturers. These platforms have the potential to become the main media terminals in the users' homes, due to their convenience and wide acceptance. Especially for users of the elderly society applications such as home automation, audio-visual communication or

[1] GUIDE EU Project - `http://www.guide-project.eu/`

continuing education can help to simplify their daily life, stay connected in their social network and enhance their understanding of the world.

Ageing and accessibility are two subjects that are highly correlated in several contexts, like interacting with electronic devices such as computers, nomadic devices or set-top boxes. Approximately 50% of the elderly suffers of some kind of (typically mild) disability such as visual, auditory or cognitive impairments, which poses several problems and challenges to social interaction. For such end-users, accessible Information and Communication Technologies (ICT) can make much more of a difference in living quality than for other citizens: it enables or simplifies participation and inclusion in their surrounding private and professional communities.

When adapted in the right way, recent advances in human-computer interfaces such as visual gestures, multi-touch as well as speech, or haptics could help to let disabled or elderly users interact with ICT applications in a more intuitive and supportive manner.

Despite these positive trends, implementation of accessible interfaces is still expensive and risky for developers of ICT applications, as the integration of accessibility requires time, knowledge and money. Among others, they have to cope with user-specific needs and limitations (including lack of ICT proficiency) as well as with technological challenges of innovative UI approaches, which require special experience and effort. Therefore, today many ICT application implementations simply neglect special needs and lock out a large portion of their potential users.

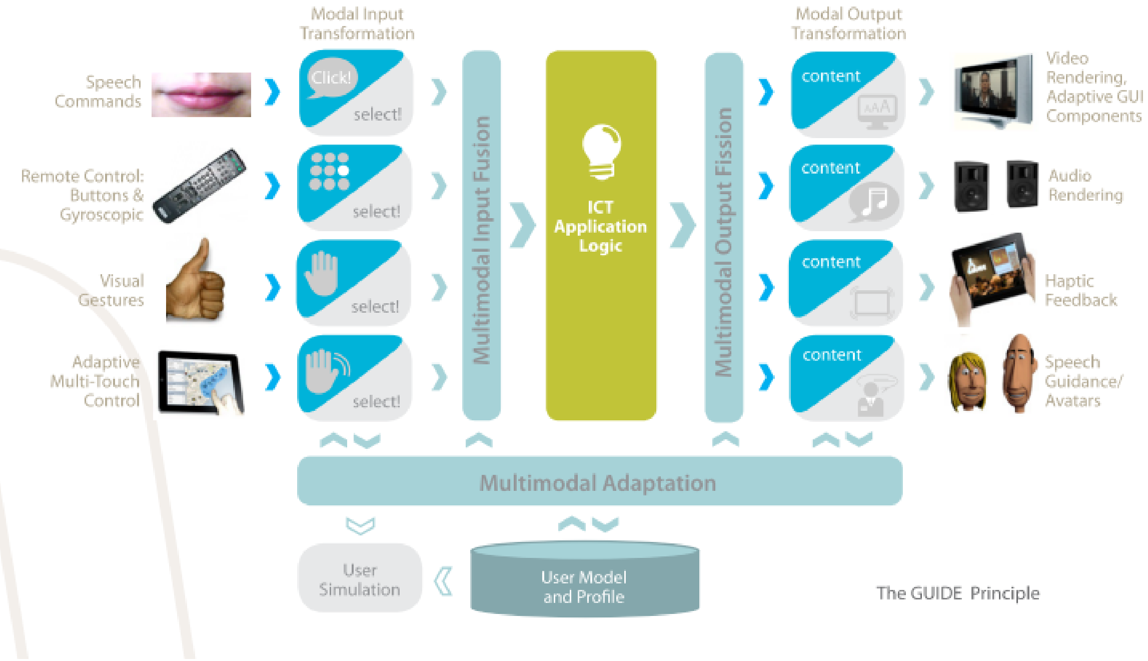


Figure 1.1: GUIDE's conceptual view

Although there are Application Programming Interfaces (API) that provide accessibility features they don't offer intelligent adaptation capable of addressing users with different kinds of impairments. Also current accessible ICT applications only provide user interface adaptation resorting to manual configuration.

GUIDE is being conceived to fundamentally change this situation by providing an open framework for developing accessible user interfaces. This framework would be able to integrate state-of-the-art accessibility features into developers' applications without the need to change their functionality.

GUIDE is expected to work with different sensorial channels adapted for best user's usability, so as figure 1.1 shows, visual, audio and haptic modes will be the main output form of feedback. Visual content like video and the graphical user's interface will be suited to user's profile, adapting the content to his or her visual, auditive or other type of disabilities. Audio from sounds or speech will be available through speakers. Haptic feedback will take advantage of the user's tactile capability to inform him of some notification, normally acting together with other modalities. To guide the user through the system's interface, there will be an avatar, developed by CCG (Centro de Computação Gráfica) that will be responsible to maintain the user on the right direction.

Aims

As mentioned before GUIDE's main target population are the elderly people though the benefits of such an adaptable system can be applied to every range of population. There are two major concerns when aiming for this target population, their resistance to learn new technologies and how to interact with them, and the capacity of those technologies to be used by users with a wide range of impairments. So one of GUIDE's goals is to implement the use of more human like ways of communication employing voice and gestures as a mean of interaction with the system, not requiring a long learning curve which is one of the reasons that keeps away new users. Additionally the use of TV as the main media channel will ease the acceptance of this system in their homes.

As aforementioned, GUIDE aims to offer its users additional interaction modalities, like speech, pointing and gesturing, besides the traditional remote control. GUIDE must adapt the operation of these modalities to benefit the user experience. GUIDE must be able to render the application content in the most appropriate modality for a given user and therefore provide alternative ways of displaying that content. GUIDE must enable the possibility to adapt its features and settings automatically demanding no skills or technological knowledge of users to configure the system, providing thus the best interaction and performance possible.

In order to offer an adapted multimodal interaction experience, GUIDE must collect information about the user such as his capabilities, characteristics and preferences. To do so, GUIDE will have to gather that information explicitly by requesting it directly from the user, and implicitly by analysing the user's behaviour and actions when interacting with the system. This user related data must be maintained and updated by GUIDE in order to be consulted by all components responsible to apply adaptation mechanisms.

On the ICT applications' developers point of view, it is understandable that they won't change their programming fashion when building TV based applications and neither have the knowledge to develop adaptive multimodal applications. In order to address these issues GUIDE won't require any changes from developers but instead introduce a middle layer between the application and the user, responsible to provide multimodal interaction and adaptive features by adjusting the user interface and providing alternative output modalities for the generated content. Thus developers will still implement their applications considering a normal TV environment, also the TV consumers will be presented with a variety of input and output modalities and with adapted content presentation.

Another goal of this project is to be able not only to deal with TV based applications (HTML) but any kind of application languages (e.g., Java, C#, etc.). GUIDE must somehow have a language independent representation of the user interface and an API to perform the communication between many different application environments.

Additionally another concern about developers and companies relates with the possibility of GUIDE being able to adjust their applications' UI changing their original visual properties into something more favourable to the user. So GUIDE must be capable of offering different levels of adaptation.

1.2.1 Role of FCUL in GUIDE

The GUIDE consortium is composed by several Universities and companies, each one with their speciality functions. FCUL is a technical member of the group and therefore it plays a major part in the development of this European project. FCUL has the responsibility to design an adaptive multimodal architecture system that follows the project requirements and implement some of the core components of that system. Those components to be implemented are the Multimodal Fusion module, Dialogue Manager, and Multimodal Fission module. In order to accomplish the goal of developing such a complex system centred on users' specificities and preferences, user trials must be realized in order to study and collect requirements. FCUL assists in the designing and preparation of these studies which involves the creation of an application to conduct and help to record the studies. Multimodal systems provide many different ways of interaction. FCUL is also integrating various input devices into the system (e.g. WiiMote, Kinect and remote control).

1.2.2 My role in GUIDE

My contribution to the GUIDE project is mainly to develop from the design to the implementation a multimodal fission component to be integrated in the GUIDE framework.

Using GUIDE a user centred design approach, trials with end-users are certain and therefore it's my incumbency to implement a multimodal prototype to perform those tests. The results of the trials will be analysed to take conclusions on how the multimodal fission can best adapt the content displayed. In order to adapt those contents the system must know the user, thus an application to assess the user's characteristics must be developed.

1.3 Goals

Understand the fundamentals of Multimodal Systems To conceive a multimodal fission component, it is my first goal to do research on how multimodal systems work and what types of adaptation there are.

Specialize in Adaptive Multimodal Fission Learn techniques and methods of multimodal fission adaptation from existing systems and develop those techniques in order to meet GUIDE requirements.

Contribute to User Studies Multimodal prototypes must be built to test with end-users in order to analyse and get some findings on users' reactions with such systems. From those findings, results a better understanding on creating a first prototype of a multimodal fission module capable to adapt content to each user.

Contribute to the process of User profiling Without registering characteristics of users, GUIDE can't adapt successfully to a specific user. Thus a tool capable to gather that information is critical to guarantee that fission module performs adaptation in a accurate way.

Integration Developing a prototype of the fission module being capable to be integrated with the other GUIDE's core components is another goal.

1.4 Contributions

From this project and thesis resulted various contributions such as articles, studies and prototypes.

Multimodal systems provide users natural ways of interaction modes, which also means different users may interact differently, combining different modalities. Studing user's behaviours and interaction patterns could increase performance and effectiveness of multimodal interaction by adapting those systems to user's abilities. To characterize user abilities we proposed an approach that resorts to a Wizard of Oz based system capable of simulating input and output actions without using any recognition technology. More information about this approach is described in ”Support for inferring user abilities for multimodal applications” [18] paper.

Based on that approach and the need to realize user studies to acknowledge user's reactions, behaviours and patterns, an multimodal application was developed. The User Trials Application is able to generate interfaces from a XML description of the UI. Input interaction was guaranteed by input pointers (e.g. Motion capture devices, Kinect and Wiimote), remote control and Speech (simulated by Wizard of Oz approach). Ouput content was presented trough text, buttons, images and videos.

From the User studies findings, user's characteristics were grouped into different clusters and user profiles defined. When a new user is presented to GUIDE it needs to gather information from him and match to one of the defined profiles. The User Initialisation Application has the aim to tutor the user on how to interact with GUIDE and test his abilities to identify his profile.

”Adapting Multimodal Fission to Users' Abilities” [14] was a result from researching information on multimodal systems. This paper describes the state of the art regarding multimodal output fission. It shows some systems capable of using the best channels to present the content taking into account user's impairments and environmental context. Also introduces GUIDE's goals to make this adaptation easier to TV based applications' developers.

After the completion of the first user trials a paper was written with the analyses and conclusions on users' reactions, preferences and abilities when interacting with different modalities (gestures, speech commands, etc.) and receiving feedback from visual elements, speech synthesizers and Avatar. User trials findings are found in ”Eliciting Interaction Requirements for Adaptive Multimodal TV based Applications” [17].

A small study was also conducted in order to find a standard UI description language independent of the application's language. Abstract and concrete representations of user interfaces were taken into account. This representation is used in GUIDE to adapt and update elements' properties, gather content to present in other modalities and infer speech vocabulary. The results of this study are presented in chapter 5.

The final contribution is the prototype of a multimodal fission module capable to decide, based on user profile and environmental context, what modalities to use, what content to split in each modality and coordinate the generated presentation.

1.5 Planning

The first scheduled task planned for this project was to create some multimodal applications prototypes from an existing framework (MMX[2]). This framework was developed by a university colleague and our goal was to verify its re-usability on GUIDE project. This process took two months and it's main value was the acknowledgement of the communication system based on publish / subscription of events implemented on MMX framework.

[2] Multimodal framework MMX - `http://frameworkmmx.sourceforge.net/`

Simultaneously and in the next four months research on multimodal systems and particularly on multimodal fission was made to learn the basics on this area such as the architecture, techniques and guidelines.

In the end of last year it was planned to conceive a first prototype of the multimodal fission component but unfortunately it was not achieved and postponed for a later stage of the project. The reasons were due to GUIDE member's indecisions over the hardware and software requirements of the set-top-box and consequently in the programming language to use. Also in these meetings was stated the importance of building an application for user studies in order to collect user's behaviours and preferences when interacting with multimodal systems. This new task would take several months to complete and update while receiving feedback from the first trials.

Planned for the beginning of this year was the selection of adaptation strategies resulting from the prototype development. But due to the aforementioned reasons it was also delayed. Instead it was decided in a new GUIDE meeting the necessity to have a tool capable of detecting new users and learn their abilities and impairments for adaptation purposes. The responsibility to make this application fell to FCUL.

GUIDE's original architecture suffered from constant modifications resulting from new meetings. These changes were made mostly due to set-top box limitations and the need to make GUIDE more scalable regarding language environments. Thus the delayed tasks only started on the beginning of Summer. From the user trials findings and research done, the development of the first prototype begun but it suffered a setback as I was diagnosed a cancer and had to be operated, postponing one more month those tasks. Completing the writing of this document originally set to end of July was delayed and finished only in September due to my radiotherapy treatments.

1.6 Document Structure

The remainder of this document is structured as follows:

Related Work This chapter describes some research done in assistive technology and in multimodal areas. The advantages and disadvantages of multimodal systems over the standard ones are presented. The highlight of this section is the techniques and architectures used to divide and adapt content by existing adaptive multimodal systems (output fission).

User Trials Application In this chapter it is described the studies done on users to identify interaction patterns, behaviours and preferences on multimodal output modalities. To do these studies (two different trials were done) an application was developed and its description is written in this same chapter.

User Initialisation Application This chapter describes an application designed to tutor and evaluate the user. It has the main goal of matching his abilities to a profile according to user's characteristics to perform the best adaptation possible.

Multimodal Fission This chapter describes the GUIDE's architecture and its components, output requirements and the description of the fission module processing. A prototype of that component is also described and analysed.

Conclusion This chapter ends the document with the main conclusions draw from this project.

Chapter 2

Related Work

This chapter presents some of the research literature related to Assistive technologies and Multimodal Interfaces, and more specifically work related to the multimodal fission of output modalities.

2.1 General Definitions

When we talk about multimodal interfaces there are a few definitions that need to be learnt for a proper understanding of the theme.

Modes are the sensorial system of a human with which he perceives the world. A **modality** is defined by the structure of the information that is perceived by the user (text, sound, vibration, etc.) and **medium** is a channel or the mean used to express the modality, i.e., the peripheral devices such as monitor or TV screen, sound columns and so on. All these three components are dependent of each other's [38].

"**Multimodal interfaces** process more than one combined input mode such as speech, manual gestures and head and body movements in a coordinated manner with multimedia output"[31] which can be **adaptive**, enabling a more natural and effective interaction whichever mode or combination of modes are best suitable to a given situation, context and according to user's preferences and abilities (user profile).

Interacting with a multimodal system means to provide input using one or multiple modalities such as speech, key pressing with keyboard or mouse, gestures, etc. and receiving information and feedback from the system by a modality or combination of modalities such as visual, audio or other output modality.

There are two types of inputs: active and passive ones. **Active inputs** are issued deliberately by the user, e.g., speech, text, while **passive inputs** are inputs that are monitored by the system without requiring explicit command from the user, e.g., facial expression or manual gesturing.

With the capacity of parallel or sequential input modalities, these systems can combine various modalities in the recognition process, this is called **multimodal fusion**. These combinations can be made at several different levels in the recognition process. To present the information there's also a concept named **multimodal fission** which divides or selects the output channels to distribute that information throughout the available outputs and usually according to the user profile and context.

2.2 Assistive technologies and techniques

Assistive technologies are technologies that have the goal to provide accessibility for disabled people. By making usable what is not, these add-ons provide specialized input and/or output capabilities for a person with impairments. Users with disabilities do not always use assistive technologies, but, nevertheless can benefit from small design changes [41].

Physical disabilities affect the ability to move, manipulate objects, and interact with the physical world. When interacting with software, no matter the cause (ageing, congenital conditions, accidents, etc.) users with physical impairments have hardware add-ons that bypass these problems like head tracking (e.g. SmartNAV[1]), joystick (e.g. 3M Ergonomic Mouse[2]) and other alternative pointing devices (e.g. BIGtrack[3]). There's also Screen Keyboards which provide the keys and functions of a physical keyboard and are normally used in conjunction with alternative pointing devices. To enhance the performance when typing there are predictive dictionaries that predict and offer a list of words. In the output point of view and of course GUIDE's view, with ageing the tactile perception capability diminishes due to degeneration of the peripheral nervous system, which means if haptic modality is to be used it must be adapted (strength level, duration, etc.) to user's abilities.

Low vision and blindness (color blindness, tunnel vision, etc.) limits the field of view as users need large fonts or have to magnify the screen aided by hardware or software, reducing the amount of information visible at one time. Users with these impairments usually lose context as their lack of field view causes the loss of events that may occur out of their field of view. Also color blind users are not capable of interpreting information that depends on colors. Blind people, who don't use visual displays at all, use Braille displays or listen to a speech synthesizer, which give them feedback about their navigation. GUIDE only features a TV screen so for this extreme visual impairment only speech and simple haptic components will help blind people to navigate through the system.

For the less extreme types of impairment it is recommended the usage of redundant information (audio, symbols, text) and customizable graphic interfaces that help users to be notified about new information and to better interpret that information. For blind users, software developers should have more attention in naming the objects for a better

[1] SmartNav hands-free cursor control - http://www.naturalpoint.com/smartnav/

[2] Ergonomic Mouse - http://www.3mergonomicmouse.com/

[3] BIGtrack `http://www.infogrip.com/product_view.asp?RecordNumber=98`

description when that object is being read by screen readers (e.g. WebAnywhere [4]) and also provide a description for all non-text objects like images, graphics or icons.

Hearing disability may not be a big issue when interacting with software applications as they rely heavily on visual presentation although this situation may change with the continually integration of computer, telephones and video. However with GUIDE's project, sound plays a more relevant role as this project is centred on TV. When possible, sound should be redundant with visual information and not used alone as we cannot assume that sounds will be heard. Examples of assistive technology are Telecommunications Device for the Deaf (TDD) that offers the capability to communicate over telephone using text terminals, Closed Captioning that provides text translation of spoken material on video media and ShowSounds[4] that provide visual translation of sound information. GUIDE's visual translation can be done via subtitles or visually transmitted sign patterns (sign language) using the avatar. Ju-Hwan Le et al study [28] also shows the importance of haptic feedback coupled with sound.

Cognitive disability is a general concept, having to do with impairments in the way humans concentrate upon, remember, plan, perceive and understand.GUIDE focus on elderly people and one disease that affect this age group (change of getting it increases with age) is Alzheimer. This disease gradually or quickly deteriorates memory and has effect on other cognitive areas too. After some cognition related tests, GUIDE is capable of assuring that users with cognitive disabilities will not be lost in their navigation as cues (text, speech, images, etc.) will be frequently presented and the avatars can also play an important role to guide the user.

Although these technologies seem to be very interesting to deliver accessibility features, GUIDE aims to use more common technologies to present the content to any user despite of his impairments. Thus we opted to develop a module capable to divide and combine that content into different modalities (visual, auditory, haptic, etc.) with its role described in the next sections.

2.3 Multimodal Systems

Multimodal systems are defined in [19] as "computer systems endowed with multimodal capabilities for human-computer interaction and able to interpret information from various sensory and communication channels." These systems offer users a set of modalities to allow them to interact with machines and "are expected to be easier to learn and use, and are preferred by users for many applications" [31].

As stated before, these systems use a different approach when compared with the classical WIMP (Window, Icon, Menu and Pointer) interfaces because multimodal interfaces make use of more ways of communication and more natural to humans. Systems like

[4] Accessibility: A guide for educators - http://www.microsoft.com/education/teachers/guides/accessibility.aspx

these tend to help users by supporting and adapting the interaction to their perceptual and communicative capabilities bringing interaction advantages to both human and machine.

As opposed to unimodal systems where the interaction had to be adapted to a given application, the Multimodal User Interfaces (MUI) are flexible and offer users the capability to change between different modes for expressing different types of information. The advantages are obvious, users with different skills, age, native languages and physical or cognitive impairments are able to interact more effectively with computer systems that are able to adapt to different situations and to a context in constant evolution [30].

The flexibility in these systems allows them to adapt not only to users but also to the environment (context awareness). For example, in a birthday party where there are a lot of people and noise in the living room, it is wise not to force speech recognition as a form of information input to the system. As for the system it will be better to raise the volume or show the information needed by text and notify the user using haptic feedback. Other example is the usage of speech to warn or present information in an eyes-busy situation.

Systems that combine outputs evolved since the early nineties where text and graphics were combined (e.g. COMET [22]). More recent systems combine speech, haptic, graphics, text, 2D/3D animations or avatars (e.g. SmartKon [35]; MIAMM [36]). Although most applications use few output modalities and consequently straightforward fission techniques, dealing with the above-mentioned combination of outputs can make the presentations more complex and difficult to coordinate and make them always coherent.

There are also studies which conclude that unimodal feedback alone is not enough unless combined with other modalities, i.e., visual interface combined with haptic or auditory feedback have shown to improve performance of users. Although the performance differs from computer experienced users and less experienced ones, auditory-haptic feedback was the most effective [20].

Oviatt and Cohen [32] states that the combination of multiple modalities on the input and output side of a multimodal system makes it more robust, reducing errors in the communication.

There are not only advantages when using multimodal interfaces, because adding several modes and mixing them increases the complexity of the application as each modality has its own interface and distinct human behavior.To know how to make them work together, their properties and the amount of information that is transmitted in each modality must be learnt.

Cognitive studies on users have shown that there are different integration patterns when interacting with multimodal systems and have identified two types of users (sequential vs. simultaneous). Those who have sequential patterns interact not overlapping inputs and with an interval between them. Those who have simultaneous patterns produce multimodal signals that overlap. As a consequence by learning from users mistakes and adapting thresholds for different user groups it is possible to increase integration accuracy and speed [19, 33].

Multimodal capability brings new issues as a MUI receives different input streams of modalities. These inputs must be first interpreted by probabilistic recognizers (HMM, GMM, SOM, etc.) [19], then, as mentioned above, a simple temporal event line isn't expected as we must consider synchronized parallel processing of these continuous streams in order to perceive the intention of multimodal combination or separated use of modes. So time sensitivity plays an important role to determine the order of commands and if they are processed in parallel or in sequence. To ensure the synchronization and efficiency of computation, normally a distributed architecture is used to deal with these issues.

In the next section these and other architectural issues and solutions will be covered.

2.3.1 Architecure

This section describes what components should be present in a multimodal architecture. The first multimodal systems, like the 80's "Put that there" [5] system, worked on a basis of a control structure where multimodal integration took place during the process of parsing spoken language. But while this is doable for processing a multimodal point and speak system it isn't for more complex inputs like the use of gestures or facial expressions. In such complex systems the control structure used in the first multimodal systems is not capable of interpreting complex unimodal or combined multimodal inputs.

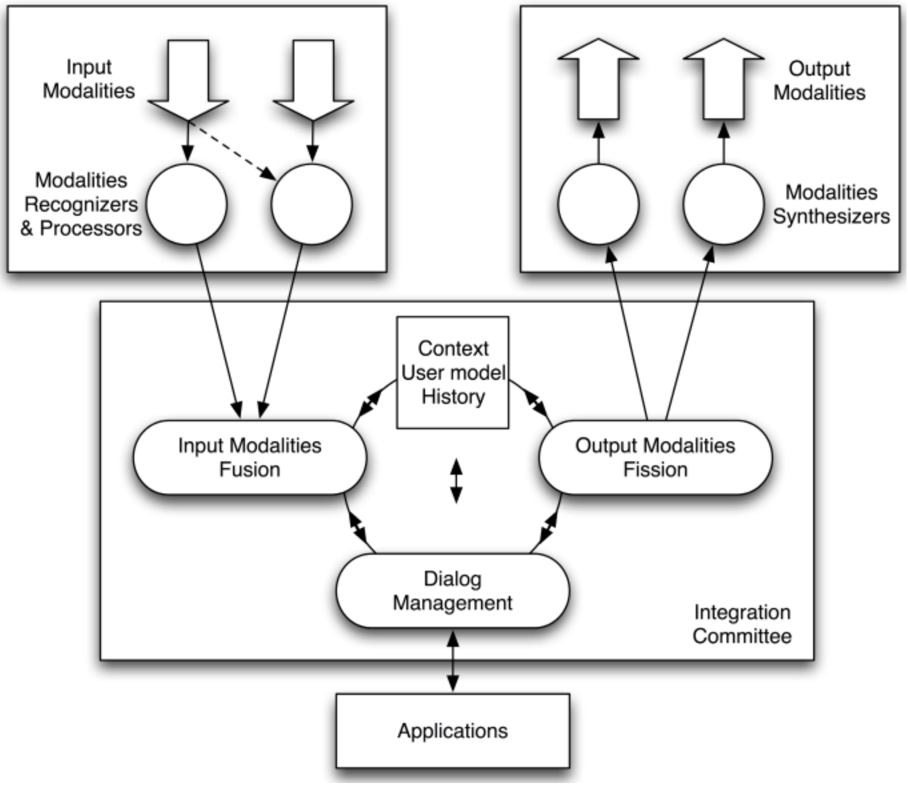


Figure 2.1: The architecture of a multimodal system and its generic components [19]

As [19] describes, the generic components for handling multimodal integration (integration committee) are a fusion engine (combination of modalities), fission module (divide information through active outputs), a dialogue manager and a context manager.

Figure 2.1 illustrates the components and the processing flow between these components. The various input modalities are all first perceived by their respective recognizers which each of them sending to the fusion module the results of their processing. This module is responsible for interpreting the information given by the recognizers and making its decision through fusion mechanisms. The computed fusion result is communicated to the Dialogue management that identifies the dialogue state and the action to perform and communicates it to an application and/or the fission module. This module returns the information to the user using the most adequate and available modality or combination of modalities according to the user profile and context. These are responsibility of the Context manager, which must be aware of the user profile and environmental context changes.

Multimodal fusion

As stated before multimodal fusion is the combination of various input modalities and that is an important difference between unimodal and multimodal interfaces. Extracting a meaning from a set of input modalities is the fundamental reason of this engine's existence. There are three main architecture types for multimodal systems:
Early fusion – this type fuses input signals at early stages or at feature level. This architecture is good for input signals that are highly dependent and closely synchronized such as speech and lip movements. This type of fusion is computationally complex and has a high training cost [40, 5].
Examples of this architecture can be found in Pavlovic [34] and Bregler, et al [7].

Late fusion – Also known as semantic level fusion, it is in contrast used for loosely coupled input signals. These signals provide different but complementary data that has been preprocessed. Late fusion systems use independent recognizers to train each input mode (unimodal data) which makes it easier to collect and scale up the number of inputs [43].

However, when using this architecture, it is assumed that each input mode is independent of the others modes and each mode is equally reliable, which is an issue that must be considered. A system using this architecture is Cohen et al. Quickset [13].

Hybrid fusion – This approach combines statistical feature fusion methods with unification-based methods to process inputs from different modalities. The hybrid technique is capable of achieving a more robust performance than the above mentioned architectures and a proof is that the system presented in [46] achieved 95% correct recognition performance.

Dialogue Manager

Dialogue management is one of the most important features of a multimodal system architecture (figure 2.2) as it is responsible for coordinating the dialogue between human and machine. According to [44] Dialogue management has to comply with some tasks regarding the update of the dialogue context on the basis of interpreted communication. It also must provide context-dependent expectations for interpretation of observed signals as communicative behavior. Dialogue management must also communicate with task / domain processing to coordinate dialogue and non-dialogue behavior and reasoning. Finally it decides what content to express based on interpreted events and when to express that content.

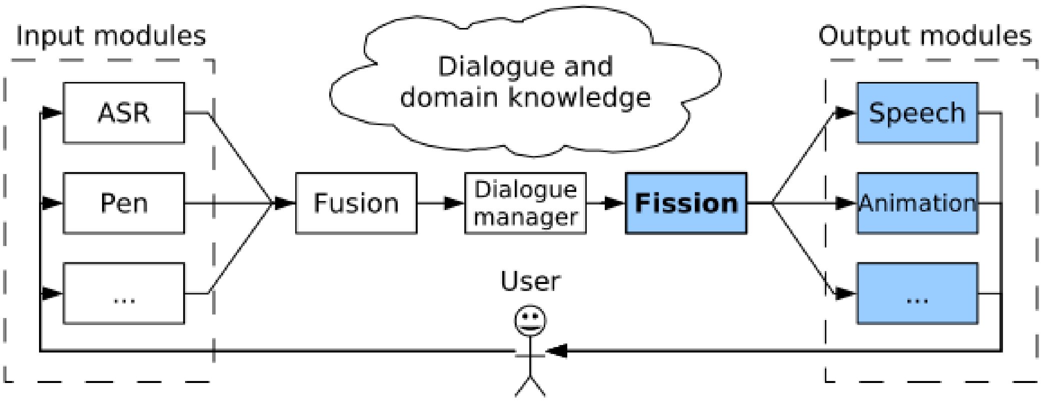


Figure 2.2: High-level architecture of a multimodal dialogue system [23]

P.Cohen [12], WeiqunXu et al.[47], M.McTear [29],R. Catizone et al. [42], have all classified and categorized strategies and approaches for dialogue management. According to Bui [8] there are four categories :

- **Finite state-based and frame-based approaches** – this model is the simplest model used to develop a dialogue management. It uses a state machine structure to represent the system's utterances. Although this approach lacks flexibility (frame-based models which are an extension of state-based models overcome this disadvantage using a slot-filling strategy in which a number of predefined information sources are to be gathered), naturalness, and applicability to other domains, it is most suitable for well-structured tasks.

- **Information state-based and probabilistic approaches** – These approaches try to describe human-machine dialogue making use of information states, divided in four main components [44]– a description of informational components such as participants, linguistic and intentional structure, beliefs, intentions, etc., formal representations of that components, a set of dialogue moves that initiate the modification to an updated information state and an update strategy for deciding which rule(s) to apply at a given point from a set of applicable ones.

- **Plan-based approaches** – These approaches are based on the plan-based theories of communicative action and dialogue which claim that the speaker's speech act is part of a plan and that it is the listener's job to identify and respond suitably to this plan [8].
- **Collaborative agent-based approaches** – these approaches are based on viewing dialogues as a collaborative process between intelligent agents that work together with the goal of obtaining a mutual understanding of the dialogue.

Context Manager

The context manager has the responsibility to keep track of the location, context and user profile maintaining a continued communication with the fusion module, fission module and dialogue manager to update them with any changes that may influence the understanding, interpretation and generation of communicative behavior. H. Bunt [9] grouped the different information dimensions into five categories:

- **Linguistic context** – Surrounding linguistic material, "raw" as well as analyzed.
- **Semantic context** – State of the underlying task; facts in the task domain.
- **Cognitive context** – Participants' states of processing and models of each other's states.
- **Physical and perceptual context** – Availability of communicative and perceptual channels; partners' presence and attention.
- **Social context** – Communicative rights, obligations and constraints of each participant. The context manager should have a set of data with a log of the recorded dialogue steps of decisions and interpretations made, a representation of the information to be gathered in the dialogue, general background information that supports any commonsense reasoning required by the system, a set of specific information, user's model information about his age, gender, preferences, user goals, beliefs, intentions and physical and mental states.

2.4 Output Fission

A multimodal system should be able to flexibly generate various presentations for the same information content in order to meet the individual user's requirements, environmental context, type of task and hardware limitations. Adapting the system to combine all this time changing elements is a delicate task (e.g. SmartKom[**?**]). The fission module and fusion engine are crucial to making possible the usage of multimodal applications for all users, as it takes advantage of multimodalities to overcome sensory impairments that users may have. When considering this in the scope of GUIDE's vision and elderly people, its target user group, that is the issue number one to be treated.

In other words this module is responsible for choosing the output to be presented to the user and how that output is channeled and coordinated throughout the different available output channels (based on the user's perceptual abilities and preferences). To do this according to the context and user profiles, the fission engine follows these three tasks that will be described further: Message construction; modality selection; output coordination.

Based on the What-Which-How-Then (WWHT) conceptual model of Cyril Rousseau et al. [38], who created this model to offer the capability to make adaptive and context aware presentations, the fission module will be described along the following three sections. The authors define three main components as being the means of communication (physical and logical) between human and machine: Mode, Modality and Medium.

It is also important to refer that there are primary and secondary relations between components. Primary relations are for example in haptic systems the "vibration" created by the system and the user's tactile mode, but as a side effect you can hear the vibration, making a secondary relation between the audio mode and vibration modality.

The WWHT model is based on four basic concepts and they will be described in detail in sections 2.3.1 , 2.3.2 and 2.3.3:

- **What** – What information to present
- **Which** – Which modality(ies) to choose to present that information
- **How** – How to present that information using that modality(ies)
- **Then** – How to make the presentation change

2.4.1 Message Construction

The presentation content to be included must be selected and structured, i.e., it is necessary to decompose the semantic information issued from the dialogue manager into elementary data to be presented to the user.

As it's shown on figure 2.3 the information is divided into n basic elements.This phase is called by the authors as "Semantic Fission" (What).

There are two main approaches for content selection and structuring that can be employed - schema-based or plan-based [16]. However, in some systems, selecting and

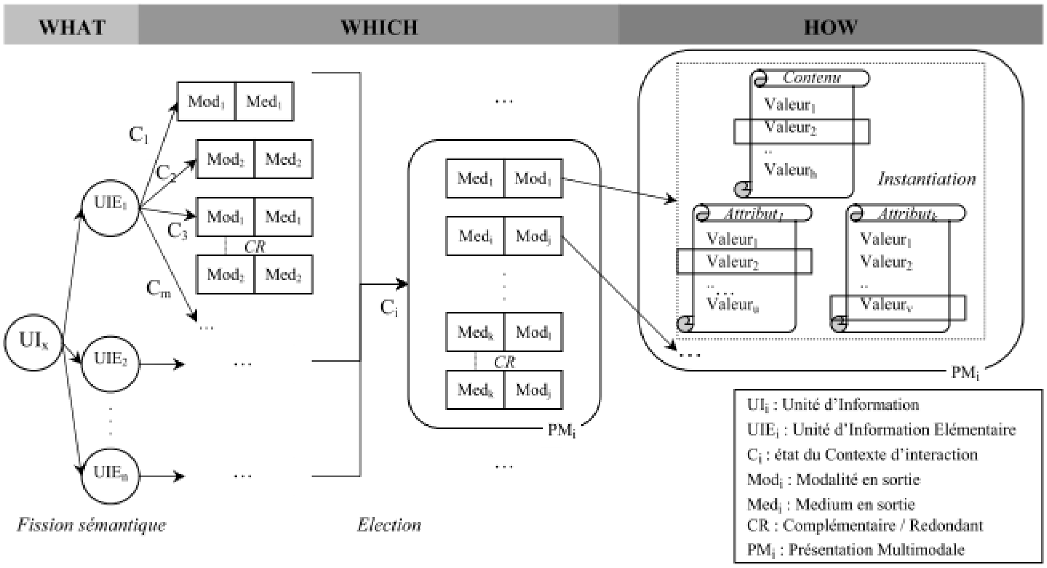


Figure 2.3: WWHT conceptual model for context aware and adaptive interaction [38]

structuring the content is done before the fission module process begins. An example of that is MAGPIE [24].

In [16] the author states the Schema-based approach "encodes a standard pattern of discourse by means of rhetorical predicates that reflect the function each utterance plays in text. By associating each rhetorical predicate with an access function for an underlying knowledge base, these schemas can be used to guide both the selection of content and its organization into a coherent text to achieve a given communicative goal". Schema-based systems like COMET [22] or PostGraphe [21] determine from a set of existing schemas the best suitable for the user's intentions assigning a weight for each intention using heuristics. From that step results a list of schemas ordered by efficiency. Next stage is the generation of elementary information. If a listed schema fails to generate the data then the next best efficient schema is tested. This is where it fails against plan-based approaches, because it is impossible to extend or modify a part of the schema.

Plan-based approaches use a goal-driven, top-down hierarchy planning mechanism which receives communicative goals and a set of generation parameters, such as user's profile, presentation objective, resource limitations and so on. Systems that use this approach then select parts of a knowledge base and transform them into a presentation structure. The structure's root node is the communicative goal, i.e., a somewhat complex presentation goal (e.g. describing a process like the TV programming) and its' leafs are elementary information to present (e.g. text, graphics, animation, etc.) [26]. WIP [45] is a presentation generation system that follows this approach receiving as input a presentation goal and then it tries to find a presentation strategy which matches the given goal. A refinement-style plan is then generated in the form of a directed acyclic graph (DAG). Figure 2.4 shows the presentation planner which generates the DAG where its leafs are specifications for elementary acts of presentation which are then sent to the respective module that is capable of handling them.

The advantage of this approach over schema-based approach is that it is able to represent the effects of each section of the presentation. Also, mode information can be easily incorporated and propagated during the content selection process. Plan-based approaches facilitate the coordination between mode information and content selection process as mode selection can run simultaneously with content selection and not only after.

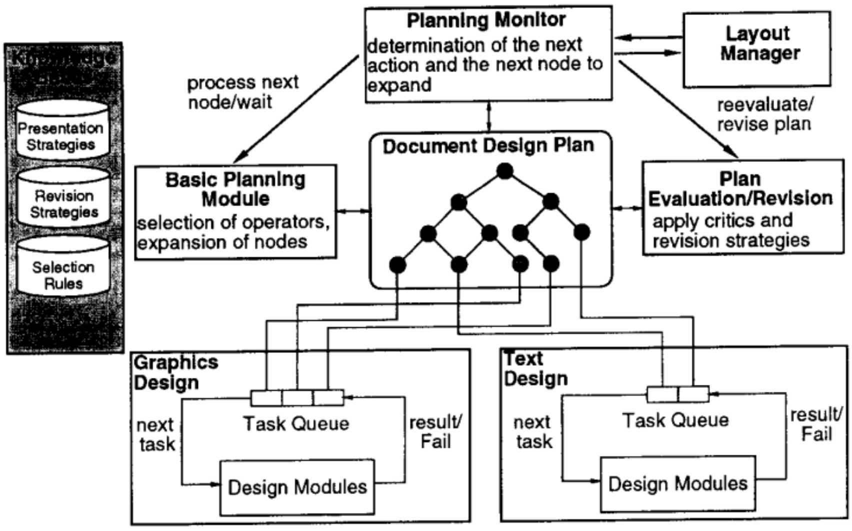


Figure 2.4: Presentation planner of WIP [45]

2.4.2 Modality Selection

After the message construction, the presentation must be allocated, i.e., each elementary data is allocated to a multimodal presentation adapted to the interaction context as presented on figure 2.3 (“Election phase” or Which). This selection process follows a behavioral model that specifies the components (modes, modalities and medium) to be used.

The available modalities should be structured according to the type of information they can handle or the perceptual task they permit, the characteristics of the information to present, the user’s profile (abilities, skills, impairments and so on) and the resource limitations. Taking this into consideration is necessary for optimal modality selection. For the compliance of this goal there are some approaches:

Composition

The system tries to combine selected primitives or operators, using predefined composition operators.The first criterion of modality selection is how efficiently and accurately each modality is likely to be perceived by the user. Depending on the modality(ies) chosen it will decide how the presentation is presented. The second step is to choose and combine a complete set of modalities that follows the specified message structure [10, 21].

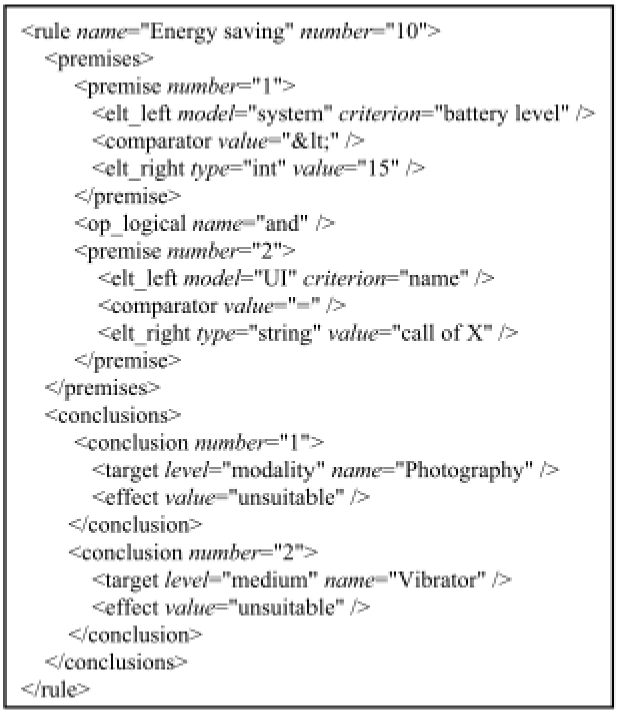
Rules

A set of election rules allocate the components of the presentation among the modalities using simple instructions (if . . . then . . .).

The premise of a contextual rule describes a state of context interaction environment (e.g. Noise level superior to 100 dB) and the contextual rule's conclusion is based on the weight of each premise.

Other type of rules is criterion-referenced ones. These rules allow a selection of rules based on global criterion (language, age, abilities, etc.) [3, 22, 38].

Figure 2.5 is an example of an election rule of ELOQUENCE [39] where the premises test the battery levels when an incoming call is triggered. If the battery is low, if the premises are true the conclusion is to decrease the vibration levels and restrict the use of photography modality to save more battery.



```
<rule name="Energy saving" number="10">
  <premises>
    <premise number="1">
      <elt_left model="system" criterion="battery level" />
      <comparator value="&lt;" />
      <elt_right type="int" value="15" />
    </premise>
    <op_logical name="and" />
    <premise number="2">
      <elt_left model="UI" criterion="name" />
      <comparator value="=" />
      <elt_right type="string" value="call of X" />
    </premise>
  </premises>
  <conclusions>
    <conclusion number="1">
      <target level="modality" name="Photography" />
      <effect value="unsuitable" />
    </conclusion>
    <conclusion number="2">
      <target level="medium" name="Vibrator" />
      <effect value="unsuitable" />
    </conclusion>
  </conclusions>
</rule>
```

Figure 2.5: Example of an election rule [39]

Coutaz et al [15] define some rules which allow the allocation of presentations with multiple pairs of modality-medium under redundancy and/or complementarity criterions. These rules follow four properties: Equivalence, Assignment, Redundancy and Complementarity.

Equivalence expresses the availability of choice between two or more modalities but does not impose any of temporal constraint on them (e.g. to show the same information message we can use text or speech synthesizer). Assignment expresses the absence of choice, which means there's no other modality choice or it is defined to use one and only modality for that specific case. Redundancy and Complementarity consider the use of combined multiple modalities under temporal constraints. Redundant express the equivalence between two or more modalities (same expressive power) and are used within the same temporal window (repetitive behavior) without increasing its expressive power. Redundancy includes sequential and parallel temporal relations. Parallelism puts restrictions on the types of modalities that can be used simultaneously as a human mode cannot be activated in parallel.

Complementarity is used when one modality isn't enough to reach the goal of the presentation and therefore more modalities are combined to reach the intended goal. Examples of systems that use CARE properties are MATIS [15], MEMO [6] or Rousseau's platform [37].

Agents

Competitive and cooperative agents plan the presentations. MAGPIE [24] system implements a set of agents that communicate with each other in order to reach a presentation goal. This system enables the dynamic creation of modality-specific agents needed to select and integrate basic components of the data presentation.

2.4.3 Output Coordination

Once the presentation is allocated, it needs to be instantiated, which consists in getting the lexical-syntactic content and the attributes of the modalities (How). First a concrete content of the presentation is chosen and then attributes such as modality attributes, spatial and temporal parameters, etc., are fixed.

For a coherent and synchronized result of the presentation, all used output channels should be coordinated with each other's. The consistency of the presentation must be verified as structural incoherencies (some modalities are indeed chosen to express multiple basic elements in one single presentation but that isn't always possible) and instantiation incoherencies (problems in the defined modality attributes) may occur. Output coordination abides by the following aspects:

- **Physical layout** – When using more than one visually-presented modality, the individual components of the presentation must be defined [22, 24].

- **Temporal Coordination** – When using dynamic modalities like voice synthesizers, videos or sounds, these components must be coordinated in order to achieve the presentation's goal. Because the order and duration of actions are different, the dynamic modalities used need to be synchronized and coherent. SMIL, HTML+TIME and HTIMEL are some of the languages built to ensure spatial and temporal synchronization of multimedia presentation where there is a structured temporal composition (sequential, parallel, exclusive) with time-based operations like timing (begin, end, duration) or repeating (repeat count, repeat duration) control.

- **Referring expressions** – Some systems will produce multimodal and cross-modal (interaction between two or more sensory modalities) referring expressions. This means making references using multimodalities or referring to another part of the presentation which need some coordination work [22, 2, 27].

Coordination and consistency are also necessary through the presentations as user's and environmental context may change along the evolution of the presentations output. This is important to not get outdated content, and if so invalidating it and get an updated version of the presentation (Then).

Chapter 3

User Trials Application

This chapter describes the user trials conducted in order to achieve a better knowledge about the users and how they interact and also their preferences on visual parameters. The chapter describes two user trials accomplished by implementing an user evaluation application for that effort. The following sections will describe the goals of the trials and the user trials application itself, the architecture of the application and of course the analyses and results of these trials on the adaptive multimodal output focus.

3.1 Introduction

Being the elderly people the main target for this project, GUIDE has to define its developing strategy following a user centred methodology to offer the best adaptive experience for this specific age range. This methodology aims to meet the users' requirements, behaviours and specificities studying and analysing their interaction with a multimodal system in the most likely end user environment. To meet this design approach a multimodal application (as GUIDE will be) had to be developed in order to extract useful data about users' interaction requirements, how they interact and combine different modalities, and most important, perceive how their different abilities and impairments can affect their perception with such systems. Therefore this application must have the capacity to provide different modes of input interaction and also multiple ways of giving feedback.

To use non technological materials (low fidelity prototypes) was out of the question as we need to get the most authentic reactions from the users as possible when using a multimodal system and the advantages inherited of this technology. Thus instead we have implemented a hybrid system with some fully working functionalities and other more complex functionalities were developed resorting to a Wizard of Oz approach.

The user trials consisted in set of scripted tests, where the users are introduced to a multimodal system and are able to experiment every device and every individual or combined modality available for navigating and selecting items in different menus, as well as giving opinions about their preferences on visual configurations of the UI (e.g.

text sizes, background colors,color contrasts and several more interface and interaction aspects) and what modalities are more appreciated to present the content to the user (e.g. visual and audio, only visual, only audio, etc.).

These trials had the goal to collect information about the end-users interaction and find common patterns, behaviours, skills (e.g. visual, audio, motor and cognitive impairments) and preferences in order to group and identify different clusters (each cluster has the characteristics that define the type of user) and to understand how target users interact with a system offering multimodal capabilities. The clusters will help to conceive a user model for GUIDE while the interaction understanding will assist in the development of guidelines for TV based multimodal applications.

To gather that data the User Trials Application has the aim of providing the users all possibilities of a more natural interaction by experimenting different devices and different ways of interaction with a TV based system and record those interactions for further analysis.

3.2 Architecture

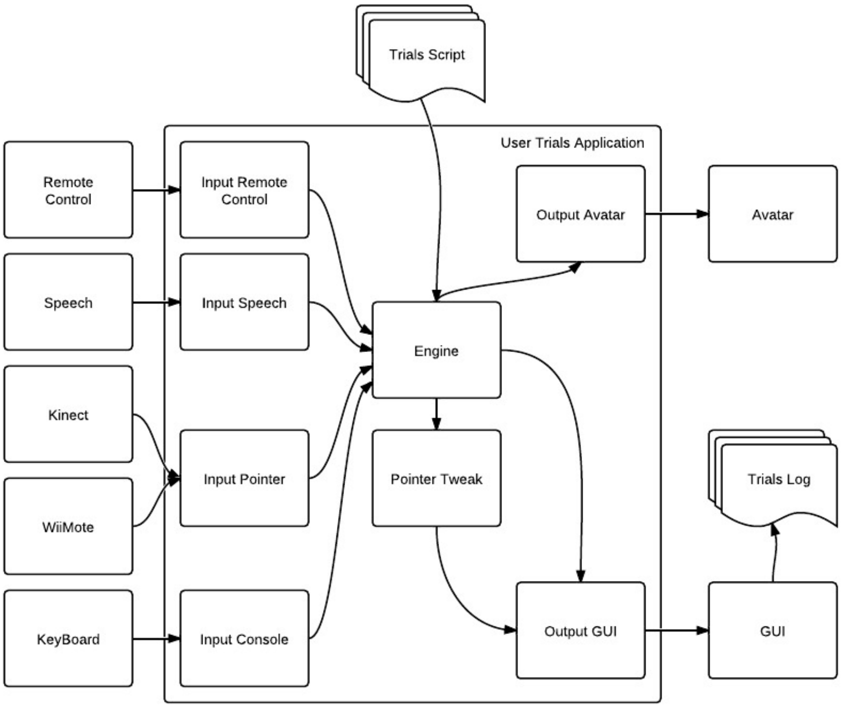


Figure 3.1: Architecture of the User Trials application

The figure 3.1 shows the architecture of the User Trials Application and all its components. The input devices that were available and implemented are many and they all work differently so input interfaces (e.g. Speech, Pointers, Wiimote, etc.) were developed to treat each device's data and communicate with the system. The same happens with the outputs: The GUI component, receives a structured data with the elements to be visually rendered, i.e., the user interface representation as well as the new coordinates of the pointer after adaptation is made by the Pointer Tweaker component. The Avatar receives a message setting up the configuration (e.g. volume, zoom level, gender, etc.) and the lines to be synthesized. All this information is controlled by the main component of the system, the Engine. This core element has the same function as a Dialogue Manager which sets up the states described in the XML based script files and manages the communication between all components.

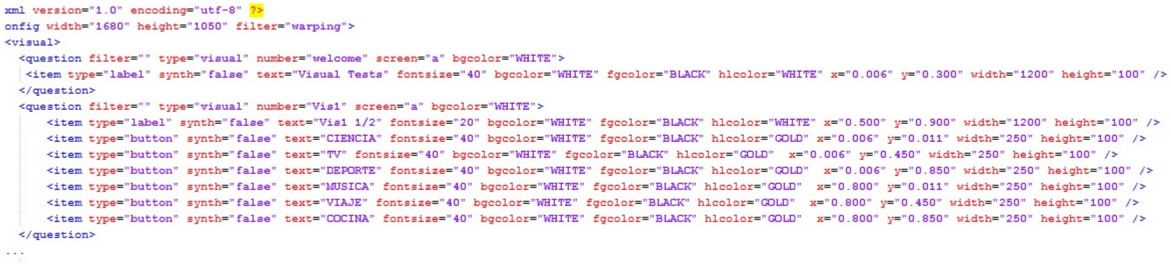
The communication between components is guaranteed by a publisher / subscriber based system. Input components publish their events using specific XML messages and the Engine, who subscribes to all input events, receives these messages and treats them accordingly. Additionally, the Engine publishes events for output modules to render the content. This approach permits adding more input devices in a easy way if they obey the type and format of the messages to publish.

The Input Console component is responsible for treating the Wizard of Oz input commands (e.g. select specific test to run, simulate button selections, evaluate user actions, save log, etc.).

All the actions (list of buttons navigated, selected, etc.), time used to complete each script and other relevant data are written in a log. The GUI component is responsible for it as it as the knowledge of all elements and cursor positions.

As mentioned above, to render the output, the system loads a script containing all the tests of the trials. These tests are a set of visual elements, audio and avatar settings. This approach and its advantages are described in the next section.

3.2.1 XML interface approach



```
xml version="1.0" encoding="utf-8" ?>
onfig width="1680" height="1050" filter="warping">
<visual>
  <question filter="" type="visual" number="welcome" screen="a" bgcolor="WHITE">
   <item type="label" synth="false" text="Visual Tests" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="WHITE" x="0.006" y="0.300" width="1200" height="100" />
  </question>
  <question filter="" type="visual" number="Vis1" screen="a" bgcolor="WHITE">
      <item type="label" synth="false" text="Vis1 1/2" fontsize="20" bgcolor="WHITE" fgcolor="BLACK" hlcolor="WHITE" x="0.500" y="0.900" width="1200" height="100" />
      <item type="button" synth="false" text="CIENCIA" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD" x="0.006" y="0.011" width="250" height="100" />
      <item type="button" synth="false" text="TV" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD"  x="0.006" y="0.450" width="250" height="100" />
      <item type="button" synth="false" text="DEPORTE" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD"  x="0.006" y="0.850" width="250" height="100" />
      <item type="button" synth="false" text="MUSICA" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD"  x="0.800" y="0.011" width="250" height="100" />
      <item type="button" synth="false" text="VIAJE" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD"  x="0.800" y="0.450" width="250" height="100" />
      <item type="button" synth="false" text="COCINA" fontsize="40" bgcolor="WHITE" fgcolor="BLACK" hlcolor="GOLD"  x="0.800" y="0.850" width="250" height="100" />
  </question>
...
```

Figure 3.2: XML example describing test screens

We decided to implement in the User Trials Application a highly flexibly scenario creation as it would provide user trials supervisors the ability to make last changes, experiments or creating new trials as they please without having to program anything. Thus we have chosen to take an XML interface approach where the user of this application can easily build its tests. This XML files are then read by the application to render the output. Figure 3.2 shows a example of a Script. A Question is equivalent to a single test or screen. Items are visual interactive elements (e.g. labels, buttons and images) which have their properties such as the text, text size, background color, location in the screen, etc. Buttons and Labels have the particularity to be read by a speech synthesiser. A tag named Avatar is used to call it in order to read a designated text.

This approach used in the implementation of the application offers the capability not only to easily build user trials for study purposes but also to simulate a future multimodal application in order to collect and analyse users reactions and interaction patterns, useful to enhance the software design in early stages of development.

3.2.2 Wizard-of-oz approach

The Wizard-of-Oz approach is a method to simulate the behaviour of a theoretical intelligent and fully functional system. This is done with the help of an experimenter (the ”wizard”) that intercepts all communications between the users and the system and performs the actions that were expected. This method has proven to be very useful to simulate for instance the voice recognizer and let the users feel free to use any words without restrictions. To perform this control over the system, the ”wizard” has a simple console with menu based options. This console offered the ”wizard” the possibility, among others, to load script files, move to the next test, classify the user performance in a task and simulate a target selection.

3.2.3 Input and Outputs

Device	**Input and Output on the User Test Application**
Remote Control	Button selection (input)
Wii Remote + Wii Sensor bar	Pointing, gesture and button selection (input)
Kinnect / Led Camera Sensor	Pointing, gesture and button selection (input)
Avatar Engine	Audio and visual output
Speech Synthesis	Audio output
Simulated Speech Recognition	Audio input (Console input)
Tablet (Apple’s Ipad)	Touch screen input and visual and haptic output

Table 3.1: I/O modalities available

As stated before the User Trials Application provides the use of several input and output modalities and devices. Most of these modalities were completely implemented, but for those who weren’t feasible to implement until the start of the first trials it was used the Wizard of Oz Approach (e.g. Speech recognition). The Input and Output devices that were available are shown in table 3.1 although some of this devices weren't implemented from the beginning (e.g. the Kinect device replaced the led camera sensor in later versions of the application, remote control was also added).

3.3 First User Trials

The script included tasks to exercise perceptual, motor and cognitive skills. These tests aimed at exploring how users perceive and prefer visual elements (font size, font and background colors, object placement, etc.) and audio stimulus (volume). Also, a set of tasks to exercise the user's cognitive abilities were included. Furthermore, and even though in some tasks users were free to interact in any way they wished, including combining modalities, we included tasks for users to explicitly use more than one modality, in order to observe what integration patterns may appear. The following sections explain these trials setup, tests and the analysis over the findings.

3.3.1 Setup

The first trials were conducted over a period of one week. Nine elderly people participated in these trials: five men and four women. Their average age was 74.6 years old, with the youngest being 66 years old and the oldest being 86 years old. Sessions lasted about 40 minutes, in which participants familiarized themselves with the available means of interaction and executed the required tasks.

The system was installed in a PC with two screens: one regular monitor for the "wizard" and one TV for the user. The devices available for interaction in these first trials were the WiiMote and a Led Camera Sensor, which detects user's hand motion. Also, speech commands were accept as input.

3.3.2 Tests implemented

As mentioned previously the trial's script took in consideration visual, audio and motor requirements. Moreover, in some tasks participants were free to elect how they wished to interact with the User Trials Application while in other tasks they were specifically requested to interact with a specific modality or combination of modalities. In most of the script's questions participants had to perform the same task with different modalities or with a different rendering parameter and after performing the task they had to report their favourite configuration for that task. In some other questions participants were just asked to select between one of differently rendered presentations without having to explicitly perform a task. Figure 3.3 shows a user performing one of the tasks, using the motion recognizer to point at the screen.

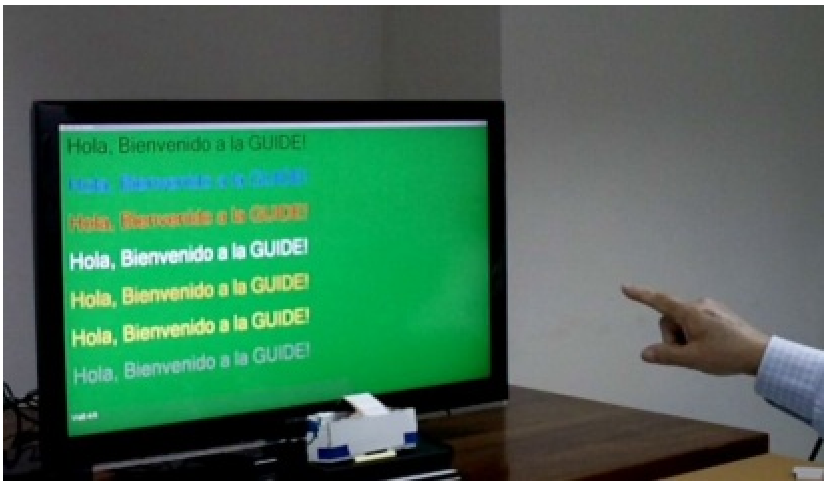

Hola, Bienvenido a la GUIDE!
Hola, Bienvenido a la GUIDE!
Hola, Bienvenido a la GUIDE!
Hola, Bienvenido a la GUIDE!
Hola, Bienvenido a la GUIDE!

Figure 3.3: User performing a task

The trials begun with a simple explanation task, where participants were asked to select one of four options presented on screen as four buttons. They has to do it through finger pointing, using Wii remote to point, speech and with a combination of speech and one of the pointing options. Afterwards, different presentations were explored: target placement (near the center or near edges of the screen), target size, target spacing, target color, text size and text color. Only for the color presentations participants simply had to select their preferred option (see example in figure 3.3). In all other, they had to perform a task, such as selecting one of the targets or reading aloud the text. The trial then proceeded with an audio rendering task where participants were asked to repeat the text that was rendered to them through a TTS (text to speech) with different volume levels. These were followed by motor tests, where users had to perform gestures (not simply pointing) both without surface support or using a tablet emulating surface. Afterwards, users had to perform a selection task in all combinations of input (options presented visually, aurally or both combined) modalities. Finally, they had to assess what would be their preferred modality to be alerted to an event when watching TV and when browsing photos on the TV. Options included on screen text, TTS, the avatar or combinations of the three. The test ended with a comparison between the avatar and a video of a person presenting the same content, to assess the avatar's ability to generate empathy with future system users.

Table 3.2 summarizes the type of tests, the tasks and the modalities users could use.

3.3.3 Analysis

This section describes the analysis of results in what concerns adaptation of multimodal outputs obtained in this trial based on the participants expressed opinions and presentation preferences but also on observations, in situ and of trials' recordings. The full analyses of the trials results can be read in [17].

Type of tests	Tasks	Devices/Modalities
Modalities and devices experimentation	Answering to questions related with preferences of interaction, experimentation of each device and modality. Menu items selection and navigation.	Input: WiiMote, Motion sensor. Output: visual menus and Avatar.
Visual capabilities and preferences	Answering to questions related with interface visual configuration (font size and color,etc). Menu item selection and navigation.	Input: one or more devices choosen by the user. Output: visual menus.
Audio capabilties and preferences	Answering to questions related with audio preferences (audio volume).	Input: Speech. Output: Audio.
Cognitive capabilities	Localization of different items on the screen (cognitive scientific tests), measuring time of response.	Input: Speech, WiiMote, Motion Sensor. Output: Visual menus and pictures
Motor capabilities and preferences	Performing gestures. Menu item selection and navigation. Interacting with the tablet. Answering to questions related with motor preferences and pointing mechanisms.	Input: WiiMote, Motion Sensor and tablet. Output: Visual menus.
Avatar preferences	Interacting with avatar. answering to questions related with avatar preferences.	Input: One or more devices chosen by the user. Output:Visual menus and Avatar.
Multimodal preferences	Menu item selection and navigation. Simulation of application contexts of use. Answering to questions related with multimodal interaction and references.	Input:One or more devices chosen by the user. Output: Visual and audio menus and avatar.

Table 3.2: Tests, tasks and modalities

Visual Perception

Two main aspects regarding visual presentation were studied: target (e.g. buttons) and text. Targets were analysed according to these three features: target placement, target size, target content.

In what concerns target placement, being the options near the edges of screen or in the center, most users (6 participants) preferred targets near the center of the screen instead the other option. Also, the majority of the participants, six of them, preferred the version with greater separation between buttons than the version with targets closer. And finally, seven of them preferred targets with larger size.

In what concerns text presentation, involving size and color properties. Six text sizes were presented to participants, being the largest an 100 pixel font (this meant that approximately five words would fill half of the TV screen), and the smallest a 12 pixel font. Intermediate sizes were 80, 64, 40 and 24 pixels. From all participants, only one preferred the largest font. Five participants chose the second largest and three participants the third largest font size. No participant preferred smaller fonts (40, 24 and 12).

	White	**Black**	**Blue**	**Green**
White	-	7	8	2
Black	7	-		5
Blue			-	2
Red	1	1		
Green				-
Orange				
Yellow				
Gray	1	1	1	

Table 3.3: Participants' preferences regarding text color (in rows) for different background colors (columns)

Text colors were assessed against different background colors. Text colors considered were white, black, blue, green, orange, yellow and gray. Background colors used were white, black, blue and green. Participants mostly opted for high contrast combinations. Table 3.3 shows the participants' expressed preferences in this test. Note that rows represent font color and columns the background color, and the values represent the number of participants that selected that particular combination.

Audio perception

The audio tasks performed evaluated participants ability to perceive audio messages in different volumes. This test employed five volume values, starting with the loudest setting and then decreasing each time by half the previous volume. Afterwords, the procedure was repeated but this time from the lowest to the highest volume.

Three participants preferred the loudest volume, with all but one participants preferring one of the three higher volumes. However, some participants noted that highest volume was too high in their opinion. One interesting finding, reported by some participants, was that their comfortable audio level was different when increasing and decreasing the volume. For instance, one participant reported she could understand the spoken message only in the first two volumes when the volume was decreasing, but she could understand the loudest three volumes when the volume was increasing. In addition to their specific preferred volume, participants reported to be comfortable with the three louder volumes.

Motor skills

Regarding their preferences and abilities when asked to perform pointing tasks with targets in the four corners of the TV screen it was not possible to identify a clear tendency in the collected results. One participant found that it was easier to point at the top right, two preferred the bottom right, one preferred the top edge, one preferred the bottom edge and four did not express any preference.

Specific Modality and Multimodal patterns

Regarding the presentation, participants unanimously expressed a preference for the system to employ redundant visual and audio presentation (all 9 participants selected this option). Observation of some participants' behaviour showed that when the system presented options both visually and aurally, they did no wait for all options to be presented, answering as soon as they perceive the right answer.

Different combinations of modalities to alert the user were also considered in the context of two different scenarios: watching TV and browsing pictures on the TV screen. For the TV watching scenario preferences were variable. Two participants preferred the alert to be rendered only using speech synthesis. Other two preferred to be alerted by avatar and text message, four preferred text and audio and one the avatar only. In the photo browsing scenario a similar variability was found. Two participants preferred alerts through speech synthesis only, two preferred the avatar and text, one preferred text and audio and two preferred the avatar only.

Avatar preferences

The use of an avatar in the system was also assessed. The participants' reaction to the avatar was not as positive as expected. There are some justifications for this, given the avatar employed had a too small size and was not properly configured in what concerns emotion expression. However, it was possible to gain some knowledge about the use of an avatar in the context of the proposed system. Some participants expressed a request for the avatar to look better, less cartoonish, in order to make them feel better about it. An important observation was that four out of the nine participants responded to the avatar's greeting message as if they had been greeted by a person. This is an indicative sign that avatar's can be used to promote a bonding between the users and the system.

3.3.4 Discussion

It must be stressed that these were the first trials, and thus have had such a small number of participants. However, these trials were useful to understand the impact of user's visual perception, audio perception, motor skills, preferences and interaction patterns can have when developing such complex project as an adaptive multimodal system. These findings support the need for adaptation mechanisms in order to provide adequate interaction to a user population with such diversity of abilities.

One example of how adaptation could be explored became evident as a result of the observations conducted in this trial: participants used their pointing hand based on where the target is on screen. This information can be explored to decide on presentation details. For instance, if the system knows the user has impairments affecting his left hand, it can display the selectable targets on the right side of the screen, offering the user a more comfortable interaction experience.

Additionally, these first trials were useful to understand what could be improved in this tool so the following trials could collect even more interesting data. The following section describes the improvements made and the new findings.

3.4 Second User Trials

As aforementioned, the User Trials application suffered some improvements taking in account the observations made by the GUIDE experimenters during the trial and the conclusions made from the first trial.

One of the enhancements implemented were the inclusion of adaptation algorithms for cursor movement. Those algorithms perform different strategies to ease the cursor movement by the users. The first algorithm is called ”Exponential Averaging” and what it does is averaging the current cursor position values (x and y) with the previous ones, giving user the feel that cursor movement is smoother. Another algorithm is named ”Damping” and it intends to oppose the fast movement of the cursor, so if the user performs a sudden movement it will suffer some attrition reducing the speed. Finally, the last algorithm is ”Warping” and it uses the information about the elements displayed on the screen to ”attract” the cursor to them. For instance, if the user moves the mouse near a button, it will slowly bent towards the button's center.

Another improvement made was the enhancement of the ”wizard”'s console. As the experimenter wasn't always a technical expert the console needed to be more simple and fool-proof. Additionally, it was added more options such as saving the results on the log whenever the ”wizard” wanted and not necessarily in the end of the trial and the possibility to select tests by name.

Related with tests, experimenters reported the need to randomize tests screens in order to not influence the participants' answers. Also, some bugs were corrected and new tests made.

New input devices were added in these trials: Microsoft Kinect and a Remote Control. And a new item was made available to display in tests, the video.

The log was also improved allowing to keep track of the test results in a more accurate way: how much time the participant take to perform a task, track of which elements were selected and what time and classification score.

The main goals didn't change much and aim to identify viable interaction methods (gestures, speech commands) of novel and traditional UI paradigms for the different impairments in the target groups using a multimodal system in realistic scenarios and collect extensive data about users in order to group types of users into clusters. The following sections explain these trials setup, tests and the analysis over the findings.

3.4.1 Setup

Twenty elderly people participated in these trials: five men and fifteen women. Their average age was 68.5 years old, with the youngest being 55 years old and the oldest being 84 years old. Sessions lasted about 60 minutes, in which participants familiarized themselves with the available means of interaction and executed the required tasks. The measurements taken in these trials were the number of errors, time to finish a task and the observation of the participants' actions.

The system was installed in a PC with two screens: one regular monitor for the "wizard" and one TV for the user. The devices available for interaction in these second trials were the WiiMote, Microsoft Kinect and a Remote Control. Also, speech commands were accept as input. The iPad was tested but independently from the system.

Everything was previously setup in one computer in the Faculty of Sciences in Lisbon, where it was tested out by the local team involved in the project, and only then was sent to Ingema in Spain to make sure the system worked as it was supposed.

Also, one person from Ingema was in Germany at IGD to learn the new features of the user test trials application. With this it was made sure that with the knowledge of the system obtained in the first trials and perceiving the modifications made on the several interaction modules, learning the system would be a very short step before beginning with the main study in Spain.

Concerning the technical modifications in the user initialization application, as mentioned in previous section, there is one major change related with the substitution of the Led Camera Motion Sensor with the Microsoft's Kinect Sensor, which gave a lot more precision to pointing interaction. Also: the Wii remote interaction was improved to give a better precision to the user experience; the Avatar engine was improved to give a more friendly interaction to the user; it was added visual feedback to voice interaction; and finally the test script was modified so that original tests that were excluded from the previous trials could be integrated in this second trials.

3.4.2 Tests implemented

The tasks the users performed and also the questions about their preferences covered six categories: visual, audio, motor, cognitive, avatar and modality requirements. These tasks were essentially the same described in the previous section about the first trials. However, some changes were made (see table 3.4):

Type of tests	**Tasks**	**Devices/Modalities**
Modalities and devices experimentation	Answering to questions related with preferences of interaction, experimentation of each device and modality. Menu items selection and navigation.	Input: Remote control, WiiMote, Kinect. Output: visual menus and Avatar.
Visual capabilities and preferences	Answering to questions related with interface visual configuration (font size and color,etc). Menu item selection and navigation.	Input: one or more devices choosen by the user. Output: visual menus.
Audio capabilities and preferences	Answering to questions related with audio preferences (audio volume).	Input: Speech. Output: Audio.
Cognitive capabilities	Localization of different items on the screen (cognitive scientific tests), measuring time of response.	Input: Speech, WiiMote, Motion Sensor. Output: Visual menus and pictures
Motor capabilities and preferences	Performing gestures. Menu item selection and navigation. Interacting with the tablet, entering numbers, making gestures and interacting with a map. Answering to questions related with motor preferences and pointing mechanisms.	Input: WiiMote, Kinect and tablet. Output: Visual menus.
Avatar preferences	Interacting with avatar. answering to questions related with avatar preferences.	Input: One or more devices chosen by the user. Output:Visual menus and Avatar.
Multimodal preferences	Menu item selection and navigation. Simulation of application contexts of use. Answering to questions related with multimodal interaction and references. Target selection in different positions in the room.	Input:One or more devices chosen by the user. Output: Visual and audio menus and avatar.

Table 3.4: Tests, tasks and modalities

In the first trials only one question was made to choose the most adequate volume of the voice. In these trials two more tests were added, the ability to hear messages at different volumes with the TV sound on and the preferred volume, and preference between female and male voices.

Participants had to perform some additionally tasks in the tablet, such as entering numbers with a virtual keyboard, make different gestures (circular, swipe, pinch gestures) and drag and manage a map using those gestures.

Another new task consisted in the participant to point at a concrete target on the screen from different positions (close versus distant; sit versus stand; front to the screen versus side to the screen) and then say from which position was pointing most comfortable.

Improvements in the avatar were made and now the participants were asked to compare it with text output and then with audio. Also, three pictures of the avatar were shown: head only, upper body and whole body and the participants were required to express their favourite image.

3.4.3 Analysis

The second trials were conducted in the same fashion as the first ones, data was collected through questionnaires, interaction with the system was recorded in a log and behaviours were observed. Once more the presented analyses is more focused in output perceptions and preferences of the users. The complete results can be found in [11].

Visual perception

Text color was evaluated against different color backgrounds. Text colors considered were black, white, red, blue, pink, yellow and orange. The background color used were black, white, blue and green. Table 3.5 shows that white color was preferred for backgrounds black, blue and green and black color was preferred for the white background. Regarding button colors participants could choose between five combinations: white over yellow, white over blue, white over black, black over white, and red over white. Participants

	White	**Black**	**Blue**	**Green**
White	-	14	13	10
Black	15	-	6	6
Blue	4	1	-	3
Red	1	1	0	1
Green	0	0	0	-
Orange	0	1	0	0
Yellow	0	3	0	0
Pink	0	0	1	0

Table 3.5: Participants' preferences regarding text color (in rows) for different background colors (columns)

preferred black letters over white background (8 participants), being the least preferred curiously the opposite combination.

In what concerns text presentation size and color were both again evaluated. Users were asked to read text presented in different sizes, having the largest 100 pixels and the smallest 24 pixels. 9 out of 20 participants preferred the third largest value (64 pixels), being the most voted option.

When asked where users preferred the positioning of the buttons there wasn't a clear conclusion, 8 participants chose peripheral positioning, while 7 chose the centre of the screen, the rest were indifferent. Regarding button's size, the majority preferred big buttons with 70% participants choosing this option. In what concerns inter-button spacing it was presented four screens with different buttons display. One screen with 4 buttons with high inter-spacing, one with 4 buttons with low inter-spacing, other with 8 buttons with high inter-spacing and finally one with 8 buttons and low inter-spacing. Participants preferred always more separation between buttons, but with more buttons presented on screen (10 users).

Audio perception

The audio tasks evaluated the participants ability to perceive messages in different volumes, but this time regarding two different contexts: perceiving audio messages without the sound of TV in background and with the TV sound on. In both contexts the preferred volume was the second loudest volume.

The gender of the voice reading the messages was also assessed, however no clearly preference was shown. 6 participants chose male, 9 chose female and 5 both.

Figure 3.4 shows the percentage of people who could not perfectly heard all the messages given by the avatar in different volumes in both contexts. Aud1 being the one with no TV sound and Aud2 with the TV sound on. It doesn't show any difference between the two cases.

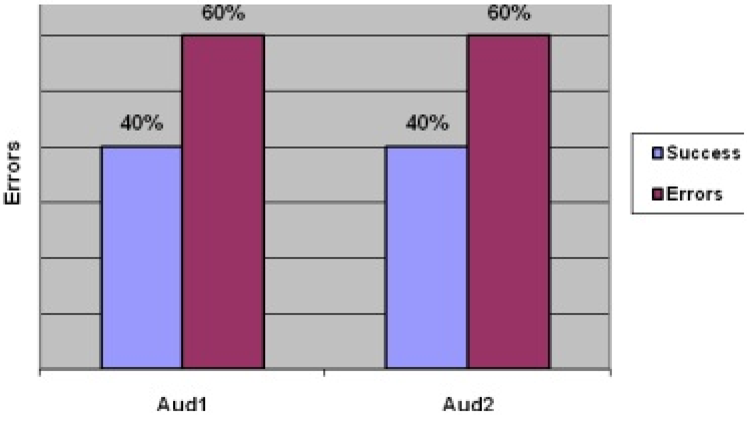


Figure 3.4: Errors made in audio tests

Motor skills

In this category preferences regarding gestures were explored. According to the result found on the first task, 10 participants preferred to select something at the screen with the hand, while just 4 preferred the WiiMote. One of the reasons of this preference can be that when they point at the buttons located in the four corners of the screen they found that these four positions are easier to reach. However, when they were required to use the WiiMote the easiest place to point at was the top right.

Target position	Using hand	Using WiiMote
Top left	1	2
Top center	0	0
Top right	3	7
Bottom right	5	4
Bottom center	1	0
Bottom left	0	0
Left	0	0
Right	0	1
None	2	1
All	8	5

Table 3.6: Participants' preferences regarding target positions using free hand pointing and WiiMote

Regarding the target position preferences when using free hand pointing and WiiMote, participants didn't show any clear preference. However, we can see in table 3.6 that when using the hand all positions seems accessible for most participants (8 users), while using WiiMote the target positioned on the top right side of the screen was preferred.

Cognitive capabilities

Regarding cognitive capabilities two tests were made. One to assess the participants' visual memory, where it was presented different objects in the screen and the participant had to recall their position. The other test evaluates the attention memory, where this time the participant had to focus his attention on one of the objects and recall its position.

As it can be seen in Figure 3.5 the percentage of errors is higher on those questions assessing visual memory (Cog1, Cog2, Cog3) than in those which evaluate the visual attention (Cog4, Cog5, Cog6). In the visual memory questions the more items were required to memorize the higher the error rate was. However, in the visual attentional questions the pattern was the inverse one.

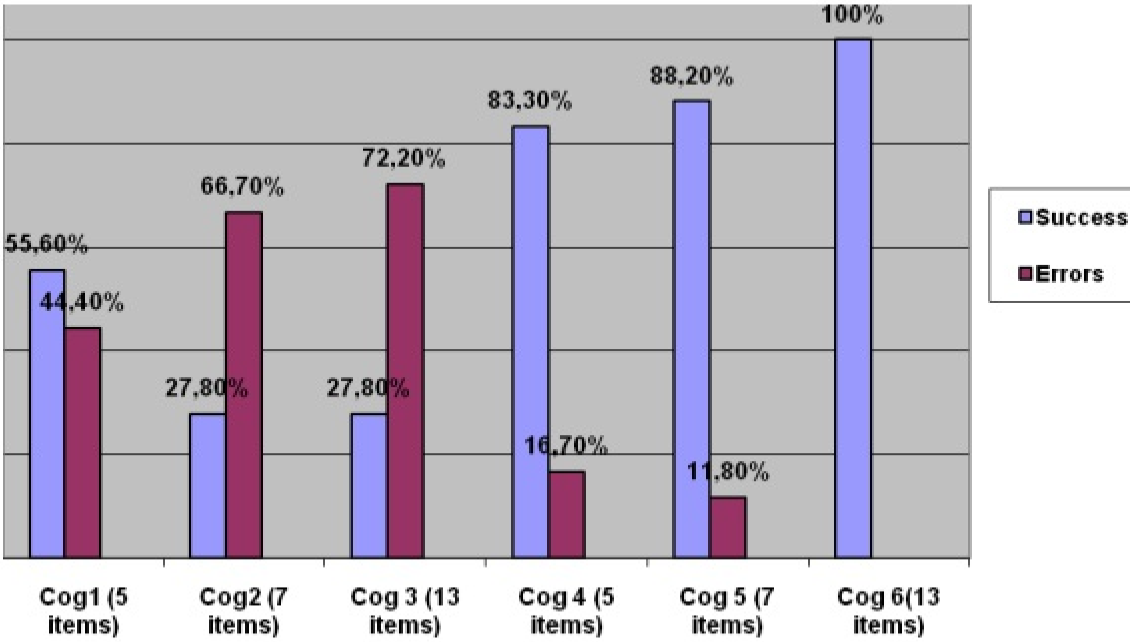


Figure 3.5: Percentage of success in the cognitive tests

Multimodal preferences

Regarding the preferred modality for output feedback, most of the participants wanted to receive system notifications only by audio, though this question was asked without giving a background context (e.g. While seeing TV, GUIDE warns user of an incoming call). But after these questions, different modalities for receiving feedback were tested in two different contexts: watching TV and browsing images (see table 3.7).

Output modalities	While watching TV	While browsing photos
On-screen text	1	1
Avatar speech	2	1
TTS speech	5	5
On-sreen text and avatar	7	4
On-screen text and TTS	4	8
Indifferent	1	1

Table 3.7: Participants' preferences regarding output modalities in different contexts

In the first context there was not a consensus but the majority preferred to receive them by means of a text shown on the screen and read by the avatar at the same time (7 participants). However, this was not the preferred modality when they were browsing the images. In this situation they prefer the text shown on the screen but read for an artificial voice (8 participants). According to what the users reported, they preferred the avatar when they are watching TV because this catch up their attention while in the context of browsing photos they are not so absorbed so they can more easily switch the focus of their attention.

Avatar preferences

Regarding the preference about the use of avatar versus other modalities (table 3.8), we found that choosing between avatar and text output, 11 out of 20 participants preferred the avatar. While, when asked to opt between avatar and TTS (text-to-speech), there wasn't a clear winner.

Output modalities	**Avatar versus Text**	**Avatar versus Speech**
Avatar	11	8
Text	5	-
TTS speech	-	6
Both equality	4	6

Table 3.8: Participants' preferences regarding output modalities versus avatar

In what concerns avatar presentation, participants preferred to see the whole body against upper body or just the head of the avatar.

Disabilities

When trying to comprehend what adaptation is needed for a specific user and his abilities it is necessary to register his abilities and impairments in order to get a more precise study about user's behaviours and preferences when interacting with multimodal systems. Thus unlike the first user trials, in these trials it was taken in account the disabilities of the participants and their preferences. This section describes the most important results.

Participants with visual impairments preferred the biggest button and text sizes, regarding buttons placement they opted for 8 buttons with large inter-spacing. Additionally, they preferred the use of modalities redundantly both for input(e.g. pointing and speech) and output(e.g. text and audio). Also, as expected, they preferred the use of the avatar (whole body) over plain text. However, there wasn't a consensus in some relevant variables such as preferred device for interaction, central or peripheral location of the buttons, text colors, etc.

Participants with hearing disabilities preferred the highest volumes on both scenarios (TV off and on), audio modality always complemented with visual. Curiously they preferred also the avatar over the use of text. Also, it wasn't found a consensus in some relevant variables such as preference about female or male voices and avatar versus TTS.

Regarding participants with cognitive impairments there was no clear tendency in output preferences that could be conclusive.

3.4.4 Discussion

These trials involved more participants than the first one, also some new interaction devices were integrated, bugs were fixed and some new tests were added.

Making a comparison between these two trials we can find some similarities in the results as well some variances. In what concerns visual perception results we can find, regardless of having visual impairments, that the majority of users preferred the use of bigger buttons and wider space between them. However, we can find that in the second trials the participants when asked about the button's color preferences the ”white on black” option was no longer one of the popular choices despite the high contrast. Regarding text presentation, we can assume that users prefer bigger sized fonts but not exaggerated and with high contrast.

When comparing free hand pointing device and WiiMote, the common result was the great acceptance of motion recognition permitting users to interact with the hands in free air in spite of the familiar feel of holding a WiiMote, which is similar to a remote control. Also we can conclude that using different input pointer devices changes the ability to reach some parts of the screen. Thus visual content (e.g. Menus) should be adapted taking this issue into account (e.g. When using Kinect change position of menus to the right corner of the screen).

In what concerns output feedback participants liked the integration of both visual and audio modalities in a redundant way. However we can find in the second trial that it can depend on the application context.

3.5 Conclusion

The multimodal adaptation architecture, as well as the results obtained so far from the different user trials, have clear implications in the development of accessible applications for elderly users. These implications can be relevant to developers as well as to the GUIDE Framework points of view.

Applications should always present a short number of interactive elements for each screen, focusing on big buttons. If developers make complex UIs, GUIDE has to be capable of dividing one screen in multiple screens (and provide navigation through them), or present options to user in alternative modalities.

Applications should make sure both text size and audio volume are configurable by the user at the beginning as well as in the middle of an interaction. If the application by itself doesn't offer this option, GUIDE UI adaptation should offer this possibility.

The existence of a strong relation between arm and item location on screen, will influence the way developers design the layout of their applications, as it also affects the configuration and parametrization of GUIDE presentation manager (fission module), as both have to contemplate the existence of this user-UI relation.

If system feedback and presentation could only be performed in one modality (Avatar, Audio or Visual information), the way to do it would depend on the interaction and application context but also from the user's preferences and capabilities. This is also true for the type of Avatar: head only avatar would be more suitable for talking with the user or giving simple feedback, while half and full-body Avatar would be suitable for instructing the user on how to do certain gestures, or on how to perform certain tasks. However, and independently of which output modality chosen, output should be repeatable every time the user asks for it again, solving problems derived from lack of attention or changes in the context of interaction.

This chapter has shown the importance of the user trials and their results for a user centred system such as GUIDE. Trials showed that each user has its own characteristics and interaction patterns, but they can be grouped into different clusters. Somehow, information about the user must be collected by the system in order to perform the right adaptation. Thus, arose the need to build an application capable of tutoring the user on how to interact with the system and also gathering the user's abilities and preferences in order to match the user to a certain cluster. The result was the User Initialization Application which is described in the next chapter.

Chapter 4

User Initialization Application

The need of making a user initialisation application (UIA) has emerged from the user trials' planning and preparation because each user has different characteristics and studying them wouldn't be enough to make an well driven adaptive experience. The user trials findings resulted in the definition of three types of user profile clusters: Low, medium and high impairments profiles.

In the following sections the UIA will be presented, a TV based application that will run on the GUIDE platform to introduce a new user to GUIDE's interaction process and insert him in a defined profile considering his characteristics. The application is the first contact with the system and therefore needs to tutor the user on how to use the system at the same time it evaluates him.

4.1 Introduction

The User Initialisation application is started when a new user is detected by GUIDE system. As stated before, its first goal is to present a tutorial informing the interactions possibilities with all the input devices available. After feeling comfortable the users have now to follow some instructions with the main focus of evaluating his physical, sensory and cognitive abilities. The aim of UIA is to lead users to think they are just practising their ability to interact with these new technologies (e.g.,pointing and speech) and setting up their environment (e.g. font-size, background and foreground colors, audio volume, etc.). But what the UIA also does is to collect the data it needs to be able to understand the abilities of the user.

As aforementioned GUIDE has defined three clusters profiles from the user trials, these clusters are divided by modalities (Vision, Hearing, Cognition and Motor) and each have their own characteristics deduced also from the trials. Vision is evaluated by close vision, distant vision, general eyesight, seeing at distance, seeing at night and color perception values while the hearing capabilities are evaluated from specific audio frequencies and background noise disturbance. Cognition is evaluated by the cognitive executive function level and Motor capabilities are characterized by a mobility diagnosis, muscular weakness, the ability to write without difficulties, among others. A user for example can belong to a low level vision cluster, medium level motor cluster and high level of impairment of audio sensory. By the end of UIA execution, it has to be able to identify what clusters of impairments the user belongs to.

This will enable a good initial fit, which means the user experience with GUIDE can be adapted right from the beginning, towards a more usable and satisfying interaction with the system. Afterwards, while using other applications, the interaction is monitored and the defining characteristics' values are fine-tuned, so that the user experience can improve constantly.

4.2 Architecture

In its final version UIA is a web based application which runs on GUIDE's environment like the others envisaged applications. However, because none of the GUIDE core components are yet fully implemented this prototype runs as a simplistic version of the system. This prototype is a web based application (HTML and Javascript) that uses an applet (Web Browser Interface in figure 4.1) to perform the communication between the UIA and this simple ”GUIDE core”. All the actions performed in the application are sent to the ”GUIDE core”, for instance the preferences collected by the UIA are received by ”GUIDE core” which uses it to perform changes in the user interface (changing the CSS file).

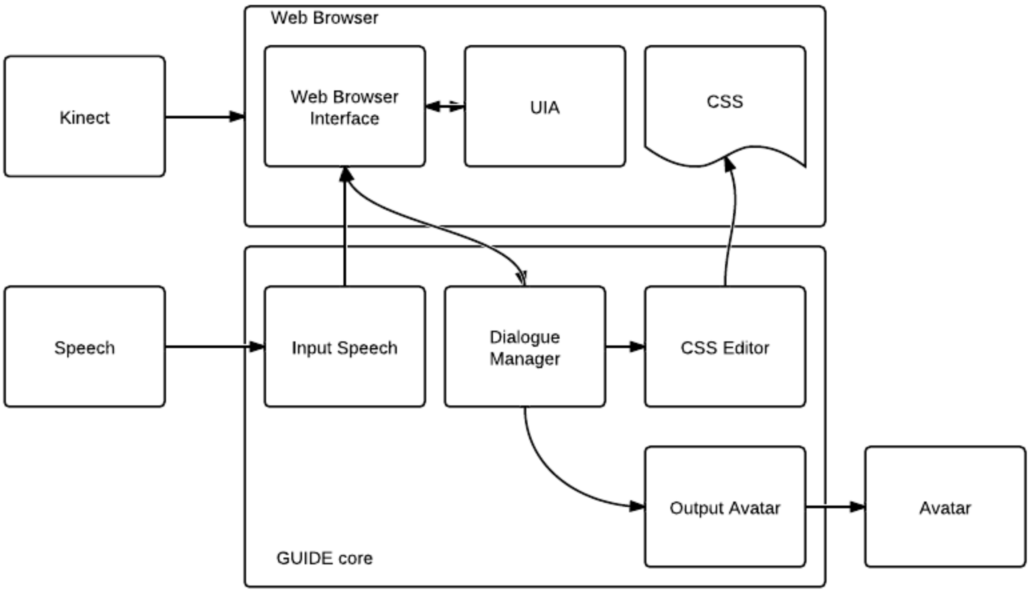


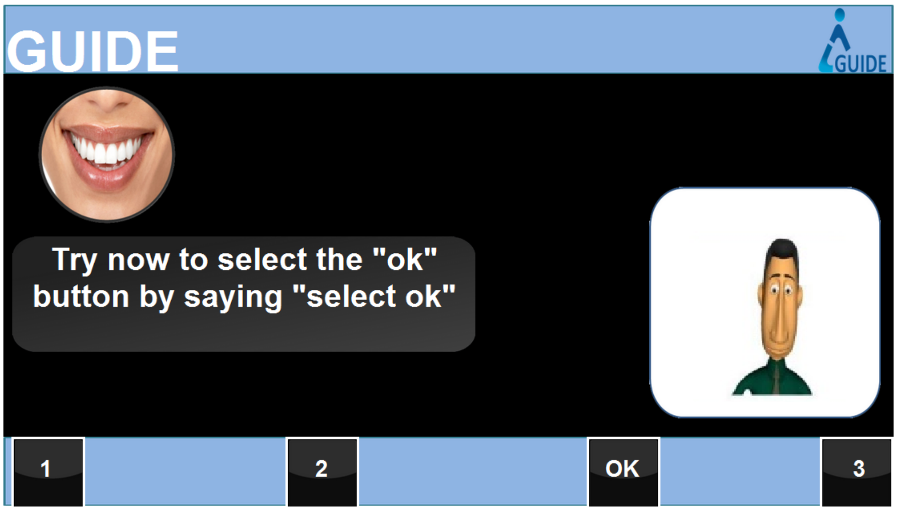
Figure 4.1: Architecture of the User Initialization prototype

The fusion is simplistic and implemented in Javascript outside the Java implementation. We took this option since the events and element information were intercepted by this Javascript component, as at this time we did not have a way to represent the User Interface in the system.

The dialogue manager implemented is also basic and it's main goal is to know which state is to be presented next, in this case a simple list of HTML pages and to deliver other messages to the other components such as the avatar and CSS editor.

The fission module is composed by the avatar communication component and a CSS editor that adapts the user interface according to the information received in run time.

4.3 Tutoring the user



Try now to select the "ok" button by saying "select ok"

Figure 4.2: Tutorial screen example

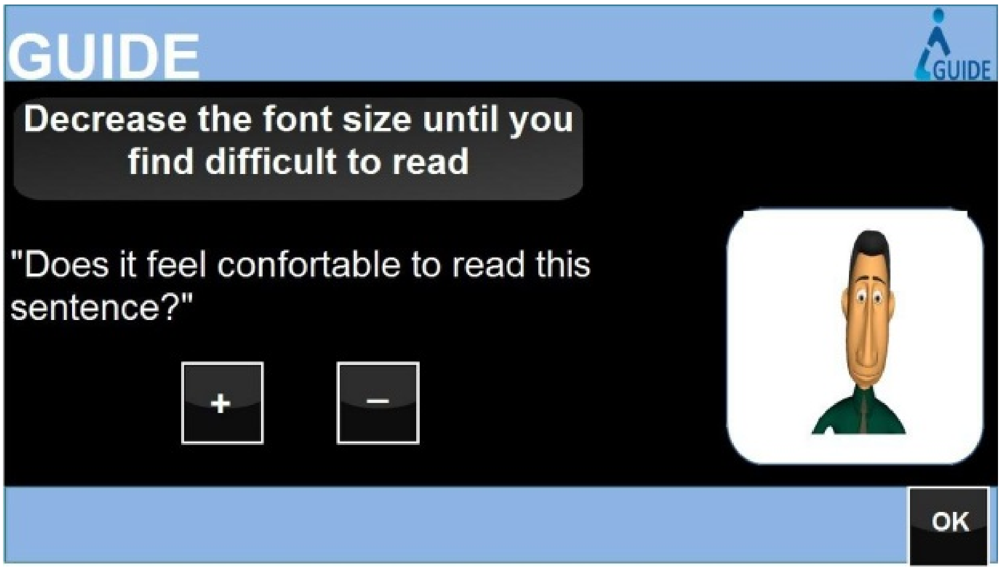
Before the tests to be held the user must have knowledge on how to interact with the system. UIA starts by introducing all the devices which the user can use to interact with the GUIDE platform. UIA shows one by one all the modalities of interaction. UIA leads the user to use the remote control and navigate with it, then it shows that he can navigate using the Wiimote or the kinect pointing towards the screen and selecting the right button. Then introduces voice commands to navigate and select buttons (see figure 4.2).

Once all these are mastered, the user is tutored to combine those modalities and explained that he can use voice and gestures, pointing at a button and saying ”select” will make the same thing as pushing the ”ok” button of a remote control. After the tutorial complete it is now time to proceed to abilities and impairments testing.

4.4 Tests implemented

The UIA prototype include some tests that serves the goal to evaluate all the characteristics of each modality set on the clusters. Being a first version of this application, the tasks presented need to be better defined and although they already capture some variables needed these tasks must be designed to feel less like an exam when actually doing them.

4.4.1 Visual



Decrease the font size until you find difficult to read

"Does it feel confortable to read this sentence?"

Figure 4.3: Visual perception task on UIA

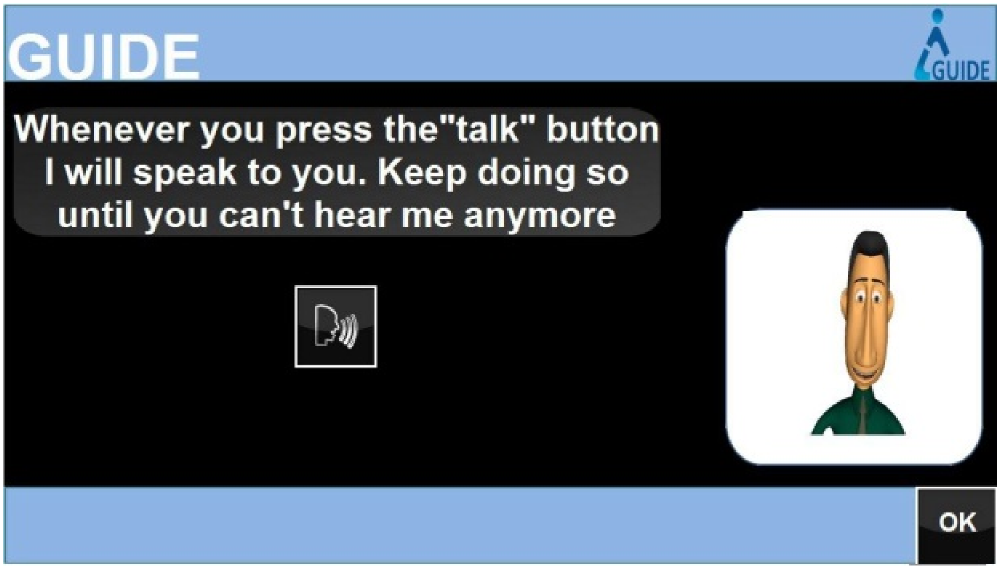
The first task related with visual perception asks the user to adjust the size of a text presented (see figure 4.3) until he or she finds it difficult to read. This task aims to find the range of font size that user is able to read and deduce close/distance and eyesight vision values. Following, it is presented several combinations of text and background colors for user to choose which one is more visible. This task has the goal to evaluate the color perception and comfort.

In what concerns button presentation, the first task asks the user to define the spacing between buttons horizontally and vertically that the user feels acceptable to be able to navigate and select. The goal of this task is to simultaneously evaluate mobility of the user as well as tingling of limbs.

Next, it is presented different combinations of text and background colors for buttons which also evaluates color vision characteristics.

4.4.2 Audio

Related to audio perception two tasks are asked to be performed. In the first one the avatar will read some sentences and each time the user presses a button the volume decreases. The user stops when he or she can't hear the voice any more. In the second one the task is similar but with background noise. This aims to gather the volume range in which user feels comfortable in noisy or non-noisy backgrounds. See figure 4.4 for this task screen.



Whenever you press the"talk" button I will speak to you. Keep doing so until you can't hear me anymore

Figure 4.4: Audio perception task on UIA

4.4.3 Motor

Concerning motor skills, the application asks the user to use his left arm to point at a target on the top right side of the screen, and then the inverse. After each task the UIA asks if it the arm is tired and if it hurts in order to get values for muscular weakness and mobility and rigidity difficulties.

4.4.4 Cognitive

This task (figure 4.5) consists in a game where it is presented on the screen six images representing objects that disappear after a short time period. Then the user is asked to locate a random object. This task aims to evaluate the visual memory and attention capacity of the user.

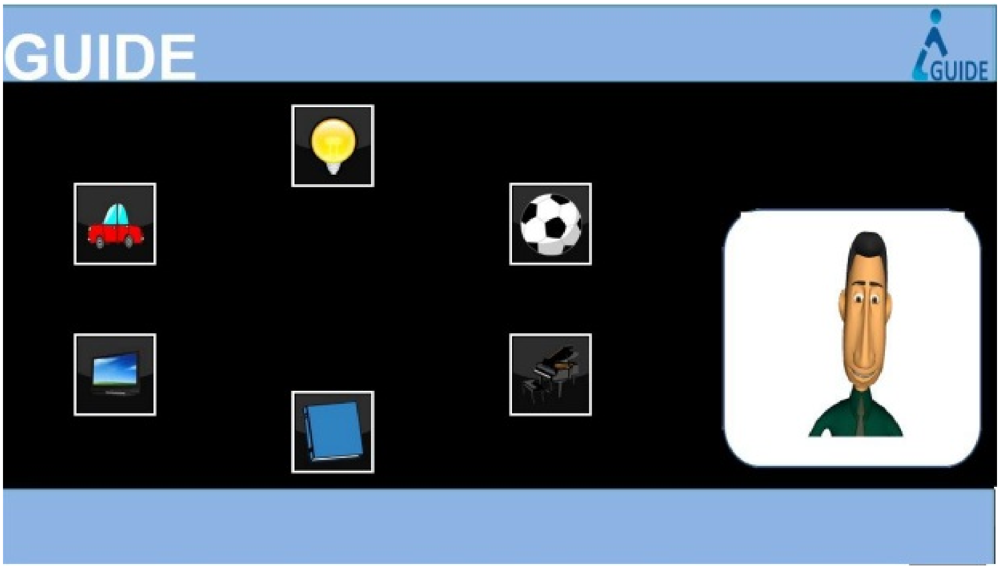


Figure 4.5: Cognitive skill task on UIA

4.5 Conclusion

The User Initialization Application is a key feature of GUIDE to explicitly gather information about the user, as it is the main driver for the adaptation mechanisms included in this project. The information collected is stored in a User Model that will be maintained and updated by the GUIDE core components. This initial prototype will lead to new improved versions with new and enhanced tests, more entertaining and less intrusive.

Chapter 5

Multimodal Fission

This chapter describes the multimodal output fission module's design and implementation and the fission processing decisions on adaptation's depth, presentation management and coordination. The last section describes a prototype of the Fission component and its validation.

5.1 GUIDE core Architecture

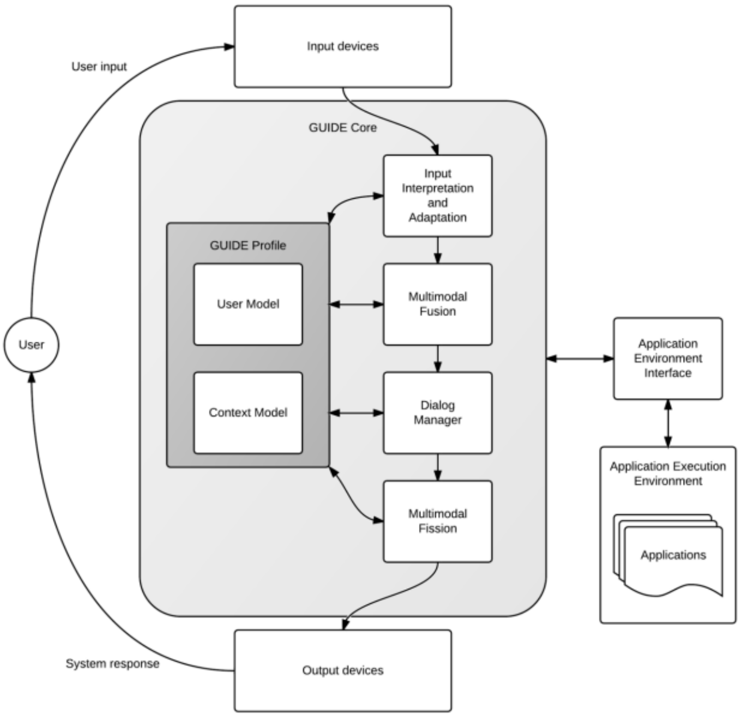


Figure 5.1: Conceptual diagram of GUIDE core

This section describes GUIDE's architecture which is composed by core components responsible for making the adaptation of input and output devices and other components that provide the communication between GUIDE's core and the application, a representation of these components can be consulted in figure 5.1. The components that integrate the GUIDE's core and the information flow between them are described in the remainder of the section.

5.1.1 Input Adaptation

The input adaptation module facilitates pointing in electronic interfaces by users with motor impairment. It attracts the pointer in the middle of a button, if the pointer is in vicinity of the button. So even if the user points towards the edge of a button, the pointer will automatically move to the centre of the button.

5.1.2 Multimodal Adaptive Fusion

In any system that supports multiple modalities for input, such as gestures or speech, there is a need for a multimodal fusion module. This component of the system is responsible for receiving the incoming input from different sources; combine that information according to specific context or user sensitive- information and making an interpretation out of it.

The main task of the fusion engine of GUIDE is to potentially combine any incoming input from the recognizers, make an interpretation of that data and forward it to the Dialogue Manager. The key to provide the most suitable interpretation is to take into account critical information that is provided by three main sources: the user model, context model and input events. In order for Multimodal Fusion to know if the user is trying to select a button, for instance, a list of interactive elements of the current state is firstly sent by the dialogue manager, then fusion requests the fission's adapted UI representation that contains the updated elements' properties (e.g., sizes and positions). For each identified element of the list the Fusion module prepares a set of triggers that can be activated with specific user actions such as pointing for more than two seconds at a button or by saying the name of that button. Then these actions are interpreted as a command (e.g. ”select element x”) and sent to the Dialogue Manager.

5.1.3 Dialogue Manager

The Dialogue Manager (DM) coordinates the activity of several components of the GUIDE Core and its main goal is to maintain a representation of the current state of the on-going dialogue the user maintains with the application or with GUIDE itself for changing some interaction preferences. To this end, the DM receives the interpreted commands from the fusion component and reacts accordingly. It has the responsibility to decide when and which state of the dialogue can be triggered.

When a application starts it is responsible for sending the initial state's user interface representation to the DM. This phase, when GUIDE is waiting for the content to be ren-

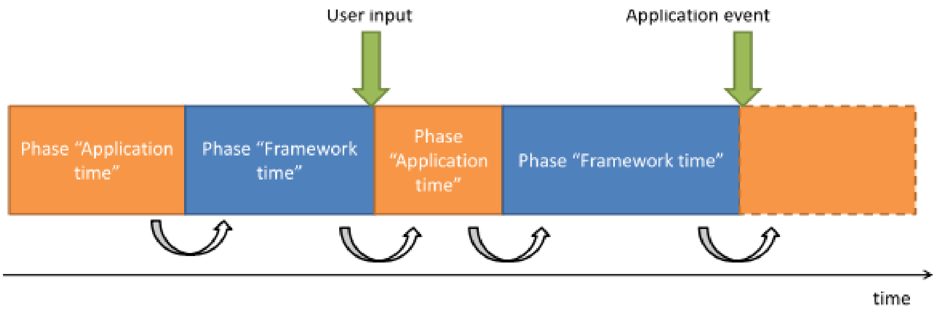


Figure 5.2: Time sharing phases

dered, is called ”Application time” (see figure 5.2). After DM receiving it, it now starts the ”Framework time”. The Dialogue Manager uses the UI representation of the application to instantiate a set of forms which are composed of a list of slots that need to be filled in order to trigger an action (usually a change of state). At the same time the UI representation is also sent to the Fission Module.

5.1.4 Multimodal Adaptive Fission

A multimodal adaptive system should be able to flexibly generate various presentations for the same information content in order to meet the individual user's requirements, environmental context, type of task and hardware limitations. Adapting the system to combine all this time changing elements is a delicate task. The fission module is crucial to make that possible as it takes advantage of multi modalities to overcome sensory impairments that users may have.

When the Fission Module receives the UI representation it parses it and identifies what content and how it is to be presented depending on the User and Context Model's information available. The content may be presented using alternative or complementary modalities. Also the elements of the UI can be modified to meet the user's specificities. Then it sends the instructions to the right output devices and coordinates them originating a coherent presentation. Further information about all the Fission process is fully described on the ”Fission Process” section.

5.1.5 User Model

The User Model in GUIDE stores the information about the user's characteristics, behaviours and preferences. This data is then used to change and adapt parameters in the fusion and fission components in GUIDE core. For instance, fission module can ask the User Model to give a range of recommended values for UI elements properties corresponding to the user characteristics.

This information about the user is gathered initially by GUIDE when it recognises a new user to the system. But after determining the initial user's profile, GUIDE will keep updating the user model to better adapt to its user. The most challenging aspect of this adaptation is how to learn the users' characteristics in a way that can be helpful to them. The adaptation, or the user model updating, will take place after explicit user actions, like, for instance, asking to raise the volume, or asking for larger font sizes. However, GUIDE wishes to go beyond adapting to explicit user instructions, and look for other opportunities to improve interaction with the applications.

5.1.6 Context Model

The context model in the GUIDE Core provides a facility for processing and storing contextual information from the different context sources: User context, Application context, UI components, Hardware context. In what concerns User context we can have, for instance, a change in the user's preference for GUIDE to render application content in a different modality, or to increase the volume due to noise produced by a fan, or GUIDE detects that another user entered the living room.

The Application context results of some dynamic behaviour in the application's logic, e.g. a pop- up, an animation, or some unexpected behaviour such a crash.

All UI input components interact with the GUIDE Core using context events that wrap information representing the input made by the user via some input modality (gesture via pointing device, button signals via a remote control or through speech input). Context information about state of the underlying hardware platform/STB. An example could be that the STB lost TV signal reception, or internet connection, or a new Display has been connected to the STB, etc.

5.2 Output Features

Due to the already mentioned GUIDE's context environment and objectives these are the following output components that will be expected to be used in order to satisfy the project requirements:

- Video rendering equipment (e.g. TV);
- Audio rendering equipment (e.g. Speakers);
- Tablet supporting a subset of: video and audio rendering
- Remote control supporting a subset of: audio rendering, vibration feedback (e.g. Wii remote)

5.2.1 Video rendering

The main medium used for video rendering is obviously the TV. Here is where visual presentations will occur, be them the channels themselves, the adaptive user interface or video streams. A tablet may also be used to clone the TV screen or complement information displayed on the TV screen (e.g. context menus) but essentially is used as a secondary display.

The main user interface should be able to generate various configurable visual elements such as text (e.g. subtitles, information data, etc.), buttons for navigation purpose, images/photos, video (e.g. video conference or media content) and an avatar. In order for the UI be adapted to the user's needs these elements are necessarily highly configurable and scalable (vector-based). Size, font, location, and color are some attributes needed to maintain adaptability. These graphical elements enable the system communication with the users by illustrating, answering, suggesting, advising, helping or supporting through their navigation. The 3D avatar plays a major role for elderly acceptance and adoption of the GUIDE system. An avatar able to do non-verbal expressions like facial expressions and gestures gives the system a more human like communication. Initially a cartoon like Avatar was developed due to the hardware limitations of the Set-top box, though after the user trials almost all participants asked for a more realistic avatar. In order to get a good performance with this restriction, a hybrid avatar is being developed, with some pre-rendered (video) frames and still having run time rendering.

5.2.2 Audio rendering

Audio feedback will be available from TV, tablet or remote control through audio speakers. Audio outputs can be from "simple" non-speech sounds, i.e., rhythmic sequences that are combined with different timber, intensity, pitch and rhythm parameters to speech synthesizers that produce artificial human speech.

Besides the obvious audio-visual output from TV channels or other media (video), GUIDE UI will provide non-speech audio feedback for alarm, warning or status messages or input selection and navigation feedback. These audio signals can act as redundant information to the visual feedback in order to strengthen their semantics.

Synchronized and desynchronized audio-visual presentation will be provided by text-to-speech interfaces. The avatar uses lip-synchronized text-to-speech in order to communicate with the user but in case of hardware limitations or high processor workload some predefined recorded sound files will be playing instead (TTS can be hardware demanding). TV audio replicated by the tablet can act as an enhancer for users with hearing impairments because the mobility of the device makes it possible to be closer to the user ears or use headphones.

5.2.3 Haptic Feedback

Haptic output feedback is done using vibration features present on remote control and/or tablet devices. This modality perceived by the user's tactile sensory is used to add new or redundant information in complement of other modalities for example with visual and audio modalities when an alert or warning message is triggered or used as an input feedback. This mode is a parallel sensory channel to visual and audio sensors which means using them together will not increase the cognitive load.

5.3 Fission Process

5.3.1 Adaptation Level

In order to select the most appropriate modalities to use, it is necessary to define how deep the adaptation level will be. There are three levels of adaptation of the interaction and presentation interface, which can be characterized as Augmentation, Adjustment and Replacement. These levels represent an increasing change to the visual presentation defined by the application, from no change to the visual rendering to a, possibly, complete overhaul. Given that GUIDE aims to support legacy applications (with changes as small as possible to their code) we must consider that these applications have been developed without accessibility concerns towards impaired users. Ideally, GUIDE would be able to evaluate if the application's presentation is close to the recommended presentation parameters for the current user and context specifications (e.g., the text sizes are between the values perceived by the user's vision), and based on that analysis select which adaptation level to apply. In practice, this represents a loss of control for application developers and publishers which they might agree to. As such, the level of adaptation of an application might be limited by the application publisher.

Augmentation

Augmentation is the lightest form of adapting the interface implemented by the developer, which is not subject of change as GUIDE only complements it with other modalities. Usually, applications are developed using primarily visual presentation mechanisms. As a consequence, audio modalities will, foreseeably, be the most used in such situations in the form of redundant information (e.g. speech synthesis of content presented on screen). The UI (its HTML/JavaScript-based embodiment, in the case of Web applications) should be enriched with UI mark-up (e.g., WAI-ARIA), so that GUIDE can extract semantic information of UI elements and render a user-specific multi-modal UI augmentation of it. The fission module also decides when and what information to render based on intercepted events (e.g., new state, mouse hovers, etc.).

Adjustment

Adjustment is the level of adaptation where the interface rendering is adjusted to the abilities of the user and can also be combined with augmentation. Once again, considering applications are primarily developed taking into account visual rendering, this corresponds to adjusting several parameters of the visual rendering (e.g. font size and contrast). If other modalities are employed, their parameters can also be target of adaptation (e.g. adjusting audio volume). What Fission does is to obtain all the recommended visual properties set to a specific user and evaluate the application's ones, then setting the right values (small adjustments can be done for presentation arrangement purposes). The resulting UI representation is now sent to the Application Environment Interface which translates those changes to a format the application is capable of rendering (e.g. translating the UIML representation to HTML or CSS changes that can be applied to the DOM, for Web applications).

Replacement

Replacement level is the most complex adaptation scheme as it means that, not only presentation changes can be made to the application's interface (i.e. the adjustment level), but it can also result in the replacement of some interactive elements for others (e.g. menus for buttons) or even in the distribution of content over different modalities or different screens in case of visual rendering. This level is extremely useful for users with cognitive impairments as in their navigation through an application they are often lost due to the tangle of menus and buttons displayed. The content of an application state can be simplified and divided by various screens or rendered through other modalities such as the Avatar or speech synthesis. Due to its complexity this level of adaptation is still being assessed in order to know its feasibility and integration in the project.

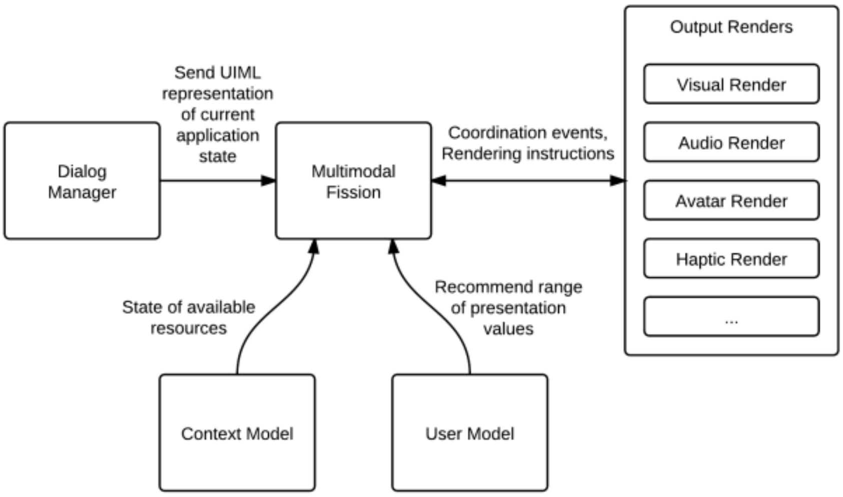

Output Renders
Visual Render
Audio Render
Avatar Render
Haptic Render
...
Send UIML representation of current application state
Dialog Manager
Multimodal Fission
Coordination events, Rendering instructions
State of available resources
Recommend range of presentation values
Context Model
User Model

Figure 5.3: Information for modality selection during multimodal fission

5.3.2 Modality Selection and Evaluation

After the selection of the adaptation level best suited to the situation, the modalities to render the content are chosen through weights selected in accordance to the availability or limitations of resources (e.g., speech synthesizer, avatar, haptic render, visual render, etc.) contained in the Context Model and with the user specificities documented in the User Model, as represented in Figure 5.3.

Regarding the example of visual presentation parameters, assuming that a user has visual impairments and the UI needs to be adapted to his situation, the elements' properties cannot be simply altered (text or button sizes for example) and hope for the presentation to still be coherent. As aforementioned the context model has the data about the resources and their settings (e.g. TV screen resolution). Using that information the fission module is capable of calculating the best values within the recommended ones by the User model maintaining the coherency in the presentation (e.g. assure that bigger buttons will not overlap each other).

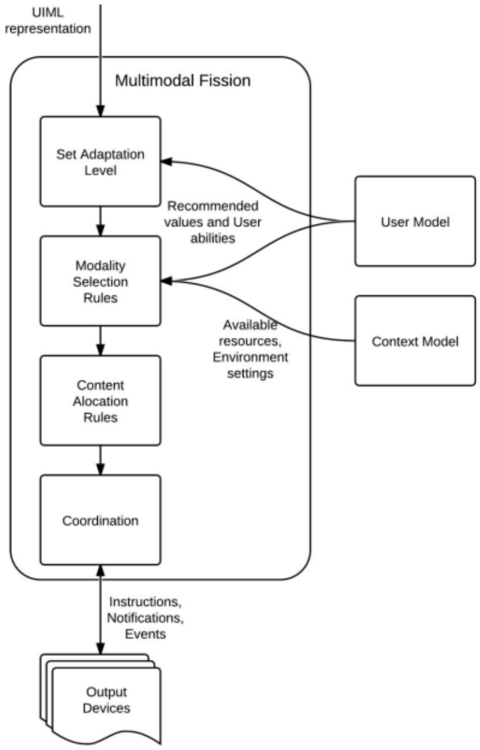


Figure 5.4: Multimodal fission presentation generation process

Once the presentation is allocated to the different modalities, it must now be instantiated. Output coordination abides by the following aspects: physical layout (when using more than one visually presented modality, like text, images or video, the individual components of the presentation must be defined), temporal coordination (when using dynamic modalities like voice synthesizers, videos or sounds, these components must be coordinated in order to achieve the presentation's goal) and referring expressions (when producing multimodal and cross-modal – interaction between two or more sensory modalities – expressions that make references using multimodalities or referring to another part of the presentation will need coordination work). This process is represented in Figure 5.4.

5.3.3 Coordination

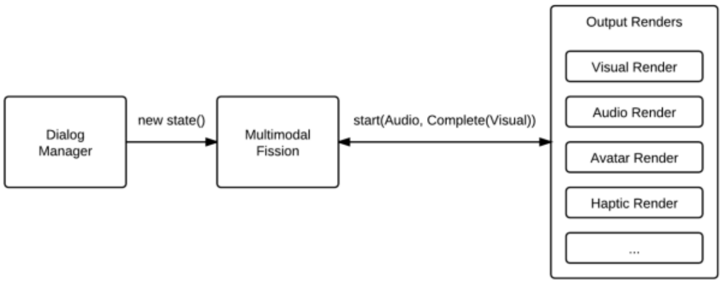


Figure 5.5: Example of rendering synchronization

To synchronize the presentation flow, coordination events will be sent to the bus in order to start or stop rendering, or to be notified when a render is completed. Figure 5.5 shows a simple example of render synchronization where the audio synthesis (e.g., button label) will only start after the visual rendering of the button is complete. The synchronized events are useful because of the need to interrupt the presentation flow when a user makes a decision on the interaction (e.g. selects a button) and a new state needs to be rendered before the rendering of the previous one is completed.

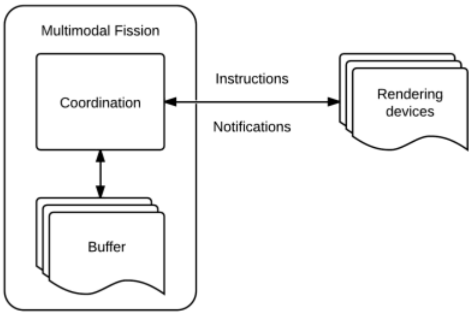


Figure 5.6: Synchronization between multimodal fission and rendering devices

The rendering instructions could be sent in two ways: in the first one the rendering instructions are sent all at once and each output would have a buffer to handle each instruction; in the second one the buffering is made in the Fission component, which sends only one instruction for each device at a time. The device will then respond with a notification of completion or failure (Figure 5.6). The second option presents a set of advantages:

- Instructions that do not get a chance to be rendered because a new state needs to be loaded due to user intervention are not even sent to the rendering devices, saving bandwidth.

- Coordination can be guaranteed by the Fission component based on notification of completion and failure. If all instructions were sent to rendering devices at once, coordination would be much more complex because rendering devices cannot be expected to be aware of the state of other rendering devices.

5.3.4 Communication Events

This module communicates with other components of the GUIDE core, using the context bus, but also with output components connected to the core, using the output bus. The fission component requires information from the Dialogue manager for updating presentation (reflecting changes in the state of the dialogue between user and application), from the User model with the most adequate presentation options for the current user and from the context model for deciding the most adequate presentation based on the current interaction environment. The Fission component communicates information to the various output devices and to the application which is responsible for updating its visual interface. A summary of the events produced and received are presented on the table 5.1.

Name	In/Out	Bus	Description
InputEvent	IN	Context	Published by Dialogue Manager, User Model and Context Model
ContextEvent	IN	Context	Published by Context Model
OutputEvent	IN/OUT	Output	Subscribed by Dialogue Manager and Output devices
ServiceRequest	OUT	Service	Subscribed by Web Browser Interface

Table 5.1: List of events published and subscribed by Fission Module

Input

Fission module needs to have knowledge of the Application's UI and that information must be structured and has to contain all elements and their properties in order to be possible to adapt that content to the user. So several abstract and concrete user interface mark-up languages were taken into account to be used as the UI standard representation for GUIDE:

XForms [1] is an XML application that represents the next generation of forms for the Web, and has introduced the use of abstractions to address new heterogeneous environments. When comparing XForms with HTML Forms, the main difference, apart from XForms being in XML, is the separation of the data, from the markup of

[1] Xforms - `http://www.w3.org/MarkUp/Forms/`

the controls. This not only makes XForms more tractable, by making it clear what is being submitted and where, but it also eases reuse of forms, since the underlying essential part of a Form is no longer bound to the page it is used in. A second major difference is that XForms, while designed to be integrated into XHTML, is no longer restricted only to be a part of that language, but may be integrated into any suitable markup language. In the XForms approach, forms are comprised of a section that describes what the form does, called the XForms Model, and another section that describes how the form is to be presented.

UsiXML [2] which stands for USer Interface eXtensible Markup Language, is a XML-compliant markup language that describes the UI for multiple contexts of use such as Character User Interfaces (CUIs), Graphical User Interfaces (GUIs), Auditory User Interfaces, and Multimodal User Interfaces. In other words, interactive applications with different types of interaction techniques, modalities of use, and computing platforms can be described in a way that preserves the design independently from peculiar characteristics of physical computing platform.

XIML [3] is an extensible XML-based specification language for multiple facets of multiple models in a model-based approach, developed by a forum headed by RedWhale software. It was introduced as a solution that enables a framework for the definition and interrelation of interaction data items.

UIML [4] User Interface Markup Language, is an example of a language that has addressed the multi-device interface issue, an XML-compliant language that supports a declarative description of a user interface in a device-independent manner. In UIML, a user interface is a set of interface elements with which the user interacts. These elements may be organized differently for the different categories of users and types of appliances. Each interface element has data (e.g. text, sounds, images) for communicating information to the user, and it also receives information from the user using artefacts (e.g. a scrollable selection list) from the underlying application. Since such artefacts can vary from device to device, the actual mapping (rendering) between an interface element and the associated artefact (widget) is done using a style sheet.

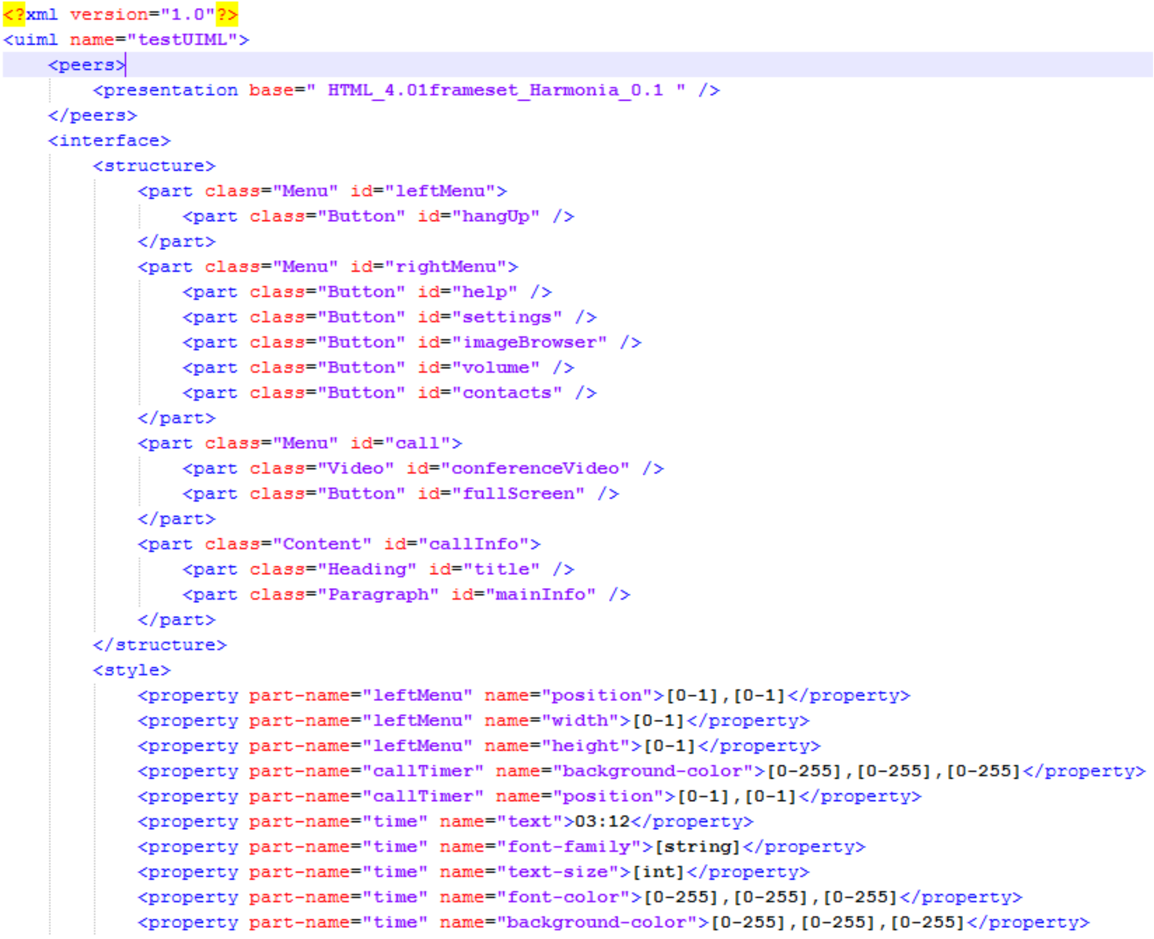
XForms cannot cover all the requirements for GUIDE as it is incapable of describing more specific properties or attributes of certain elements (Buttons, Images, etc.) such as position values, size, colour or other style properties. Although much more complete than XForms, UsiXML fails in giving the location of objects which is an important property

[2] UsiXML - `http://www.usixml.eu/`

[3] XIML - `http://www.ximl.org/`

[4] UIML - `http://www.uiml.org/`

for Fusion and User Model components in the GUIDE Core. Although the tag and property names are a bit sloppy they are indeed complete but the main drawback connected to XIML is the fact that, differently from the majority of the other User Interface description languages, it is developed within a software company, and therefore its use is protected by copyright. UIML specification does not define property names. This is a powerful concept, because it permits UIML to be extensible: one can define whatever property names are appropriate for a particular element of the UI. For example, ”colour” might be a useful property for a button, while ”text-size” might be appropriate for a label. This flexibility allows us to define all the properties needed for all GUIDE components (Input Adaptation, Fusion, Fission, Dialogue Manager, User Model, etc.). Additionally, they might be used to represent the information developers might provide using WAI-ARIA markup tags. UIML seems to be the most complete and flexible UI representation, therefore it was chosen as the standard language to be used inside the GUIDE Core.



```
<?xml version="1.0"?>
<uiml name="testUIML">
    <peers>
        <presentation base=" HTML_4.01frameset_Harmonia_0.1 " />
    </peers>
    <interface>
        <structure>
            <part class="Menu" id="leftMenu">
                <part class="Button" id="hangUp" />
            </part>
            <part class="Menu" id="rightMenu">
                <part class="Button" id="help" />
                <part class="Button" id="settings" />
                <part class="Button" id="imageBrowser" />
                <part class="Button" id="volume" />
                <part class="Button" id="contacts" />
            </part>
            <part class="Menu" id="call">
                <part class="Video" id="conferenceVideo" />
                <part class="Button" id="fullScreen" />
            </part>
            <part class="Content" id="callInfo">
                <part class="Heading" id="title" />
                <part class="Paragraph" id="mainInfo" />
            </part>
        </structure>
        <style>
            <property part-name="leftMenu" name="position">[0-1],[0-1]</property>
            <property part-name="leftMenu" name="width">[0-1]</property>
            <property part-name="leftMenu" name="height">[0-1]</property>
            <property part-name="callTimer" name="background-color">[0-255],[0-255],[0-255]</property>
            <property part-name="callTimer" name="position">[0-1],[0-1]</property>
            <property part-name="time" name="text">03:12</property>
            <property part-name="time" name="font-family">[string]</property>
            <property part-name="time" name="text-size">[int]</property>
            <property part-name="time" name="font-color">[0-255],[0-255],[0-255]</property>
            <property part-name="time" name="background-color">[0-255],[0-255],[0-255]</property>
```

Figure 5.7: Example of an UI representation in UIML

In a UIML document (figure 5.7 is an example) there are two main sections for UI representation, Structure and Style. The Structure section is an abstract representation of the UI elements (parts) where we can give them some contextual meaning such as grouping buttons in a ”Menu” or setting what is a textual ”Content”. In the Style section it is described concrete representation properties such as font-sizes, colors, positions, etc.

Each state of the application is sent has an input event and received by the fission module through a Context Bus in which it is subscribed. Input events are not only requests from the dialogue manager to update the output to a new state, they are also responses from the User Model with the most appropriate settings for rendering to the current user and responses from the context model with the current contextual conditions to use in selecting the best modalities. Other type of event is a context event where the fission module receives contextual notifications such as updates on the current interaction from the context model (e.g., silent environment changes to a noisy one), this meaning updating the presentation (e.g., increase volume or avoid speech synthesis). The fission module is also subscribed to output events which are responses from output devices to coordination operations (e.g. Avatar successfully renderedthe content).

Output

The fission module obviously needs to publish events which are sent to the Output bus and Service bus. The fission component produces output events to instruct the different output devices to render the appropriate output, as defined in the fission processing. Output events are also sent to output devices in order to start, cancel or stop rendering.

The fission component uses the service bus to communicate with the application, which is responsible for updating its visual status, through the Web Browser Interface (WBI). The change requests consists of a description of the new UI, represented in a platform agnostic language (UIML) that is to be translated to the UI language of the execution platform by an appropriate component located within the WBI.

5.4 Fission Prototype

5.4.1 Architecture

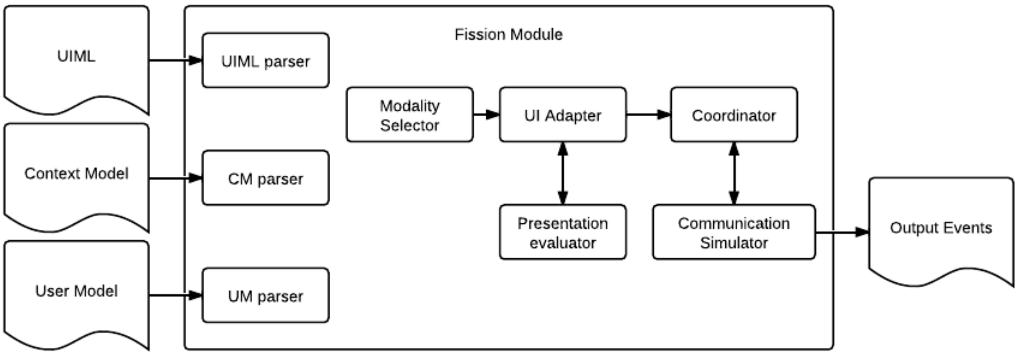


Figure 5.8: Architecture of the Fission Module prototype

This section describes the architecture of the fission prototype seen on 5.8.

UIML file This XML document represents a state of an application to be rendered. This content should be an event sent by the DM but while the core components and the communication system are not fully developed yet, this prototype reads and parses the UI representation from a file.

Context Model file Context Model should be a component and not a XML file but once more in this prototype it is read and parsed. This file contains data about the devices available and their configuration (e.g. TV resolution).

User Model file This prototype loads the XML and parses it. This file contains the set of modalities and user's level of impairment for each modality. Then for each modality it is described the recommended values for visual elements, audio and avatar renders (e.g. minimum and maximum font and element sizes for a button, minimum and maximum volume levels, etc.).

Output Events file This file is the result of output events generated by the fission module which contains the messages, sent and completion intervals.

		Complement		Adapt	
Modality	Impairment Level	Visual	Auditive	Visual	Auditive
Visual	none	0	0	0	0
	mild	0	1	1	0
Auditive	none	0	0	0	0
	mild	0	0	0	1
Cognitive	none	0	0	0	0
	mild	0	1	1	1
Motor	none	0	0	0	0
	mild	0	0	1	0
Visual & Auditive	mild	0	1	1	1
Visual & Motor	mild	0	0	1	0
Auditive & Motor	mild	0	0	1	1
Auditive & Visual & Motor	mild	0	1	1	1

Table 5.2: Selection of modalities to complement impairments and adapt

Modality Selector This class is responsible for selecting the modalities to use for a specific user. First the level of adaptation (augmentation, adjustment or replacement) is chosen ideally deducing from the compatibility of the user interface with user's characteristics. But in this prototype, the level of adaptation is described on the user model. The next step is to access impairment levels (none or mild) and decide what are the modalities to use (e.g. with mild visual impairments, a redundant audio description of the presentation should be rendered). Table 5.2 describes all possibilities and combinations of modalities that this component assess. ”Complement” modalities are the modalities used to give redundant information to the user. ”Adapt” modalities are the modalities that needs to be adapted, i.e., a user with mild visual impairments needs the visual content adapted to his characteristics while Audio complements the visual impairment and as the user has not auditive impairments there is no need to adapt it. This simple approach was opted due to processing restrictions imposed by the set-top-box. Other approaches can be assessed after doing performance evaluation using a set-top-box.

UI Adapter The UI Adapter component requires the data to adapt the content to present. For example, the volume level for the TTS or Avatar is calculated based on the recommended levels from UM, environment context from CM and the preferences of the user. Though in this prototype some values are calculated by average between the minimum and maximum values. Visual elements are adapted but need to be evaluated to ensure presentation coherency.

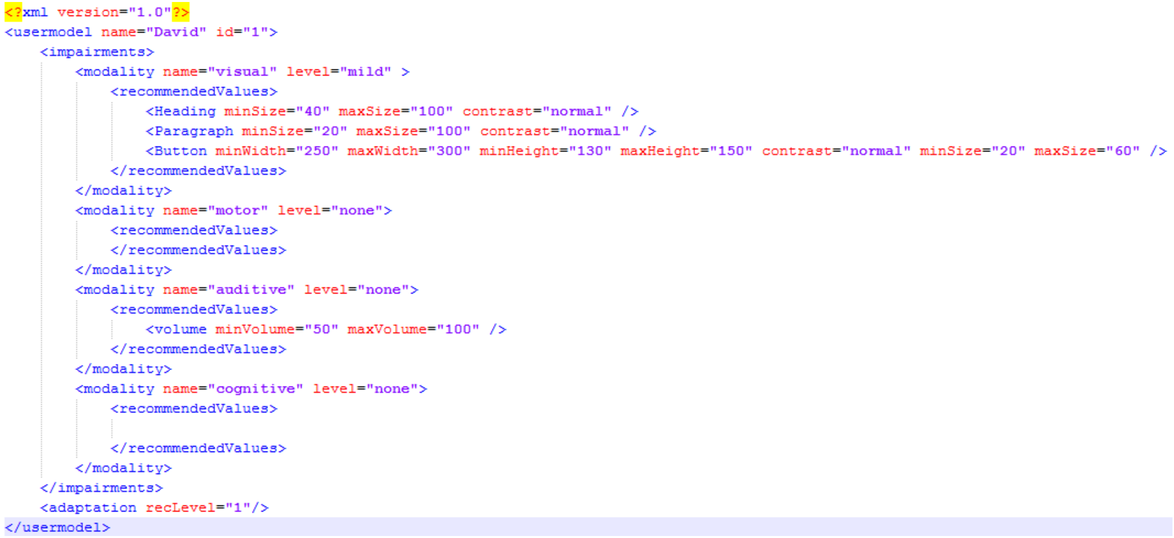
Presentation Evaluator This class is responsible to assess the validity of the presentation (visual content). UI adapter sends the adapted UIML to the evaluator and it calculates if there is any element that is out of the screen boundaries or overlapping other element. In the case of not satisfying these requirements it is declined and the UI adapter calculates new sizes or positions for the elements. The current prototype is only capable to evaluate if an element is exceeding the screen limits.

Coordinator This component is responsible for laying out the output events in the right order (e.g. visual content should be first rendered before the avatar starts to speak). The messages are inserted into a buffer and sent one at time waiting to receive a response from the device (completion or failure rendering).

Communication Simulator This class simulates the reception of events and rendering of devices by randomly attributing times for render and sending the confirmation of completion to the coordinator component. After all events are completed it records them in the output events file.

5.4.2 Use case

This section describes a use case of adaptation using the prototype developed. The user model in figure 5.9 shows that the user suffers from mild visual impairments and the level of adaptation for the presentation screen is set to 1 which means ”Adjustment”, therefore visual adaptation is recommended. Thus consulting the table 5.2 for persons with mild visual impairments the system will adapt visual content to user’s characteristics and redundant auditive modality will complement the impairment. Fission module now requests from the user model the recommended values for visual elements. Figure 5.9 shows that

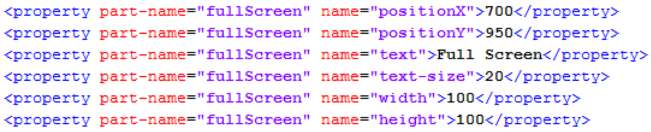


```
<?xml version="1.0"?>
<usermodel name="David" id="1">
    <impairments>
        <modality name="visual" level="mild" >
            <recommendedValues>
                <Heading minSize="40" maxSize="100" contrast="normal" />
                <Paragraph minSize="20" maxSize="100" contrast="normal" />
                <Button minWidth="250" maxWidth="300" minHeight="130" maxHeight="150" contrast="normal" minSize="20" maxSize="60" />
            </recommendedValues>
        </modality>
        <modality name="motor" level="none">
            <recommendedValues>
            </recommendedValues>
        </modality>
        <modality name="auditive" level="none">
            <recommendedValues>
                <volume minVolume="50" maxVolume="100" />
            </recommendedValues>
        </modality>
        <modality name="cognitive" level="none">
            <recommendedValues>

            </recommendedValues>
        </modality>
    </impairments>
    <adaptation recLevel="1"/>
</usermodel>
```

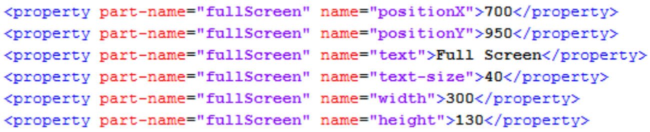
Figure 5.9: Example of a User Model

text headings, content text and buttons should have text sizes between a minimum and a maximum where the maximum value is the most suitable for the user. Buttons also have a recommended size for this specific user.



```
<property part-name="fullScreen" name="positionX">700</property>
<property part-name="fullScreen" name="positionY">950</property>
<property part-name="fullScreen" name="text">Full Screen</property>
<property part-name="fullScreen" name="text-size">20</property>
<property part-name="fullScreen" name="width">100</property>
<property part-name="fullScreen" name="height">100</property>
```

Figure 5.10: Example of a button in UIML before adaptation

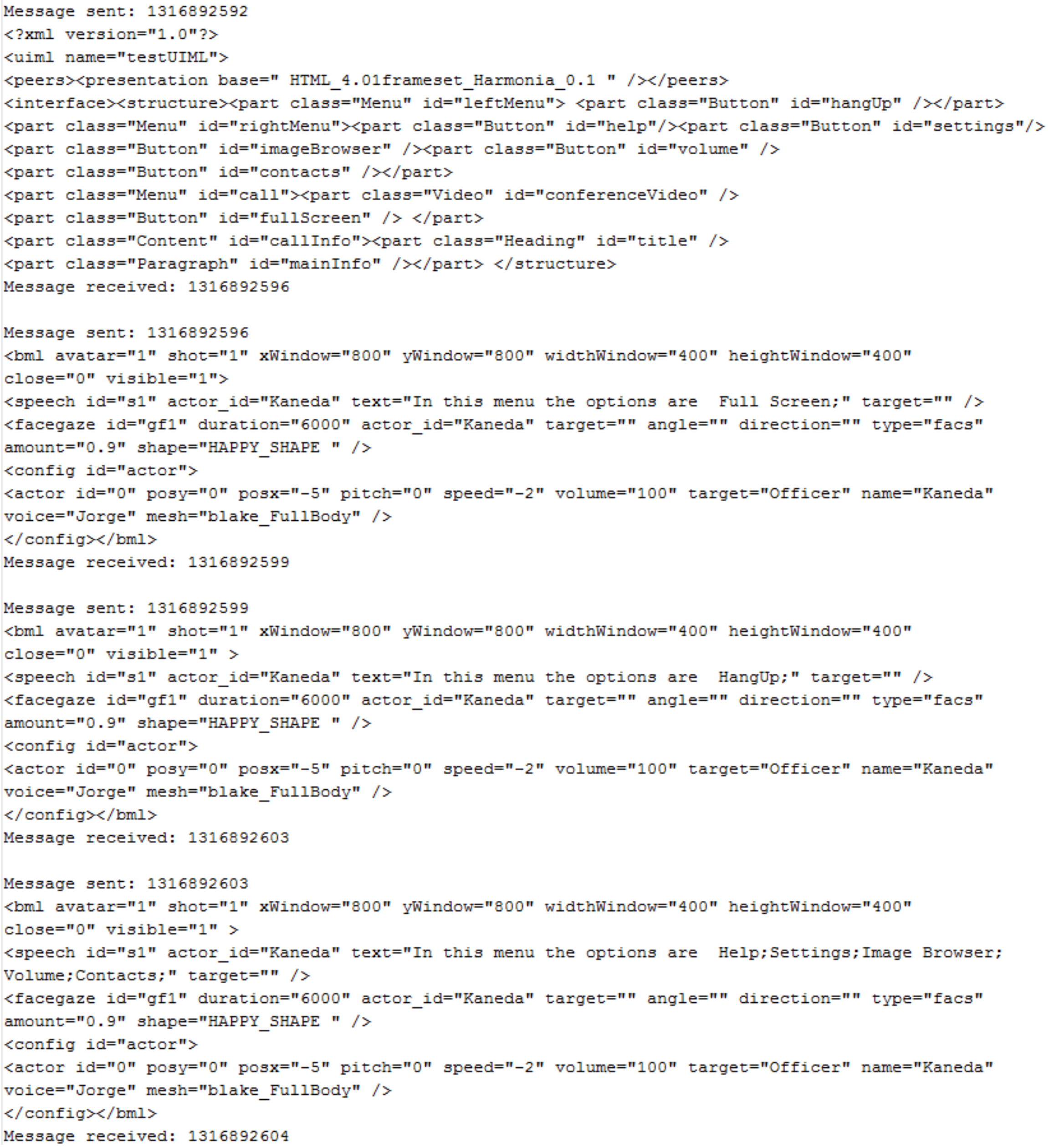


```
<property part-name="fullScreen" name="positionX">700</property>
<property part-name="fullScreen" name="positionY">950</property>
<property part-name="fullScreen" name="text">Full Screen</property>
<property part-name="fullScreen" name="text-size">40</property>
<property part-name="fullScreen" name="width">300</property>
<property part-name="fullScreen" name="height">130</property>
```

Figure 5.11: Example of a button in UIML after adaptation

Figure 5.10 shows an excerpt of UIML visual representation before performing adaptation. When the fission module starts adapting the visual content it evaluates the compatibility of changing element's properties. Thus it requires CM to give relevant data about the visual output device settings, in this case the TV. Fission module calculates the best button size (between the recommended ones) in order for none of the visual elements to reach the screen limits. If the minimum value still doesn't achieve the requirements to a coherent presentation, the original values from UIML will be used. Figure 5.11 shows that the button's width is the highest recommended value from the user model but the height is the minimum recommended value as higher sizes would be incompatible. For the text size of the button it was performed an average between the two recommended values, but ideally it should be also related with the button's size.

Finally figure 5.12 shows the events generated from the coordination process. It is visible that visual content was first sent to be rendered and then the events to render auditive description of the presentation were next. Each message was only sent after receiving confirmation of render completion.



```
Message sent: 1316892592
<?xml version="1.0"?>
<uiml name="testUIML">
<peers><presentation base=" HTML_4.01frameset_Harmonia_0.1 " /></peers>
<interface><structure><part class="Menu" id="leftMenu"> <part class="Button" id="hangUp" /></part>
<part class="Menu" id="rightMenu"><part class="Button" id="help"/><part class="Button" id="settings"/>
<part class="Button" id="imageBrowser" /><part class="Button" id="volume" />
<part class="Button" id="contacts" /></part>
<part class="Menu" id="call"><part class="Video" id="conferenceVideo" />
<part class="Button" id="fullScreen" /> </part>
<part class="Content" id="callInfo"><part class="Heading" id="title" />
<part class="Paragraph" id="mainInfo" /></part> </structure>
Message received: 1316892596

Message sent: 1316892596
<bml avatar="1" shot="1" xWindow="800" yWindow="800" widthWindow="400" heightWindow="400"
close="0" visible="1">
<speech id="s1" actor_id="Kaneda" text="In this menu the options are  Full Screen;" target="" />
<facegaze id="gf1" duration="6000" actor_id="Kaneda" target="" angle="" direction="" type="facs"
amount="0.9" shape="HAPPY_SHAPE " />
<config id="actor">
<actor id="0" posy="0" posx="-5" pitch="0" speed="-2" volume="100" target="Officer" name="Kaneda"
voice="Jorge" mesh="blake_FullBody" />
</config></bml>
Message received: 1316892599

Message sent: 1316892599
<bml avatar="1" shot="1" xWindow="800" yWindow="800" widthWindow="400" heightWindow="400"
close="0" visible="1" >
<speech id="s1" actor_id="Kaneda" text="In this menu the options are  HangUp;" target="" />
<facegaze id="gf1" duration="6000" actor_id="Kaneda" target="" angle="" direction="" type="facs"
amount="0.9" shape="HAPPY_SHAPE " />
<config id="actor">
<actor id="0" posy="0" posx="-5" pitch="0" speed="-2" volume="100" target="Officer" name="Kaneda"
voice="Jorge" mesh="blake_FullBody" />
</config></bml>
Message received: 1316892603

Message sent: 1316892603
<bml avatar="1" shot="1" xWindow="800" yWindow="800" widthWindow="400" heightWindow="400"
close="0" visible="1" >
<speech id="s1" actor_id="Kaneda" text="In this menu the options are  Help;Settings;Image Browser;
Volume;Contacts;" target="" />
<facegaze id="gf1" duration="6000" actor_id="Kaneda" target="" angle="" direction="" type="facs"
amount="0.9" shape="HAPPY_SHAPE " />
<config id="actor">
<actor id="0" posy="0" posx="-5" pitch="0" speed="-2" volume="100" target="Officer" name="Kaneda"
voice="Jorge" mesh="blake_FullBody" />
</config></bml>
Message received: 1316892604
```

Figure 5.12: Output events resulted after running fission module prototype

5.4.3 Discussion

The prototype developed is capable to choose the best modalities to perform adaptation based on the user model using a simple binary rule method. It evaluates the content in order to change visual properties with compatible values. It also coordinates the presentation flow, ordering which and when devices should start rendering.

Though it is far from complete as it only addresses visual and auditory modalities and ignore adaptation needs from Cognitive (e.g. relevant content should transit from one state to another to maintain the user on track) and Motor impairments (e.g. add more button's inter-spacing to facilitate selection), this first prototype is ready to be integrated into the GUIDE framework to coordinate with its different elements, and be used in the next user trials. While the integration process goes on, the development of the missing features will occur in parallel.

The current prototype resorts to the use of XML files simulating the inputs and a Communication Simulator simulating the coordination with output devices. This was done in order to be able to start evaluating the performance of the multimodal fusion component. The integrated version of this component will incorporate the communication events currently under definition in the scope of the prototype, but benefiting from the debugging already conducted thanks to the use of the aforementioned components.

Chapter 6

Conclusions

GUIDE tries to deliver accessibility features to TV based applications without any effort from the developers as the adaptation process is the framework's responsibility. It was showed that GUIDE provides a set of multimodal devices capable to assure that a user with some kind of functional limitation or impairment (vision, hearing, motor and/or cognitive) can still enjoy his interaction with the system (home automation, media access, video conferencing, tele-learning, etc.). With these applications GUIDE can help to simplify elderly daily lives, expand their social network and enhance their understanding of the world lowering the e-exclusion and loneliness levels.

Two phases of user trials were performed with the goal of collecting preferences feedback as well as raw data to study this specific target of population. The major outcomes were user requirements, preliminary guidelines for prototype improvement of the multimodal fission module and other core components, user preferences on multimodality interaction and clustering of user models based on user's abilities.

To help to perform the user studies, the user trials application was developed and provided multimodality features (input and output). Users were capable to interact using and combining a wide range of devices and the content was generated from a UI description file (visual and audio elements). The applicability of this prototype is beyond user trials and questionnaires as it can be used to simulate the behaviour of any multimodal project besides GUIDE in early stages.

Based on some of the findings found in the user trials, the design and implementation of a first multimodal fission prototype was developed and assessed. This prototype is capable to perform different levels of adaptation and decide the best modalities to render the presentation content. The modality selection is based on the profile specifications of the current user in order to give him the best user experience when interacting with GUIDE applications.

The first prototype does not yet meet the planned goals, due to several delays in several specifications of the different components comprising the GUIDE core and communication languages. Nevertheless, it already addresses two of the major limitations (visual and audio) that are expected to have an impact in users performance while using the GUIDE framework. It should be stressed, that the delays have raised the need to develop extra components (e.g. the communication simulator) in order to guarantee that the debugging process of the multimodal fission component could begin before the integration with other components in the GUIDE core. While some of these components will be reused (UIML parsing will still be required, even when the message arrives from the Dialog Manager instead of an XML file) others will not appear in the final version of the multimodal fission component.

Future work will address the development of rules for complementing and adapting to motor and cognitive limitations, but also to deal with the third level of adaptation mentioned before, the Replacement level. This level has been left out of developments in the project so for, given the reluctance shown by application developers in having a ”foreign” framework taking over the rendering of their applications. As such, the project has committed no resources so far to this goal, which is why no references to it have been made in this document. However, it is envisioned that new algorithms will have to devised to that end, since the variables at play are substantially different than the ones for the two first adaptation levels.

Glossary

CCG	Centro de Computação Gráfica
CSS	Cascading Style Sheets
DM	Dialogue Manager
FCUL	Faculdade de Ciências da Universidade de Lisboa
GUI	Graphical User Interface
GUIDE	Gentle User Interfaces for Elderly People
HTML	HyperText Markup Language
ICT	Information and Communication Technologies
UI	User Interface
UIML	User Interface Markup Lanaguage
WBI	Web Browser Interface
XML	eXtensible Markup Language

Bibliography

[1] E Abrahamian, J Weinberg, M Grady, and C Stanton. The effect of personality-aware computer-human interfaces on learning. *Journal Of Universal Computer Science*, 10(1):27–37, 2004.

[2] Elisabeth Andre. Employing ai methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*, 13(4-5):415–448, 1999.

[3] John Bateman, Jorg Kleinz, Thomas Kamps, and Klaus Reichenberger. Towards constructive text, diagram, and layout generation for information presentation. *Comput. Linguist.*, 27:409–449, September 2001.

[4] J P Bigham, C M Prince, and R E Ladner. *WebAnywhere: a screen reader on-the-go*, page 73–82. ACM New York, NY, USA, 2008.

[5] Richard A. Bolt. Put-that-there: Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.*, 14:262–270, July 1980.

[6] Jullien Bouchet and Laurence Nigay. Icare: a component-based approach for the design and development of multimodal interfaces. *Language*, pages 1325–1328, 2004.

[7] C Bregler and Y Konig. Eigenlips for robust speech recognition. *Proc ICASSP*, pages (II–669) – (II–672), 1994.

[8] Trung H Bui. Multimodal dialogue management - state of the art. *Manager*, 2(TR-CTIT-06-01), 2006.

[9] Harry Bunt. Dialogue pragmatics and context specification. In *In Abduction, Belief and Context in Dialogue; studies in computational*, pages 81–150. John Benjamins, 2000.

[10] Stephen M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph.*, 10:111–151, April 1991.

[11] Jose Coelho, Carlos Duarte, Pradipta Biswas, and Pat Langdon. Developing accessible tv applications. *Proceedings of the ACM SIGACCESS Conference in Accessibility (ASSETS)*, 2011.

[12] K Bretonnel Cohen, Giovanni Varile, Antonio Zampolli, Ronald Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue. *Survey of the State of the Art in Human Language Technology*, volume 76. Cambridge University Press and Giardini, 1997.

[13] P R Cohen, Michael Johnston, D McGee, Sharon Oviatt, J Pittman, I Smith, L Chen, and J Clow. *QuickSet: multimodal interaction for distributed applications*, pages 31–40. ACM, 1997.

[14] David Costa and Carlos Duarte. Adapting multimodal fission to user's abilities. *Proceedings of the 14th International Conference on HumanComputer Interaction HCII2*, pages 347–356, 2011.

[15] Joëlle Coutaz, Laurence Nigay, Daniel Salber, Ann Blandford, Jon May, and Richard M Young. Four easy pieces for assessing the usability of multimodal interaction: the care properties. *Proceedings of*, 95(June):1–7, 1995.

[16] Carlos Duarte. Design and evaluation of adaptive multimodal systems. *Inform*, 2008. Phd Thesis.

[17] Carlos Duarte, José Coelho, Pedro Feiteira, David Costa, and Daniel Costa. Eliciting interaction requirements for adaptive multimodal tv based applications. *Proceedings of the 14th International Conference on HumanComputer Interaction HCII*, pages 42–50, 2011.

[18] Carlos Duarte, Daniel Costa, David Costa, and Pedro Feiteira. Support for inferring user abilities for multimodal applications, 2010.

[19] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. *Human Machine Interaction*, 5440(2):3–26, 2009.

[20] V K Emery, P J Edwards, J A Jacko, K P Moloney, L Barnard, T Kongnakorn, F Sainfort, and I U Scott. Toward achieving universal usability for older adults through multimodal feedback. *ACM SIGCAPH Computers and the Physically Handicapped*, 73-74(73-74):46–53, 2003.

[21] Massimo Fasciano and Lapal Guy. Intentions in the coordinated generation of graphics and text from tabular data. *Knowl. Inf. Syst.*, 2:310–339, August 2000.

[22] Steven K. Feiner and Kathleen R. McKeown. *Automating the generation of coordinated multimedia explanations*, pages 117–138. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1993.

[23] Ellen Foster. Interleaved preparation and output in the comic fission module, 2005.

[24] Yi Han and Ingrid Zukerman. A mechanism for multimodal presentation planning based on agent cooperation and negotiation. *Hum.-Comput. Interact.*, 12:187–226, March 1997.

[25] Yi Han and Ingrid Zukerman. A mechanism for multimodal presentation planning based on agent cooperation and negotiation. *Human-Computer Interaction*, 12(1):187–226, 1997.

[26] G. Herzog, E. Andre, S. Baldes, and T. Rist. Combining alternatives in the multimedia presentation of decision support information for real-time control. In *In IFIP Working Group 13.2 Conference: Designing Effective and Usable Multimedia Systems*, 1998.

[27] W Lewis Johnson, Jeff W Rickel, and J C Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11(July 1999):47–78, 2000.

[28] Ju-Hwan Lee and Charles Spence. Feeling what you hear: task-irrelevant sounds modulate tactile perception delivered via a touch screen. *Journal on Multimodal User Interfaces*, 2(3-4):145–156, 2009.

[29] Michael F McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys*, 34(1):90–169, 2002.

[30] Sharon Oviatt. *Mutual disambiguation of recognition errors in a multimodel architecture*, pages 576–583. Number May. ACM Press, 1999.

[31] Sharon Oviatt. The human-computer interaction handbook: Fundamentals, evolving technologies and emerging application. chapter Multimodal interfaces, pages 286–304. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003.

[32] Sharon Oviatt and Philip Cohen. Multimodal interfaces that process what comes naturally. *Information Systems Journal*, 43(3):1–23, 2000.

[33] Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson, and Lesley Carmichael. Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 44–51, New York, NY, USA, 2003. ACM.

[34] Vladimir Pavlovic and Thomas S Huang. Multimodal tracking and classification of audio-visual features. *Proceedings*, 1998.

[35] Norbert Reithinger, Jan Alexandersson, Tilman Becker, Anselm Blocher, Ralf Engel, Markus Lockelt, J Muller, Norbert Pfleger, Peter Poller, Michael Streit, and et al. *SmartKom: adaptive and flexible multimodal access to multiple applications*, pages 101–108. ACM, 2003.

[36] Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. Miamm - a multimodal dialogue system using haptics. In Jan van Kuppevelt, Laila Dybkjaer, and Niels Ole Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.

[37] Cyril Rousseau, Yacine Bellik, and Frédéric Vernier. Multimodal output specification / simulation platform. *Proceedings of the 7th international conference on Multimodal interfaces ICMI 05*, page 84, 2005.

[38] Cyril Rousseau, Yacine Bellik, and Frédéric Vernier. Wwht: un modèle conceptuel pour la présentation multimodale d'information. In *IHM*, volume 264 of *ACM International Conference Proceeding Series*, pages 59–66. ACM, 2005.

[39] Cyril Rousseau, Yacine Bellik, Frédéric Vernier, and Didier Bazalgette. A framework for the intelligent multimodal presentation of information. *Signal Processing*, 86(12):3696 – 3713, 2006. Special Section: Multimodal Human-Computer Interfaces.

[40] Vatikiotis-Bateson E. Rubin, P. and C. Benoit. Audio-visual speech processing (special issue). 1998.

[41] M Schneider-Hufschmidt. Human factors (hf): Multimodal interaction, communication and navigation guidelines. *Proceedings of the 19th International Symposium on Human Factors in Telecommunication*, 1:1–53, 2003.

[42] A Setzer, Y Wilks, and Draft Version. State of the art in dialogue management. *Management*, (September), 2002.

[43] D.G. Stork and M.E. Hennecke. Speechreading by humans and machines. 1996.

[44] David Traum and Staffan Larsson. The information state approach to dialogue management. *Current and New Directions in Discourse and Dialogue*, 4:325–353, 2003.

[45] Wolfgang Wahlster, Elisabeth Andre, Wolfgang Finkler, Hans-Juergen Profitlich, and Thomas Rist. Plan-based integration of natural language and graphics generation. *Artificial Intelligence Special Volume on Natural Language Processing*, 63(RR–93–02):387–427, 1993.

[46] L Wu, S L Oviatt, and Philip R Cohen. Multimodal integration-a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, 1999.

[47] Weiqun Xu, Bo Xu, Taiyi Huang, and Hairong Xia. Bridging the gap between dialogue management and dialogue models. *Discourse*, (July):201–210, 2002.