Adapting Multimodal Fission to User's Abilities

David Costa and Carlos Duarte

LaSIGE and Informatics Department,
Faculty of Sciences of the University of Lisbon,
Campo Grande, 1749-016 Lisboa, Portugal
dcosta@lasige.di.fc.ul.pt, cad@di.fc.ul.pt

Abstract. New ways of communication are now possible thanks to adaptive multimodal systems, enabling the improvement in accessibility of ICT applications to all users. We are developing a project which combines TV with a multimodal system in order to overcome accessibility and usability problems by impaired users. This paper is focused on the fission of outputs, and how the presentations of applications running on GUIDE's environment are adapted to the user's capabilities.

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 to have a natural conversation are not the same as the ones we use during human-computer interaction and the main reason is because 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 (GUI) 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 / 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. 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.

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2011, LNCS 6765, pp. 347–356, 2011.

Our 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 it is necessary to adjust that content for each modality chosen [2].

Multimodality tries to resolve social and e-exclusion 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. [3]

1.1 GUIDE "Gentle User Interface for Elderly People"

This work is being developed in the scope of the European project GUIDE ("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 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 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, as for interacting with electronic devices like 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 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. 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. [4]

2 Multimodal Systems

Multimodal systems are defined in [2] 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" [5].

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 situation and to a context in constant evolution [1].

Systems that combine outputs evolved since the early nineties where text and graphics were combined (e.g. COMET [6]). More recent systems combine speech, haptic, graphics, text, 2D/3D animations or avatars (e.g. SmartKon [7]). Although most applications use few output modalities and consequently straightforward fission techniques, when dealing with above-mentioned combination of outputs it can turn the presentations more complex, difficult to coordinate and make them always coherent. Oviatt and Cohen [8] describes 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.

As [2] 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.

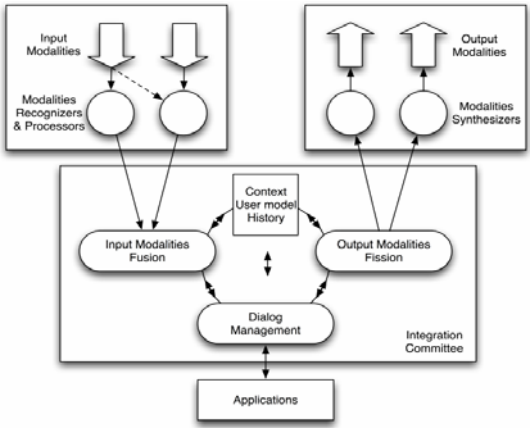

Fig. 1. The architecture of a multimodal system and its generic components [2]

Figure 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 their results of processing. This module is responsible for interpreting the information given by the recognizers and making its decision through fusion mechanisms. With the fusion result computed it is now time to communicate to the Dialogue management that identifies the dialogue state and the action to perform and communicate 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.

2.1 Adaptive Multimodal 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[7]). 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. [9], who created this model to offer the capability to make adaptive and context aware presentations, we will describe the fission module along the following three sections. The WWHT model authors define three main components as being the means of communication (physical and logical) between human and machine. Those components are 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 2.1.1, 2.1.2 and 2.1.3: **What** information to present, **Which** modality(ies) to choose to present that information, **How** to present that information using that modality(ies), **Then** – How to make the presentation change.

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 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 [10]. However, in some systems, selecting and structuring the content is done before the fission module process begins. An example of that is MAGPIE [11].

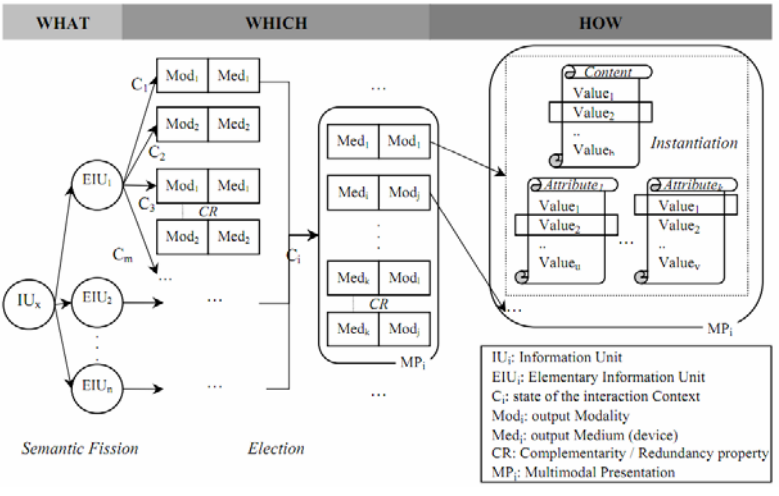

Fig. 2. WWHT conceptual model for context aware and adaptive interaction [9]

In [10] the author describes 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[6] 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 approach uses 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.) [12]. WIP[13]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) 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.

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 ("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. [14, 15]
- **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 are **criterion-referenced** ones, these rules allow a selection of rules based on global criterion (language, age, abilities, etc.) [6, 9, 16].

Coutaz et al [17] 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[17] or Rousseau's platform[18].

Agents - Competitive and cooperative agents plan the presentations. MAGPIE[11]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.

Output Coordination. Once the presentation is allocated, it is now 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 the attributes are fixed such as modality attributes, spatial and temporal parameters, etc.

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. [6, 11]
- **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.
- **Referring expressions –** Some systems will produce multimodal and cross-modal (interaction between two or more sensory modalities) referring expressions, that means making references using multimodalities or referring to another part of the presentation which need some coordination work.[6, 19, 20]

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).

3 GUIDE's Output Fission Module

Due to the already mentioned GUIDE's context environment and objectives these are the following output components that are 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).

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 to 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 GUIDE system. An avatar able to do non-verbal expressions like facial expressions and gestures gives the system a more human like communication. Although a human realistic avatar would be preferable, due to hardware limitations (set-top box) a cartoon-like representation was chosen.

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/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.

Haptic feedback. Haptic output feedback is done using vibration features present on remote control and/or tablet devices. This modality perceived by user's tactile sensory is used to add new or redundant information in complementation of other modalities for example with visual and audio 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.

Fission Architecture. There isn't much research done on fission of output modalities because most application use few different output modalities therefore simple and direct output mechanism are often used. Nevertheless based on 2.1 and on the expected GUIDE applications we will discuss the best implementation of the fission module to be integrated on GUIDE framework.

As mentioned above on section 2.1 this module usually follows three main steps, the message construction, modality selection and output coordination. This module communicates directly with the dialogue manager and not with the application itself.

Taking the example of GUIDE video conferencing, when the application initiates and it is logged with the GUIDE system the application sends to the dialogue manager the information needed to communicate (we assume it is a set of possible states and commands). The Dialogue manager is expected to send information about the content to be displayed (abstract information) and also user information in order to adapt that content to the user's capacities. Although the message construction could be made in the dialogue module as it already has the needed information, decomposition of the semantic information can also be done on the fission module. Between the two possible presentations structures we would choose the plan-based one due to its already explained advantages over schema-based. This structure represents an abstract user interface presentation and it is modality independent. The communication language used between dialogue and fission module should be XML related, however a specific language is yet to be decided (EMMA, SMUIML, UsiXML and TERESA XML are going to be considered).

The selection of a modality to present some specific data is elected using different techniques. Due to the complexity and the hardware limitation imposed by the use of a set-top box we will leave the agent based technique out for now, and we'll test the efficiency of the composition and rules approaches, resulting in a modality dependent structure. If the user has middle hearing, visual and cognitive impairments (common in elderly people), the selection phase will follow the right procedures to adapt the presentation to the user. In this case the sounds (e.g. Text-to-speech of available options description) would be avoided at least if used alone and not as redundant information (e.g. TTS of selected button description). The visual presentation for instance such as text or buttons should be altered thus settings like font size, width and height, back and front colors, contrast and even the location are all susceptible to change. Cognitive impairments are more likely to be a complex matter and more research is indeed needed to know how to treat this problem. Common sense would say to create a simplistic interface with the minimal options and important information required to the user should be always available.

After the selection is done, data need to be sent to their respective renders modules and treated. This communication is also XML based and must coordinate all the modalities in order to maintain the presentation coherency. SMIL, HTIMEL or HTML+TIME are some of the languages to be considered. It is important to refer each state the application goes through, the different presentation screens follows all the fission phases to ensure adaption of environmental changes (user left the front screen, another user appears, loud ambient, etc.)

4 Conclusion

We showed the importance of adaptive output fission to enhance the interaction between humans and computers and how it can overcome human impairments in order to give all users the possibility to interact with any application. We researched some multimodal systems and their techniques for output presentations and described the advantages and disadvantages. We also presented our project named GUIDE and the goals we pretend to achieve, and more specifically the fission module in development. We are working to choose or create the best techniques for the best possible performance due the limited GUIDE hardware restrictions.

References

1. Faraz T.: Multimodal Interfaces, http://www.cs.utoronto.ca/~faraz/projects/6326/paper.pdf
2. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal interfaces: A survey of principles, models and frameworks. In: Lalanne, D., Kohlas, J. (eds.) Human Machine Interaction. LNCS, vol. 5440, pp. 3–26. Springer, Heidelberg (2009)
3. Human Factors (HF): Multimodal interaction, communication and navigation guidelines. In: ETSI 2008 Guides and recommendations to promote e-accessibility, http://www.etsi.org/WebSite/Technologies/HumanFactors.aspx
4. Gentle user interfaces for elderly people, http://www.guide-project.eu/
5. Oviatt, S.: Multimodal interfaces. In: Jacko, J., Sears, A. (eds.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (2003)
6. Feiner, S.K., McKeown, K.R.: Automating the Generation of Coordinated Multimedia Explanations (1991)
7. Reithinger, N., et al.: 5th International Conference on Multimodal Interfaces (2003)
8. Oviatt, S.L., Cohen, P.R.: Multimodal Interfaces That Process What Comes Naturally. Communications of the ACM (2000)
9. Rousseau, C., Bellik, Y., Vernier, F.: WWHT: Un modèle conceptuel pour la presentation multimodale d'information. In: IHM 2005 Proceedings of the 17th International Conference on Francophone sur l'Interaction Homme-Machine (2005)
10. Duarte, C.: Design and Evaluation of Adaptive Multimodal Interfaces, PhD Thesis, Faculty of Sciences, University of Lisbon (2007)
11. Han, Y., Zukerman, I.: A mechanism for multimodal presentation planning based on agent cooperation and negotiation. In: Human-Computer Interaction (1997)
12. Herzog, G., André, E., Baldes, S., Rist, T.: Combining alternatives in the multimedia presentation of decision support information for real-time control. In: Proceedings of the IFIP Working Group 13.2 Conference (1998)
13. Wahlster, W., André, E., Finkler, W., Profitlich, H.-J., Rist, T.: Plan-Based Integration of Natural Language and Graphics Generation. Artificial Intelligence Special Volume on Natural Language Processing (1993)
14. Casner, S.M.: Task-analytic approach to the automated design of graphic presentations. ACM Trans. Graph. (1991)
15. Fasciano, M., Lapalme, G.: Intentions in the coordinated generation of graphics and text from tabular data. Knowledge Information Systems (2000)
16. Bateman, J., Kleinz, J., Kamps, T., Reichenberger, K.: Towards constructive text, diagram, and layout generation for information presentation. Computational Linguistics (2001)
17. Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R.M.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: the CARE Properties. In: INTERACT 1995 (1995)
18. Rousseau, C., Bellik, Y., Vernier, F.: Multimodal output specification / simulation platform. In: ICMI 2005 (2005)
19. André, E., Rist, T., Muller, J.: Employing AI methods to control the behavior of animated interface agents. Applied Artificial Intelligence (1998)
20. Johnson, W.L., Rickel, J.W.: Animated pedagogical agents: Face -to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education (1998)