Eliciting Interaction Requirements for Adaptive Multimodal TV Based Applications

Carlos Duarte, José Coelho, Pedro Feiteira, David Costa, and Daniel Costa

LASIGE, University of Lisbon, Edifício C6, Campo Grande
1749-016 Lisboa, Portugal
{cad,pfeiteira}@di.fc.ul.pt,
{jcoelho,dcosta,thewisher}@lasige.di.fc.ul.pt

Abstract. The design of multimodal adaptive applications should be strongly supported by a user centred methodology. This paper presents an analysis of the results of user trials conducted with a prototype of a multimodal system in order to elicit requirements for multimodal interaction and adaptation mechanisms that are being developed in order to design a framework to support the development of accessible ICT applications. Factors related to visual and audio perception, and motor skills are considered, as well as multimodal integration patterns.

Keywords: Multimodal interaction, Adaptation, User trials.

1 Introduction

TV is one the most ubiquitous devices in our homes. As a result, it has been explored over the years as a means to convey all types of contents, ranging from information to entertainment, to viewers in their houses. Viewing experience has been, as the name implies, passive, with interaction mostly limited to choosing what channel to watch. This trend is now changing, with offers, like Google TV, promising a future where the viewer takes on an active role, and becomes not only a consumer, but also an active content producer. Interaction possibilities will encompass browsing the web, contributing with blogs, publication of different media, like pictures and videos, exploring different applications, including social networks' access, and, keep the more traditional use cases of watching live TV, and recording and watching recorded TV shows.

All this new interaction possibilities will still happen in the same environment where we are used to watch TV. This is mainly the living room, but also includes every other location where people watch TV, like kitchens or their bedrooms. Apart from the location, the interaction means used in these settings will have to move forward from the remote controls that are currently used to interact with a TV and the devices connected to it. Although well suited to switch channels and control the audio volume, they will be inadequate for the majority of the interaction tasks that will become available. While using a remote to select a canned response might still be a possibility, using it for entering a tweet, for instance, will surely discourage users

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2011, LNCS 6765, pp. 42–50, 2011.

from doing it. This means, that a whole new interaction techniques will have to make their way into the living room. Gesture recognition will become part of the interaction experience, which is already happening, first with the Nintendo Wii, and more recently with Microsoft Kinect. But we can expect voice recognition to also become part of this new interactive TV scenario, as well as alternative "remote controls", like smartphones or tablet devices.

Besides this environmental and interaction devices variability, the target audience makes this a truly inclusive problem, since everyone can be a user of such platforms, regardless of their age, knowledge and physical, sensorial or cognitive abilities. Moreover, each user will have his or her preferences. They will be more comfortable with one interaction device than others, but even that may vary depending on the task they are accomplishing.

One approach to tackle to problems raised by all these variables, environment, devices and users, is to employ adaptation techniques. This will allow an intrinsically multimodal system, to explore to the maximum the advantages of multimodal interaction, most notably the possibility of interacting naturally, with the added benefits in terms of learnability and ease of use, and of letting users chose the modality that is the most adequate, based on their appraisal of the current situation [1].

Adaptation will allow a system to exploit the knowledge it might have about its users, in order to provide the most appropriate means of interaction for a given task in a given environment, being aware to the context defining parameters [2]. To support an efficient and effective adaptation in a multimodal setting, it is of the utmost importance to correctly identify the adaptation variables and the interaction patterns users reveal.

This paper reports on the efforts made to tackle the aforementioned issues, in the context of the European Union funded GUIDE[1] project. By employing a user centred approach, interaction requirements are being elicited to understand how user's abilities impact their perception, and also their use of different skills. This allowed for the identification of several interaction requirements, initially targeted at specific modalities, like visual presentation issues. Simultaneously, we have been observing how users integrate multiple modes of interaction when offered the possibility to explore them in a combined or independent fashion.

2 Context

The studies reported in this paper were made in the context of the European Union funded project GUIDE (Gentle User Interfaces for Elderly Citizens). GUIDE aims to provide a framework and toolbox for adaptive, multimodal user interfaces that target the accessibility requirements of elderly users in their home environments, making use of TV set-top boxes as a processing and connectivity platform.

GUIDE envisions bringing to users the benefits of multimodal interaction, empowering interaction through natural and seamless interaction modes, that do not require learning, and that will be able to convince users to adopt them [3]. This includes the modalities that are used in natural human-human communication, as speech and pointing. Additionally, and being based on TV as the central processing

[1] www.guide-project.eu

hub, GUIDE also encompasses remote controls as another interaction modality. Moreover, in an attempt to explore novel interaction scenarios, and even promote mobility in the explored interaction settings, GUIDE includes tablets as an additional device, that can be employed both for input and output modalities. The majority of the system's output will be transmitted, as expected, through the TV set. Complementing it, besides the aforementioned tablet, is the possible of haptic feedback through the remote control.

As can be easily perceived from the above description, GUIDE aims to implement a fully multimodal system, offering its users a range of modalities that they can explore in order to address, both their impairments, as well as the context variations that might come into being. In order to avoid an excessive configuration complexity that such richness might impart, GUIDE includes a set of adaptation capabilities to harness the complexity of managing its full array of modalities. As such, adaptation will impact both multimodal fusion, by taking advantage of known multimodal interaction patters and by adapting individual recognizers in order to increase their efficiency, and multimodal fission, by generating output presentation customized to the system's current user, considering a diversity of factors, with the more important ones being the user's perceptual, motor and cognitive skills and abilities. Adaptation will also impact the dialog management component, thus addressing the interaction flow with applications based on user's abilities.

In order to support such system characteristics it is fundamental to employ a user-centered design methodology. The user is the pivotal subject for the system's adaptation mechanisms. User information drives the adaptation, and knowledge about him is fundamental for the fusion and fission processes and the dialogue management. As a result, a profound characterization of its target users group is one of the project's milestones, and several iterations of user studies are planned in the scope of this project. The next section details the first of these iterations.

3 User Studies

As aforementioned, in order to collect the information required for the characterization of the target population several user studies have been planned. These will proceed in iterative fashion, intermingled with developments, allowing for an updated feedback on evolutions of the envisioned framework and its deployment. In this section we describe the set-up of the first user study conducted in the scope of the project.

Given its nature of being the first study, the possibility of using low fidelity prototypes was considered. This however, was readily discarded, given that we wished to acquire reactions to what technology could offer its users as close to reality as possible. Having already available the majority of the individual technologies envisioned for the project, we opted to build a multimodal system to explore individual reactions to interaction means made available, and to study patters of usage [4]. The recognizers already supported by the system include pointing (either free-hand through a motion sensing system, or with a remote control – in our case, we used the Nintendo's Wii remote control). Speech recognition was not included at this time, but nevertheless tasks accepting speech input were included in the studies, in order to assess how representatives of the target population employ it. Output was achieved through a TV, including text, audio and video. An avatar was also used. With it we attempted to understand if it could contribute the user's empathy towards the system.

The script included tasks to exercise perceptual, motor and cognitive skills. With them we aimed at exploring how users perceive and prefer visual elements (font size, font and background colors, object placement …) and audio stimulus (volume). We included also a set of tasks to exercise the user's cognitive abilities. Furthermore, and event tough 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 an attempt to observe what integration patters were used.

The 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 younger being 66 years old and the older 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.

In the next section we will discuss our findings based mostly on qualitative findings from the trials' observation and the participants' remarks made during and after the trials.

It should be stressed that these were the first trials, and thus have had such a small number of participants. Currently, a new of set of trials is underway, involving several scores of participants, and where findings from the first trials, here reported, have already been used to improve the system's performance.

4 Discussion

As mentioned previously the trial's scrip took in consideration visual, audio and motor requirements. Moreover, in some tasks participants were free to elect how they wished to interact with the trial's 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 favorite 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 any task.

The trials begun with a simple explanation task, where participants were asked to select one of four options presented visually on screen as four buttons. They had to do it through finger pointing, using the 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 of near the 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. In all other, they had to perform a task, be it 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 text that was rendered to them through a TTS system 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) and output (selection made through pointing, speech 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.

In the following paragraphs we report our findings, based on the participants expressed opinions but also on our observations, in situ and of the trials' recordings.

4.1 Visual Perception

Regarding visual presentation there were two main focus studied: target (e.g. buttons) and text.

Targets were analyzed according to their placement, size, separation and content. In what concerns target placement, most participants (6 participants) preferred targets near the center of the screen instead of near the edges. Most participants (7 participants) preferred the larger targets. The majority of participants (6 participants) preferred the version with greater separation than the version with targets closer. Reasons given for this preference include being easier to see (and understand) the targets, additionally to movement related issues (being easier to select). Regarding the content of targets, participants showed a clear preference for solutions that promote a great visual contrast. The more popular choices were white text on black background or blue background, and black text on white background, with a single participant electing blue text on yellow background. There seemed to be some sort of consensus among participants that strong colors are tiring.

In what concerts text presentation, size and color were both evaluated. Six text sizes were presented to participants. The larger was an 100 pixel font (this meant that approximately five words would fill half the TV screen), and the smaller a 12 pixel font. Intermediate sizes were 80, 64, 40 and 24 pixels. Only one participant preferred the larger font. Five participants preferred the second largest and three participants the third largest size. No participant preferred fonts with any of the three smallest sizes.

Table 1. Participants‘ preferences regarding text color (in rows) for different background colors (columns)

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

Text color was evaluated against different color backgrounds. Text colors considered were white, black, blue, red, green, orange, yellow and gray. Background colors used were white, black, blue and green. Participants mostly opted for high contrast combinations. Table 1 shows the participants' expressed preferences, with text color in the rows and background color in the columns. Values in the table's cells represent the number of participants that selected that particular combination of text and background color.

4.2 Audio Perception

The audio tasks evaluated participants' ability to perceive messages in different volumes. Five volumes were employed. The test started on the loudest setting, and then would decrease each time by half the previous volume and after reaching the lowest volume the procedure was repeated but increasing the volume by complementary amounts. Three participants preferred the loudest volume, with all but one participants preferring one of three largest volumes. However, some participants noted that the highest volume was too high in their opinion.

One interesting finding, reported by some participants, was that their comfortable audio level was different when the volume was decreasing or increasing. 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. Other examples of such behavior were observed and reported. In addition to reporting their preferred volume setting, several participants also reported being comfortable with the three loudest volumes.

4.3 Motor Skills

Participants found both free-hand pointing and pointing with the Wii to be natural interaction modalities, meeting our goal of providing interaction modalities that are natural and do not require learning to be used.

When comparing both free-hand pointing and Wii pointing the majority (6 participants) preferred free-hand pointing. Nevertheless, there were some interesting remarks. One participant stated she preferred the Wii remote because it reminded her of her remote control that she is used to handling when watching TV in her house. Other participants changed their preferences during the trial, after becoming more accustomed with both options. These participants typically moved from an initial preference for the remote to a preference for free-hand pointing.

For the design of pointing interaction devices it was possible to gather some relevant indications. For instance, one participant pointed almost exclusively with her finger, barely moving her arm and hand. This was especially challenging for the motion tracking system. Other participants change the pointing hand depending where the target is on screen, using the left hand to point at targets on the left side of the screen, and the right hand to point at targets on the right side of the screen. This knowledge can be used, for instance, to adapt presentation based on the ability of the user's arms.

Participants were also asked to perform a set of representative gestures (e.g. circling, swiping) both in free air and on a surface representative of a tablet. Eight of the participants preferred performing the gestures in free air. Some participants justified this preference because they felt that in that manner they could be more expressive. Others reported that doing gestures in free air was similar to do gestures when engaged in conversation with another person, thus feeling more natural. Another raised the issue that when performing gestures in the tablet she wouldn't be able to see any content that might be displayed in it. Participants were also asked if they preferred to do those gestures with just one or with the two hands. Four participants expressed no preferences in this regard, with only two preferring two handed gestures and 3 preferring one handed gestures.

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 it 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.

4.4 Specific Modality and Multimodal Patterns

One important topic to address when using both gestures and speech is how users combine them, specifically if and what use they make of deictic references.

One initial observation was that the purpose of combining modalities was not clear to all participants. This could however be attributed to their being required to combine modalities to do the same task they had just made with a single modality.

Most participants employed multiple modalities in a redundant fashion, speaking the option's text and pointing to it. From the interaction analysis it was possible to identify some integration patterns. Four participants pointed before speaking, while other four participants spoke before pointing. For the other participant no clear tendency was found. A few participants combined the pointing with spoken deictic expressions. In one of the trials the participant used the deictic reference while pointing, and then followed it speaking the option's text, introducing a different interaction pattern.

Multimodal interaction was explicitly exercised in the context of a selection task done with visual, audio and combined presentations and with selection being possible through pointing, speech and combined usage of these two modalities.

Regarding the presentation, participants unanimously expressed a preference for the system to employ redundant visual and audio presentation (all 9 participants selected this option). Observations of some participants' behavior showed that when the system presented options both visually and aurally, they did not wait for all options to be presented, answering as soon as they perceived the right answer.

Regarding the input, seven participants are satisfied with having simply speech recognition, commenting that it is much easier to perform selection in this fashion. Two participants expressed their preference for a system where they can combine pointing and speaking.

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.

Other interesting observations regarding the use of multiple modalities were made. Some participants, even in tasks where they were required to select only by pointing ended up also speaking the option's text, even without noticing it. In one case, even after being asked to point the correct option, a participant just spoke the option's text.

In what concerns the use of speech alone for selection operations it is important to understand if users simply read aloud one of the choices presented to them or if they use alternative speech. Although most participants simply read aloud one of the presented options, we also witnessed some natural utterances, like saying "the fourth" instead of "option 4" which was the text presented on screen.

4.5 Other Topics

Although not available to be used during the trial, participants were shown a keyboard and a mouse and we asked if they would like to have them available to interact with GUIDE. The majority (6 participants) did not wish to have it available. Some even expressed they could not really understand why should the system employ those devices if users could interact through the natural means they had already tried (speech and gestures), while others simply stated that those interaction means are harder to use than speech or pointing.

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.

5 Conclusions

Due to their intrinsic nature, the design of multimodal interaction systems should be strongly founded on user centred design techniques. That is exactly what is being done in the context of the GUIDE projected, and some of those initial findings have been reported in this paper.

User trials with a multimodal system are being conducted in order to characterize user abilities, and also to assist in the future evolutions of the system. These results are expected to provide impact both on the individual modality level, but also through the multimodal integration patterns found.

This paper presented a summary of the findings, and the impact that user's visual perception, audio perception and motor skills can have on the interaction design of such systems. Additionally, some multimodal interaction patterns were observed and reported.

What is clearly supported by the data acquired so far, is the need for adaptation mechanisms in order to provide adequate interaction mechanisms 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 so far: participants selected their pointing hand based on where the target is on screen. This knowledge can be explored to decide on presentation details. For instance, if we know the user has impairments affecting his left hand, we can present the selectable targets on the right side of the screen, offering him what should be a more comfortable interaction experience.

These user trials are ongoing, so further observations and insights are expected in the near future and will be reported in due time.

References

1. Oviatt, S., Darrell, T., Flickner, M.: Special issue: Multimodal interfaces that flex, adapt, and persist. Commun. ACM 47(1), 1 (2004)
2. Duarte, C., Carriço, L.: A conceptual framework for developing adaptive multimodal applications. In: Proceedings of the 11th International Conference on Intelligent user Interfaces, pp. 132–139. ACM Press, New York (2006)
3. 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)
4. Duarte, C., Feiteira, P., Costa, D., Costa, D.: Support for inferring user abilities for multimodal applications. In: Proceedings of the 4th Conferência Nacional em Interacção Pessoa-Máquina, Aveiro, Portugal (2010)