Readapting multimodal presentations to heterogenous user groups

Antonio Krüger, Michael Kruppa, Christian Müller, and Rainer Wasinger

Saarland University, Dept. of Computer Science
Postfach 151150, 66123 Saarbrücken, Germany
{krueger,mkruppa,cmueller,wasinger}@cs.uni-sb.de

Abstract. This article exploits the possibilities of mixed presentation modes in a situation where both public and private display screens as well as public and private audio channels can be accessed by the users. This will allow the users to share information with a group, while still being able to receive individual information at the same time. Special strategies are identified that readapt an already running public presentation to the interests of late arriving users. Following these strategies, the generation of multimodal presentations for both public and private devices is described.

1 Introduction and motivation

Intelligent computing environments pose new challenges on the design of computer-user interfaces. In such environments, the classical input devices like mouse and keyboard will loose importance. In contrast, more human-like communication methods will play the key role. Interaction between user and computer will include all kinds of modalities such as gestures, speech and haptics. Despite this development, large screens that may have to be shared with other users will still be used to display huge amounts of text and graphics, whereas small and portable screens will be used for the presentation of more private information. This will raise the need for multimodal interfaces that coherently mix audio and visual information in presentations and create user dialogues tailored for hybrid devices.

This article exploits the possibilities of mixed presentation modes in a situation where both public and private screens (e.g. small PDA and large wall-based displays) as well as public and private audio channels (e.g. loudspeakers and headphones) can be accessed by the users. This will allow the users to share information with a group, while still being able to receive private and individual information at the same time.

We will focus on the concerns regarding the multimodal generation of text and audio for these situations, as well as analyse the question on how to combine public and private audio presentations that run at the same time. This is realised by exploiting the so-called *cocktail-party-effect* [1] that allows users to partially focus on different audio sources. Finally, we will explain how to adapt presentations for heterogenous user groups, allowing the system to react to changing

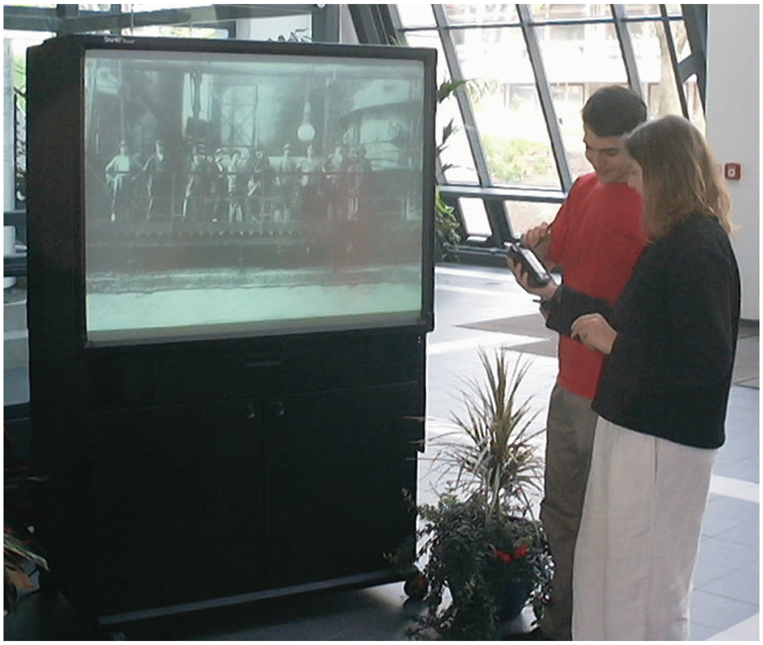

Fig. 1. Visitors in front of the virtual window and equipped with handhelds.

interests in the group. Special focus will be placed on users leaving and joining an already started presentation. A prototype implementation of an intelligent museum guide for the UNESCO world-heritage site *Old Völklingen Iron Works* in Germany, will be used throughout the paper to illustrate the usefulness of our ideas.

2 The museum scenario

This work is motivated by the PEACH project, (carried out together with IRST[1] and DFKI[2]) which aims to improve the fruition of cultural heritage sites, through the help of new computer technology. The overall goal is to enhance the visitor's personal experience when visiting a museum. For this purpose, we are closely collaborating with several museums, one of which is the UNESCO world cultural heritage site *Old Völklingen Iron Works*, an old steel manufacturing factory that after having been closed 16 years ago, has now been turned into a technical museum. Most of the exhibits have a technical nature, for example the huge engines that produced steam and that were needed for the steel melting process. The museum has installed the technical infrastructure (i.e. more then 100 IR[3] bea-

[1] Institute of Research and Technology, Trento, Italy

[2] German Research Centre for Artificial Intelligence, Saarbrücken, Germany

[3] Infra Red

cons) to enable us to test our results on a larger scale. In addition, the museum has provided us with graphics and text material on the exhibits. This allows each visitor to use a PDA that provides them with personalised information at defined locations. At some places in the museum, large information screens (*virtual windows*) with loudspeakers will be used to convey additional information. This setup is illustrated in Figure 1.

For the sake of simplicity, we assume that the visitors can be categorised into two distinct groups: visitors that are interested in technical aspects of the exhibits, and visitors that want to know about social-historical impacts of the steel factory. At the moment, this categorisation is both fixed and selected in advance for each visitor. In the future we plan to automatically categorise visitors by monitoring their behaviour. Furthermore, we assume that visitors either come on their own, or are members of a group (e.g. a family) and thus more willing to share a common presentation. The presentation on the virtual windows is adapted for the users standing in front of it. This adaption takes into account the different user interests (technical or social) and the structure of the user group (i.e. if there are users from different parties). This distinction is important in deciding how to distribute information to the virtual windows and the handheld devices.

3 Related work

A lot of research has been conducted in the area of intelligent museum and tourist guides. The most closely related project to this work is the HIPS project [2]. HIPS is a location-adaptive museum guide based on a subnotebook that provides information on the exhibits to a visitor of a museum. The notebook receives its position from IR beacons distributed throughout the environment, and then uses this position as a pointer to specific content stored in a multimedia database. Animations and audio files are conveyed to the device over a wireless link. HIPS concentrates on creating individual presentations by modelling each user's interests separately. In contrast to our own work, user groups and combined private and public presentations are not considered.

Mobile tourist guides (e.g. [3–5]) provide users with location-based information and use similar approaches to detect the user's position to provide them with appropriate multimedia content. In this context, only little work has been done to exploit the benefits for tourist groups. None of the systems have looked at mixed public and private presentations.

The area of automated generation of multimodal and coherent presentations has also received a lot of attention. One prominent class of these systems is able to generate technical documentations (e.g., WIP [6] and IBIS [7]) that are tailored to the individual user and situation. A feature of these systems is that everything is designed from scratch and no predefined multimedia material is used. This makes it necessary to use rich knowledge bases, that represent relevant tasks, the actual user's interest, and the design rules that are needed to render a presentation. The research behind these systems combines automated natural language and computer graphics generation to yield effective multimodal presentations. In regards to the planning of multimodal presentations, the work described in this paper is more similar to the EMBASSI project [8].

The support for hybrid devices is another important area of ongoing research in ubiquitous computing. Most of this work has been focusing on the technical aspects of bringing together large and small screen devices. In [9], a framework is described that enables users of small screen devices to take advantage of the large screens in their environment. The authors concentrate on how different services may be distributed on different devices, but they do not try to plan and tailor the content according to the users interests and technical configurations.

4 Readaptation strategies

As mentioned in section 2, some assumptions are made to substantiate the scenario. We assume that the visitors can be interested in either of two areas, being the social-historical aspects of the exhibits, or the technical aspects of the exhibits. Furthermore, we assume that there are two disjunct groups of visitors and that this information on the visitors' interests is given. Furthermore we distinguish between A-type visitors (those who came first) and B-type visitors (those who joined the presentation).

In our simplified scenario, the multimodality of the discourse consists of spoken explanations combined with images. Each content item is labelled with a category (general, social, technical) and the relevance (high, medium, low). The notion of relevance, although not discussed in detail, expresses how interesting, important, informative, or entertaining an item is. The images for each item are stored in an ordered list, where the first element represents the best fit. When the discourse is planned, the items can be presented in a long (default) or a short version, or they can be left out completely. The utterances are hardwired and represented as strings. Coherence in the discourse is achieved at the expense of redundancies introduced from connecting items together.

There are two audio channels for speech: a private channel (a one-ear headphone) and a public channel (loudspeakers). The images can be displayed on a private display (heldheld device) or a public display (big screen), and the public display can be split to show several images at the same time.

We have identified four different strategies that adapt content to specific user situations:

1. **No readaptation**
   When people with the same interests share the display, the presentation does not have to be readapted. In this situation one or more A-type visitors of a specific group are listening to a presentation and some B-type visitors with the same interests (e.g. technical aspects) join the presentation. In this case, it does not matter if the groups are different, and there is no need to change the presentation. However, it may be necessary to provide them with a summary of what has already been said.

2. **The conservative readaptation strategy**
   This is a situation where one or more visitors with the same profile are using a public display and an equivalent number of visitors (or less), belonging to a group with a different interest join them. When readapting the presentation to the B-type visitors, a conservative strategy should be chosen. Conservative means that the original presentation remains unchanged. The B-type visitors then retrieve additional information in the form of annotations on their private devices. An analogy to this is a private human guide that accompanies the visitor through the museum. For example, on arriving at a presentation, such a private quide (standing next to the visitor) would quietly comment on various aspects of the presentation. A private device with display can go beyond this by also showing pictures.
3. **The progressive readaptation strategy**
   When the constraint of treating the A-type visitors as the main group is relaxed, the readaptation strategy may be more progressive. This makes sense, when the number of B-type visitors is larger than the number of A-type visitors and if they belong to different groups and have different interests. With a progressive strategy, the original presentation may be changed in terms of content (speech and pictures), length, and presentation style.
4. **The merging strategy**
   In some situations, it might be adequate to readapt the ongoing presentation without using a second channel. This is especially interesting, if B-type visitors are members of the same group as the A-type visitors, but have different interests. In this case, the different presentation plans can be merged into one, while abiding to some constraints like the limitations on total time or the relevance of the content that concern the underlying presentation goals.

5 Representing content and planning presentations

Multimedia material is stored in the form of images and text, although videos and audio clips may also be considered in future work. The images required by a particular presentation are sent over a wireless network connection before the client's presentation begins. To increase the response time of the resource-limited client PDAs, the images are stored in the database in varying pixel sizes, for example 320x240 and 1024x768. This removes the need for the client to expend resources on first resizing the images to suite its own screen size. It also allows greater flexibility in the type of clients that can connect to a virtual window.

The multimedia material for each dialog component is defined in an XML content description file together with additional information such as classification and relevance.

The static dialog/text components will hopefully be generated in the future by a text generation engine. This component would have the task of creating the dialog components in a various number of natural sounding length formats. These dialogs would also be taken from a much larger source of data, for example an electronic book covering everything that the museum has to offer.

The planning process itself has to select and then distribute audio and graphical material on the virtual window and the PDA. To explain our approach let's look how the conservative readaptation strategy (i.e. the virtual window presentation remains fixed) is realised. Therefore let's assume that the presentation consists of three major blocks with several dialog components: $g_1, ...g_n, t_1, .., t_m, g_n + 1, ..., g_k$, where g_i denotes the i-th general dialog component and t_j the j-th technical dialog component. The overall time of the technical part of the presentation $T(t1, .., t_m)$ is derived by counting the syllables, words and phrases. The goal now is to find a private audio sequence $s_1, ..., s_l$ of social dialogue components with $T(s_1, ..., s_m) = T(t_1, ...t_m) + \delta$, where δ denotes a positive or negative time span. This time span has to be relatively small (2s), allowing for a precise temporal alignment of private and public audio by increasing or reducing pauses between the components. If the user is not looking at the device, a meta comment is added that informs them to do so. From the set of possible solutions, the one with maximum relevance is selected. This is currently done by simply summing all relevance factors of the sequence. In the next step of the implementation, we also plan to take into account the ordering of the sequences. The planning problem can thus be formulated as a constraints problem, including time constraints that create presentations which minimise waiting times between presentations, and relevance constraints that maximise the relevancy of all presentations. Currently we us the OZ constraint solver [10] to find a (optimal) solution.

With the conservative readaptation strategy, selecting an appropriate private graphic is easy because it is sufficient to select the most relevant image from each dialog component. However, other strategies, for example the merging strategy, ask for a more complicated selection mechanism since more than one user interest has to be considered.

6 System architecture and technical setup

This section first describes the major components of the system and then focuses on some of the technical details of the prototype implementation. As shown in Figure 2, the current implementation is based on a client-server architecture. On initialisation, the server reads in all of the dialog components from the XML content description file as described in the previous section.

As visitors begin to interact with the system (i.e. when they appear in front of the virtual window), their specific user parameters are communicated and stored on the server and used to create a presentation based on the readaptation strategies. These parameters include the visitor's interest, the group that the visitor belongs to, their arrival time and some technical details regarding the specifications of the client device. The presentation planner then chooses a suitable adaptation strategy and creates a presentation based on the constraints provided in the constraints database. The resulting presentation is then displayed to the visitor on either the public channel, private channel or both channels. In the case that the visitor is the first to arrive at the virtual window, the presen-

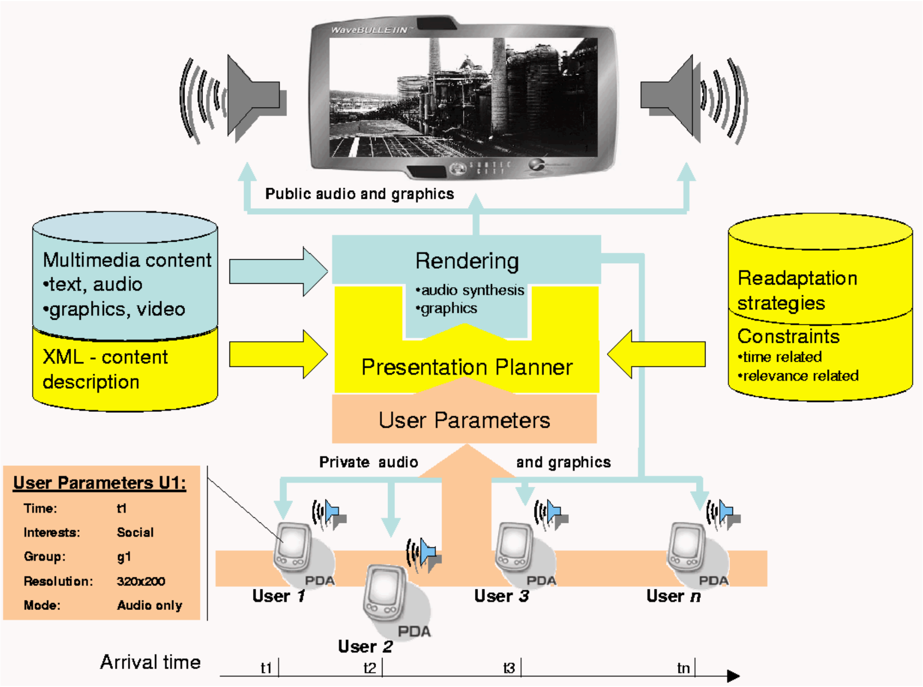

Fig. 2. Components required for the public/private media system.

tation requires no readaptation strategy and is therefore simply loaded onto the public channel in the form of audio and graphics.

Tests on the current system show that the timing involved in coordinating the presentations for both the client and server are fairly accurate. Along with the dialog length calculations, it is also possible to adjust the synthesizer speaking rates to fine tune the system.

The stationary system (virtual window) consists of a standard desktop PC with sufficient multimedia capabilities to playback both audio and video presentations. As a combined output/input device, a smartboard is also used, and speakers connected to the PC provide public speech output.

The mobile part of the system is comprised of several Compaq iPAQ client devices, each equipped with a PCMCIA Wavelan card and a 512 Megabyte SDRAM card. The originally installed Windows CE operating system was replaced by the familiar linux distribution [11]. The Java(TM) 2 Runtime Environment, Standard Edition, Version 1.3.1 was installed on the SDRAM card. Users carry the iPAQ around in their hands or may leave it hanging on a cord around their neck. We assume that the user is willing to look at the device (or already looking at it) when in their hands and that they prefer speech-only output when the device is hanging around their neck. This allows the system to distinguish between a *graphics-audio* and an *audio-only* mode.

Two audio synthesis engines are currently being used, one for the server and one that runs on each of the clients. The server is currently using a standard desktop speech synthesizer, IBM ViaVoice, which implements a female voice. The clients are running Festival Lite (Flite) from CMU, which uses a male voice. Flite is particularly good because it is fast and cheap (free). An alternative implementation could have been to use FreeTTS, which is completely Java based and equally cheap.

7 Summary and future work

Presented in this paper are the first steps towards the generation of multimodal presentations for hybrid devices and multiple users. For this purpose, we mix public and private graphics and audio streams to tailor the system's output to the interests of users standing in front of a public screen. The generation process takes into account the structure of the group, especially if users join or leave a running presentation. The planning process itself can still be considered fairly simple. It recombines pieces of multmedia material to fulfill certain constraints. In order to improve the system's output, we plan to investigate the use of text and graphics generation methods similiar to those used by the multimodal presentation systems described in section 3.

Furthermore, we plan to allow the user to interact with the system, preferably with speech input. This will require an even more advanced planning process that plans not only the presentations, but the whole dialogue situation. Provided that full text generation and dialog planning is incorporated, the mixing of private and public audio could be further exploited, for example, one could provide users with text summaries that come at the end of the presentation, or give them meta comments that indicate what will follow in the next few minutes of the presentation.

To enhance the graphical quality of our presentations we are already experimenting with a virtual character (similiar to the one used in [12]) that guides the user through private and public presentations. We believe that a presentation agent capable of "jumping" from a private to a public screen and vice versa would help to guide the user's attention to relevant presentation parts and would thus help to improve the coherence of the overall presentation.

Finally, the use of a better text-to-speech component that produces clearly distinguishable voices would help to exploit the cocktail-party-effect to a greater extent.

8 Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft, the German Ministry for Education and Research, and the Autonomous Province of Trento, Italy. We would like to thank the UNESCO-world heritage site Old

Völklingen Iron Works, Germany for the multimedia material and for being able to test our system in the environment of the museum. Credits also go to Thorsten Bohnenberger who helped with the constraint programming, and Eyeled for the use of their IR-beacons.

References

1. Arons, B.: A review of the cocktail party effect. Journal of the American Voice I/O Society (1992)
2. Not, E., Petrelli, D., Sarini, M., Stock, O., Strapparava, C., Zancanaro, M.: Hypernavigation in the physical space: Adapting presentations to the user and to the situational context. The New Review of Hypermedia and Multimedia 4 (1998) 33–45
3. Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C.: Developing a Context-aware Electronic Tourist Guide: Some Issues and Experiences. In: Proceedings of CHI 2000. (2000) 17–24
4. Long, S., Kooper, K., Abowd, G.D., Atkeson, C.G.: Rapid Prototyping of Mobile Context-Aware Applications: The Cyberguide Case Study. In: Proceedings of the 2nd ACM International Conference on Mobile Computing and Networking. (1996) 97–107
5. Malaka, R., Zipf, A.: Deep Map - challenging IT research in the framework of a tourist information system. In Fesenmaier, D.R., Klein, S., Buhalis, D., eds.: Information and communication technologies in tourism 2000. Springer, Wien (2000) 15–27
6. Wahlster, W., André, E., Bandyoppadhyay, E., Graf, W., Rist, T.: Wip: The coordinated generation of multimodal presentations from a common representation. In Ortony, A., Slack, J., Stock, O., eds.: Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues. Springer-Verlag, Berlin (1992) 121–144
7. Feiner, S., Seligmann, D.: Automated generation of intent-based 3D-illustrations. Computer Graphics (1991)
8. Elting, C., Michelitsch, G.: A multimodal presentation planner for a home entertainment environment. In: ACM Proceedings of the PUI2001, Orlando Florida USA (2001)
9. Pham, T.L., Schneider, G., Goose, S.: A situated computing framework for mobile and ubiquitous multimedia access using small screen and composite devices. In: Proc. of the ACM International Conference on Multimedia. (2000)
10. Smolka, G., Henz, M., Würtz, J.: Object-oriented concurrent constraint programming in Oz. In van Hentenryck, P., Saraswat, V., eds.: Principles and Practice of Constraint Programming. The MIT Press (1995) 29–48
11. Compaq: The familiar linux distribution. (2002) url: http://www.handhelds.org.
12. André, E., Müller, J., Rist, T.: Wip/ppp: Automatic generation of personalized multimedia presentations. ACM Multimedia (1996) 407–408