Chapter 7
Two Frameworks for the Adaptive Multimodal Presentation of Information

Yacine Bellik
Université d'Orsay, Paris-Sud, France

Christophe Jacquet
SUPELEC, France

Cyril Rousseau
Université d'Orsay, Paris-Sud, France

ABSTRACT

Our work aims at developing models and software tools that can exploit intelligently all modalities available to the system at a given moment, in order to communicate information to the user. In this chapter, we present the outcome of two research projects addressing this problem in two different areas: the first one is relative to the contextual presentation of information in a "classical" interaction situation, while the second one deals with the opportunistic presentation of information in an ambient environment. The first research work described in this chapter proposes a conceptual model for intelligent multimodal presentation of information. This model called WWHT is based on four concepts: "What," "Which," "How," and "Then." The first three concepts are about the initial presentation design while the last concept is relative to the presentation evolution. On the basis of this model, we present the ELOQUENCE software platform for the specification, the simulation and the execution of output multimodal systems. The second research work deals with the design of multimodal information systems in the framework of ambient intelligence. We propose an ubiquitous information system that is capable of providing personalized information to mobile users. Furthermore, we focus on multimodal information presentation. The proposed system architecture is based on KUP, an alternative to traditional software architecture models for human-computer interaction. The KUP model takes three logical entities into

DOI: 10.4018/978-1-60566-978-6.ch007

account: Knowledge, Users, and Presentation devices. It is accompanied by algorithms for choosing and instantiating dynamically interaction modalities. The model and the algorithms have been implemented within a platform called PRIAM (PResentation of Information in AMbient environment), with which we have performed experiments in pseudo-real scale. After comparing the results of both projects, we define the characteristics of an ideal multimodal output system and discuss some perspectives relative to the intelligent multimodal presentation of information.

INTRODUCTION

For a few years, access to computers has become possible to a large variety of users (kids, adolescents, adults, seniors, novices, experts, disabled people, etc.). At the same time, advances in the miniaturization of electronic components have allowed the development of a large variety of portable devices (laptops, mobile phones, portable media players, personnel digital assistants (PDA), etc.). New interaction situations have started to appear due to users' mobility enabled by this evolution. It is nowadays commonplace to make a phone call on the street, to work while commuting in public transportation, or to read e-mails at a fast-food. The interaction environment which was static and closed has become open and dynamic. This variety of users, systems and physical environments leads to a more complex interaction context. The interface has to adapt itself to preserve its utility and usability. Our work aims at exploiting the interaction richness allowed by multimodality as a means to adapt the interface to new interaction contexts. More precisely we focus on the output side of the interface. Our objective is to exploit intelligently all modalities available to the system at a given moment, to communicate information to the user. In this chapter, we start by presenting related work. Then we present a first framework which addresses the problem in a "classical" interaction situation. A second framework addresses the same problem in a different situation: ambient environments. After comparing the results of both projects we conclude by presenting some future research directions.

RELATED WORK

At first, multimodality was explored from the input side (user to system). The first multimodal interface was developed in 1980 by Richard Bolt (Bolt, 1980). He introduced the famous "*Put That There*" paradigm which showed some of the power of multimodal interaction. Research work on output multimodality is more recent (Elting, 2001-2003). Hence, the contextualization of interaction requires new concepts and new mechanisms to build multimodal presentations well adapted to the user, the system and the environment.

Output Multimodality Concepts

Presentation Means

When designing presentation as an output of a system, one has to choose which modalities will be used, and how they will convey information. The concept of *presentation means* represents the physical or logical system communication capacities. There are three types of presentation means: mode, modality and medium. Depending on authors, these three terms may have different meanings (Frohlich, 1991; Bernsen, 1994; Nigay, 1995; Bordegoni, 1997; Martin, 1998). In our case we adopt user-oriented definitions (Bellik, 1995; Teil, 2000). A mode refers to the human sensory system used to perceive a given presentation[1] (visual, auditory, tactile, etc.). A modality is defined by the information structure that is perceived by the user (text, ring, vibration, etc.) and not the structure used by the system[2]. Finally, a medium

is a physical device which supports the expression of a modality (screen, loudspeakers, etc.). These three presentation means are dependent. A set of modalities may be associated with a given mode and a set of media may be associated with a given modality. For instance, the "Vibration" modality can be expressed through the "Vibrator" medium and invokes the "Tactile" mode.

Interaction Context

An interaction occurs in a given *context*, although the definition for what *context* means may vary depending on the research community. In our research work we adopt Dey's definition (Dey, 2000). The interaction context is considered as any information relative to a person, a place or an object considered as relevant for the interaction between the user and the system. We use a model-based approach (Arens, 1995) to specify the elements of interaction context (system model, user model, environment model, etc.). A set of dynamic or static criteria is associated to each model (media availability, user preferences, noise level, etc.). The work presented here does not propose new ways of capturing context. Instead, we suppose that we can rely on an adequate framework, as those proposed in Dey (2000) or Coutaz (2002).

Multimodal Presentation

When one uses several modalities to convey information, the presentation of information is said to be *multimodal*. A multimodal presentation is comprised of a set of (modality, medium) pairs linked by redundancy or complementarity relations according to CARE properties (Coutaz, 1995). For instance, an incoming call on a mobile phone may be expressed by a multimodal presentation composed of two (modality, medium) pairs: a first pair (Ring, Loudspeaker) indicates the call receipt while a second pair (Text, Screen) presents the caller identity (name).

Output Multimodal Models and Systems

SRM (Standard Reference Model) (Bordegoni, 1997) is one of the first conceptual models which addressed the problem of multimodal presentation. Stephanidis (Stephanidis, 1997) improved it by integrating the interaction context within the initial design of the multimodal presentation, even though this integration was incomplete. Then, Thevenin introduced the concept of plasticity (Thevenin 1999) to describe the adaptation of interfaces. At first, this concept of plasticity addressed the interface adaptation in regard to the system and environment only, while preserving interface usability. Later, it has been extended to the <user, system, environment> triplet designing the general interaction context (Calvary, 2002) (Demeure, 2003). The concept of plasticity inspired CAMELEON-RT (Balme, 2004) which is an architecture reference model that can be used to compare existing systems as well as for developing run time infrastructures for distributed, migratable, and plastic user interfaces.

Actually, we can notice that existing systems have often addressed the problem under a specific angle (Table 1). For instance, WIP (André, 1993) explored the problem of coordinating text and graphics. This system is capable of automatically generating from text and graphics user manuals for common devices. The COMET system (Feiner, 1993) also addresses the same problem in a different application domain (diagnostic, repair and maintenance). While both systems addressed the problem of coordinating visual modalities, MAGIC (Dalal, 1996) explored the coordination of visual and audio modalities. AIFresco System (Stock, 1993) addressed the problem of natural language generation in the context of an hypermedia system. PostGraphe (Fasciano, 1996) and SAGE (Kerpedjiev, 1997) have a common approach which consists in generating multimodal presentation based on the concept of presentation goal. CICERO (Arens, 1995) introduced an ap-

Table 1. Problems addressed by some existing systems. © Yacine Bellik. Used with permission.

Systems	Addressed Problems
WIP (1993), COMET (1993), AlFresco (1993)	Visual modalities coordination
MAGIC (1997)	Visual and audio modalities coordination
AlFresco (1993)	Natural language generation
CICERO (1995)	Models management
AVANTI (2001)	User model management
PostGraphe (1996), SAGE (1997)	Presentation goals management

proach based on models (media, information, task, discourse and user). AVANTI (Stephanidis, 2001) is one of the first systems, which takes into account the interaction context even though it is mainly based on user profiles. Table 1 synthesizes the contribution of these different systems.

Towards Mobile Environments: Ambient Intelligence

For a few years, computer technology has been pervading larger parts of our everyday environment. First it spread in "technological" artifacts such as cameras, mobile phones, car radios, etc. Now, researchers consider its integration into even more commonplace objects such as clothing, doors, walls, furniture, etc. This trend is referred to by terms like *pervasive* or *ubiquitous computing*, the *disappearing computer*, *mixed systems*, *ambient intelligence*, etc. All of them describe the same kind of concept, that of giving everyday objects additional capabilities in terms of computation, wireless communication and interaction with human users.

Although the basic concept dates back to the early 1990s (Weiser, 1993), its implementation was long deemed impractical because electronic devices could not be miniaturized enough. However, recent advances in the fields of miniaturization, wireless networks and interaction techniques are quickly removing these technical barriers. Moreover, until recently, researchers in the field had to master both hardware and software, which limited the development of these systems. Now, off-the-shelf hardware platforms are readily available (Gellersen, 2004), thus software specialists can experiment with ambient systems without being hardware experts.

In consequence, more and more research groups are getting involved in the domain, and some people even think that we are at the threshold of a revolution similar to that of the 1980s when computers broke out of datacenters and spread in office environments and later at home (Lafuente-Rojo, 2007).

In 2001 the European Information Society Technologies Advisory Group (ISTAG) tried to characterize the specificities of ambient intelligence (Ducatel, 2001). It came out with three core properties:

- ***Ubiquitous computing***: microprocessors can be embedded into everyday objects that traditionally lack any computing ability, such as furniture, clothing, wallpaper, etc. Some people already envision embedding RFID[3] into construction materials (concrete, paint) or furniture (Bohn 2004).
- ***Ubiquitous communication***: these objects must be endowed with wireless communication abilities, rely on energy sources that provide them with good autonomy, and be capable of spontaneously interoperating with other objects, and without human intervention.
- ***Intelligent user interfaces***: human users must be able to interact with these objects in a natural (using voice, gestures, etc.) and

customized way. The object must therefore take user preferences and context into account.

Ambient intelligence systems interact with users when they are not in "classical" interaction situations, i.e. sitting at one's desk or using a portable device (PDA for instance). These systems must be able to discretely and non-intrusively react to user actions. This can be significantly useful in mobile situations, when it is impractical to use a device, even a handheld one such as a mobile phone. For instance, when one is finding their way at an airport, they generally do not want to hold their mobile phone, it is much more appropriate to receive information from the background, for instance from loudspeakers or stationary display screens.

MOTIVATION FOR OUR RESEARCH WORK

Our purpose in executing the research work described in this chapter was to investigate how one can design frameworks and algorithms for mobile-based multimodal presentation of information. The work was split around two categories of systems: first those in which the user is fixed with respect to the system, but both may be mobile with respect to the environment (such as a mobile car system), and second those in which the user is mobile with respect to the system itself.

The first research work addresses the issue of creating a multimodal presentation of information that takes the current context into account. It introduces a framework that is capable of choosing a combination of modalities for a given information item (in a redundant or complementary fashion according to CARE properties). The system and the user are fixed with respect to one another. They can both be stationary (e.g. a user using their desktop telephone), or both be moving together (e.g. a traveler and their mobile phone).

In contrast, the second research work focuses on the new problems introduced by the mobility that may exist between the user and the system (like in ambient environments). It does not focus on combining modalities, but rather on using a variety of devices to provide a user with information. These devices may be stationary (fixed display screens, loudspeakers in an airport), or mobile (handheld devices), though the emphasis will be on the former. Only one modality is used at a time, but the proposed framework is nevertheless multimodal because possible modalities depend on the user and on the devices considered (exclusive multimodality (Teil, 2000)). This framework provides a mechanism for choosing a device and a modality for presenting an information item for a mobile user.

FIRST FRAMEWORK: CONTEXTUAL PRESENTATION OF INFORMATION

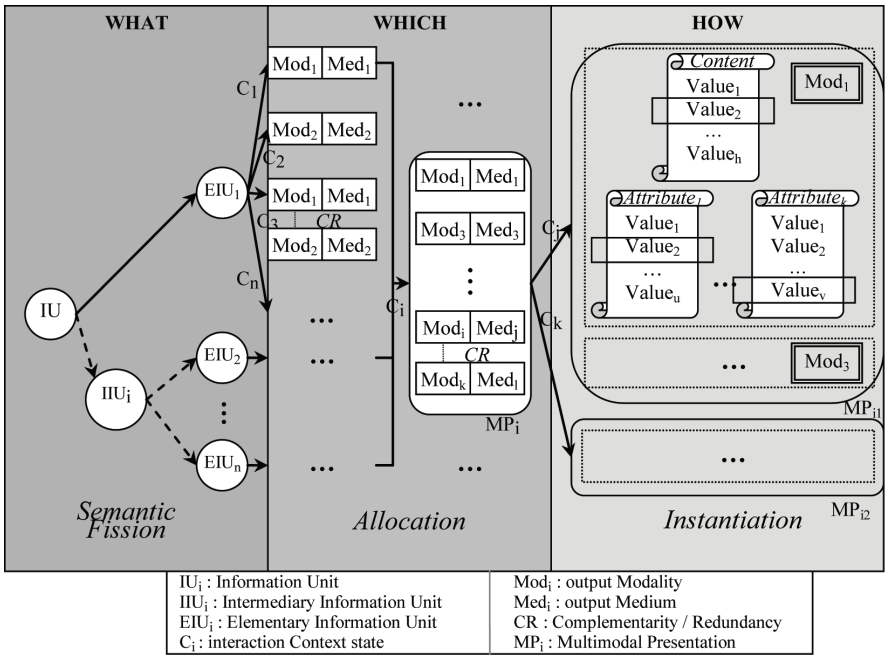
This research work aims at proposing an abstract model which enables one to organize and to structure the design process of a dynamic and contextual multimodal presentation. This model called WWHT (What, Which, How, Then) was implemented in a software platform (ELOQUENCE) which includes a set of tools supporting the designer/developer during the process of elaborating multimodal presentations. We have used this platform to develop two applications: a fighter cockpit simulator and an air traffic control simulator.

The WWHT Model Components

The WWHT model is based on four main components: information to present, presentation means, interaction context and the resulting multimodal presentation.

The information represents the semantic object the system has to present to the user. For instance, in mobile telephony, the receipt of a new call

Figure 1. Design process of a multimodal presentation adapted to interaction context. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



constitutes semantic information that the output multimodal system has to express. The presentation means, interaction context and multimodal presentation are defined as above.

The WWHT Model

The WWHT model is structured around answering four main questions:

- *What*: what is the information to present?
- *Which*: which modality (ies) should we use to present this information?
- *How*: how to present the information using this (ese) modality (ies)?
- *Then*: how to handle the evolution of the resulting presentation?

The first three questions (*What*, *Which* and *How*) refer to the initial building of a multimodal presentation while the last one (*Then*) refers to its future. Figure 1 presents the process of the initial design. The presentation evolution is described in the sub-section entitled "Then".

Further questions could have been asked such as: "*When*", "*By Whom*", "*Where*". However we limited the model to the questions that are directly relevant for a multimodal presentation module. For instance, in our software architecture, it is not the responsibility of the presentation module to decide *when* to present information. The presentation module just waits for information to present sent by the dialog manager and when the dialog manager invokes it, it presents the information. The answer to the question "*by whom* is the presentation done and decided?" is: the system (at runtime). However the system simply complies to design rules introduced by the designer. So at last, it is the designer who determines the behavior of the system. Finally, the question "*Where*" is implicitly included inside the "*Which*" and "*How*" questions since the choice of modalities, medias and their attributes will determine the location of the presentation.

What

The starting point of the WWHT model is the semantic information (Figure 1, IU) the system has to present to the user. To reduce the complexity of the problem, we start by decomposing the initial semantic information into elementary information units[4] (Figure 1, EIUi). For instance, in the case of a phone call, the information "Call from X" may be decomposed into two elementary information units: the call event and the caller identity.

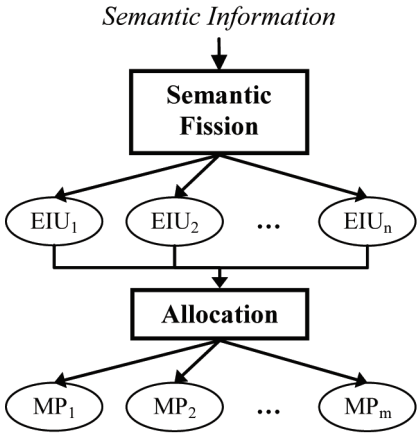
Let us underline here the fact that the term "*fission*" (Wahlster, 2003) (Nigay, 1993) is often used by opposition to the term "*fusion*" (for input multimodal interaction) to qualify the whole building process of a multimodal presentation (Figure 2). We prefer to talk about fission only during the first step of the process, which consists in splitting the initial semantic information into several elementary information units. Since the entry point of this process is semantic information, we prefer to call it semantic fission rather than multimodality fission or modality fission. The next step which consists in choosing *a modality or a combination of modalities* for each elementary information unit is then called *allocation* and is detailed in the next section.

In general, the semantic fission is done manually by the designer because an automatic semantic fission requires semantic analysis mechanisms which make the problem even harder. However, it constitutes an interesting topic for future work.

Which

When the decomposition is done, a presentation has to be allocated to the information. The allocation process consists in selecting for each elementary information unit a multimodal presentation (Figure 1, [Modi, Medj]) adapted to the current state of interaction context (Figure 1, Ci). The resulting presentation is comprised of a set of pairs (modality, medium) linked by redundancy/complementarity relations. This process may be

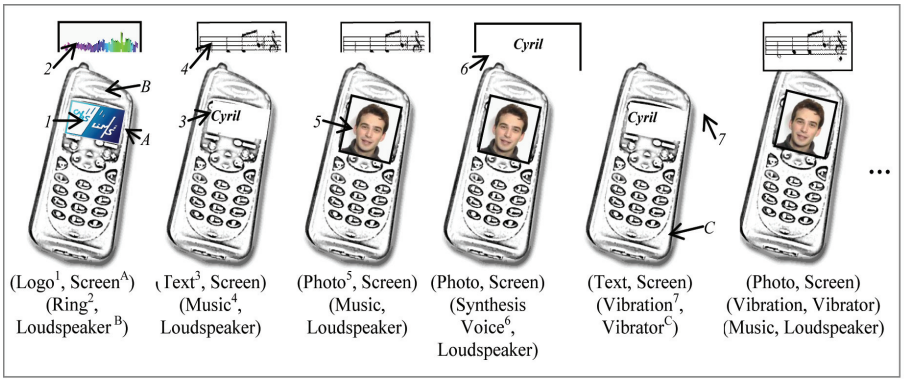
Figure 2. Place of semantic fission within the building process of a multimodal presentation. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



complex in particular in the case of applications with several communication modalities and/or applications with a high variable interaction context. Figure 3 presents examples of possible multimodal presentations to express the semantic information "Call from X" on a mobile phone.

The selection process of presentation means is based on the use of a behavioral model. The representation of this behavioral model may vary depending on the system considered: rules (Stephanidis, 1997), matrices (Duarte, 2006), automata (Johnston, 2005), Petri Nets (Navarre, 2005), etc. In the ELOQUENCE platform we have used a rule-based representation. This representation allows an intuitive design process (If … Then…instructions). However this choice introduces problems on the scalability, the coherence and the completeness of a rule-based system. A graphical rule editor has been implemented to help the designer in the design and the modification of the rules base. Mechanisms for checking the structural coherence (two rules with equivalent premises must have coherent conclusions) are

Figure 3. Different possible presentations to express the receipt of a call on a mobile phone. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



also proposed but the designer is still responsible of the completeness of the rules base.

Three types of rules are distinguished: contextual, composition and property rules. The premises of a contextual rule describe a state of the interaction context. The conclusions define contextual weights underlining the interest of the aimed interaction components (according to the context state described in the premises rule). The composition rules allow the modalities composition and so the design of multimodal presentation with several (modality, medium) pairs based on redundancy and/or complementarity criteria. Lastly, the property rules select a set of modalities using a global modality property (linguistic, analogical, confidential, etc.).

By analogy with the political world, we call our allocation process: "election". Our election process uses a rules base (voters) which add or remove points (votes) to certain modes, modalities or media (candidates), according to the current state of the interaction context (political situation, economic situation, etc.).

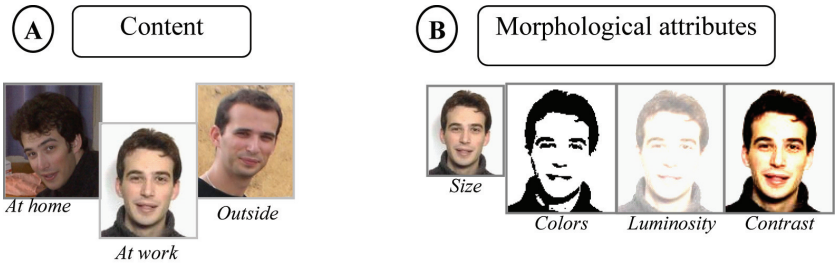
The application of the contextual and property rules defines the "pure" election while the application of the composition rules defines the "compound" election. The pure election elects the best modality-medium pair while the compound election enriches the presentation by selecting new pairs redundant or complementary to the first one.

How

When the allocation is done, the resulting multimodal presentation has to be instantiated. Instantiation consists in determining the concrete lexico-syntactical content of the selected modalities and their morphological attributes[5] depending on interaction context (Figure 1, Cj). First, a concrete content to be expressed through the presentation modality has to be chosen. Then, presentation attributes (modality attributes, spatial and temporal) parameters are set. This phase of the WWHT model deals with the complex problem of multimodal generation (André, 2000; Rist, 2005). The space of possible choices for the content and attributes values of modalities may be very large and thus the choice complex.

Ideally, the content generation should be done automatically. However this is still an open problem for each considered modality such as text generation (André 2003) or gesture generation (Braffort, 2004). For us the problem is rather to select one content between *n* possible predetermined contents and then to determine the adequate values for the morphological attributes. Figure 4 (A) shows an example of possible contents for

Figure 4. Which content and which attributes for the photographic modality? © Yacine Bellik. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission



the photographic modality while Figure 4 (B) shows possible morphological attributes for the same modality.

Then

We could think that when instantiation is done, the problem of building a multimodal presentation adapted to the current interaction context is resolved. Actually, the interaction context is dynamic and thus may evolve. This reveals the problem of presentation expiration. Indeed, the presentation may be adapted when it is built but there is a risk that it becomes inadequate if the interaction context evolves. This expiration problem concerns mainly the *persistent* presentations. Thus a multimodal presentation has to remain adapted during its whole life cycle. This constraint requires the use of mechanisms that allow the presentation to evolve according to the following factors:

- information factor,
- interaction context,
- time factor,
- space factor,
- user actions.

Information factor is a common evolution factor. For instance, the presentation of a laptop battery evolves according to its power level. In the case of the interaction context, its modifications may induce presentation expiration but it is not always the case. For instance, a visual presentation will not be affected by a higher noise level. The time factor may be important in some applications. Calendar applications are a good example. An event may be presented differently at two different moments. For instance, strikethrough characters may be used to display the event when it becomes obsolete. The space factor refers to the position and space size allocated to a given presentation. For instance, the FlexClock application (Grolaux, 2002) adapts the presentation of a clock according to its window size. The clock is displayed using graphics and text when the size window is big and only text modality when it is small. Finally, user actions may also influence the presentation. For instance, when the mouse cursor hovers a given icon, an attached text tip may appear.

We define two (non-exclusive) types of presentation evolution: *refinement* and *mutation*. On the one hand, refinement doesn't change the presentation means (modalities, media) used by the presentation. It affects only their instantiations. On the other hand, mutation induces modifications in the modalities and/or media used by the presentation. This difference is important because it requires different mechanisms to handle each type of evolution. Refinement requires a back-track to the instantiation (*how*) phase only while mutation requires a back-track to the allocation (*which*) phase. For instance, let us consider a multimodal presentation for the power level of a battery.

Figure 5. Evolution types of a presentation. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

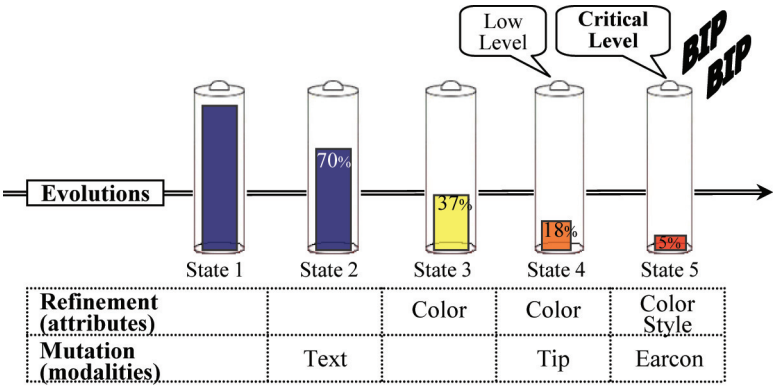


Figure 5 proposes four possible evolutions. The presentation at 70% (State 2) evolves by adding a text modality. We have a mutation here. The presentation at 37% (State 3) modifies the colors. It is a refinement. Finally, the presentation at 5% (State 5) combines both evolution types[6].

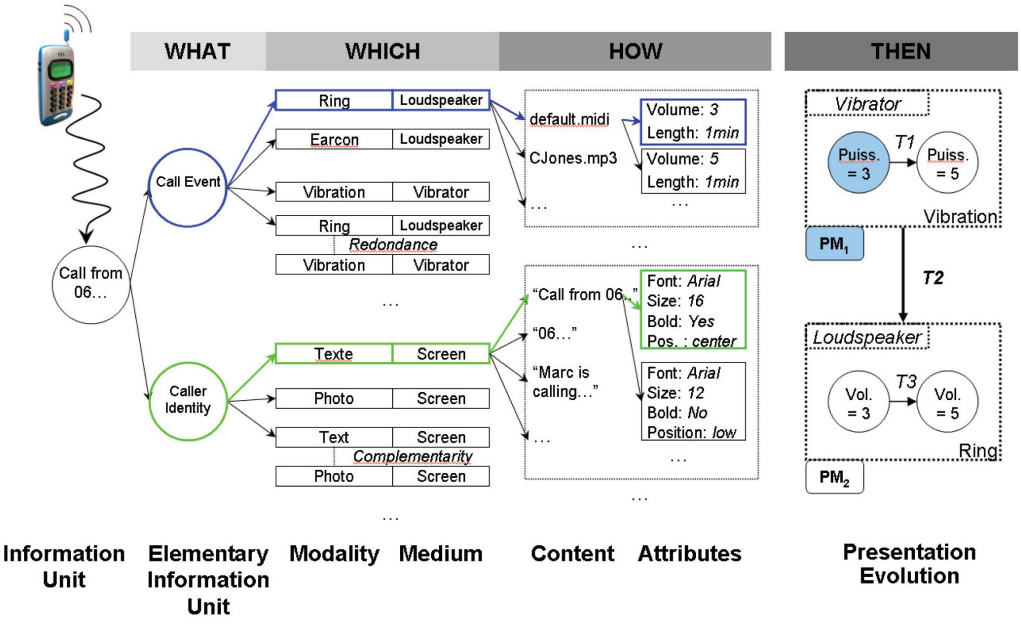
To sum up, Figure 6 shows a possible application of the different steps of the WWHT model for the mobile phone call scenario. T1, T2 and T3 show possible successive presentation evolutions.

The ELOQUENCE Platform

We have developed a software platform derived from the WWHT model to assist the designer/developer during the process of elaborating multimodal presentations. This platform, called ELOQUENCE, has been used in two different applications: a fighter cockpit simulator (Figure 7) and an air traffic control simulator (Figure 8) (Rousseau, 2006).

The ELOQUENCE platform includes two tools that respectively allow the specification and simulation of the system's outputs, and a runtime kernel which allows the execution of the final system. The specification tool (Figure 9) allows the designer to define all the elements required by the WWHT model: information units, presentation means, interaction context, and behavioral model. An analysis process must be applied to obtain these required elements. At first, it is necessary to collect a data corpus. This corpus must be composed of scenarios / storyboards (referring to nominal or degraded situations) but also of relevant knowledge on application field, system, environment, etc. Collecting this corpus must be strictly done and should produce consequent and diversified set of data. The corpus provides elementary elements needed to build the output system core (behavioral model). The quality of system outputs will highly depend on the corpus diversity. The participation and the collaboration of three actors is required: ergonomists, designers and end users (experts in the application field). Designers and users are mainly involved in the extraction of the elements while ergonomists are mainly involved in the interpretation of the extracted elements. The participation of all these actors is not an essential condition. However, the absence of an actor will be probably the source of a loss of quality in the outputs specification. Different steps should be followed to extract the required elements. The first step identifies pertinent data which can influence the output interaction (interaction context modeling). These data are interpreted to constitute context criteria and classified by models (user model, system model, environment model etc.). The next step specifies

Figure 6. Application of the WWHT model to a mobile phone call scenario. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

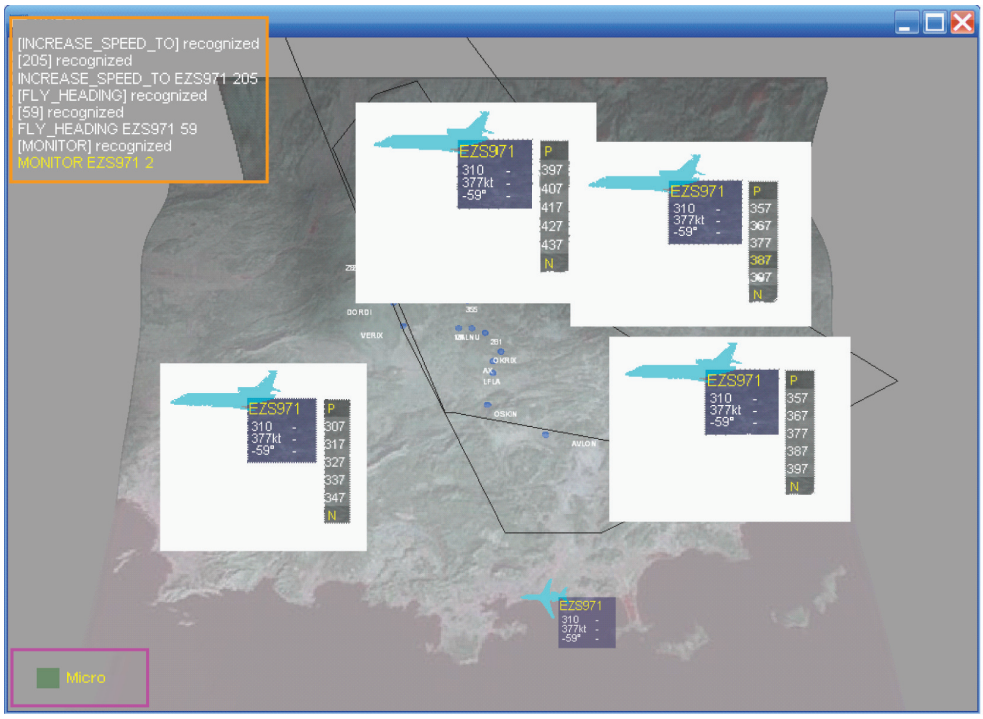


the interaction components diagram. Media are often defined first and from media it is relatively easy to identify output modes and modalities. The third step identifies semantic information which should be presented by the system. For better performance of the final system, it is recommended to decompose information into elementary semantic parts. At last, these extracted elements will allow the behavioral model definition.

The simulation tool constitutes a support for a predictive evaluation of the target application. It allows the designer to immediately check the

Figure 7. The fighter cockpit simulator. "© Thalès. Used with permission.

Figure 8. The air traffic control simulator. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



results of the specifications and thus makes the iterative design process easier. Figure 10 presents the simulation tool in the case of an incoming call on a mobile phone. It is composed of four parts. A first interface (A) simulates the dialog controller. More precisely it allows to simulate incoming information units from the dialog controller. A second interface (B) simulates a context server allowing the modification of the interaction context state. These two interfaces are generated automatically from the above specification phase. A third window (C) describes the simulation results in a textual form. Finally, a last interface (D) presents with graphics and sounds a simulation of the outputs results.

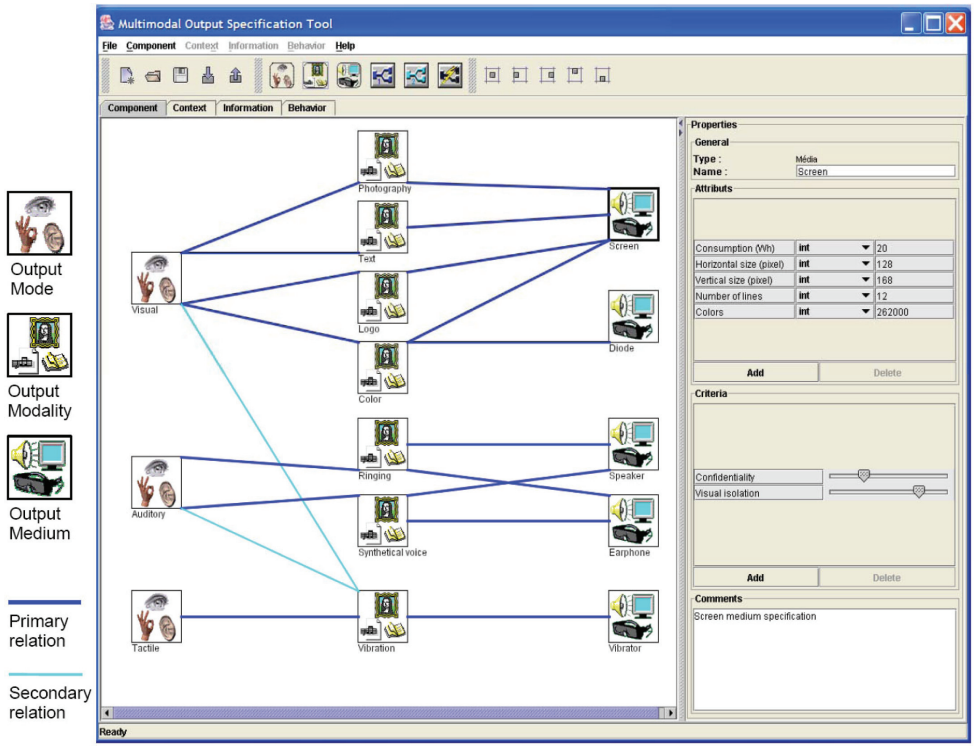
Finally, the runtime kernel is integrated in the final system. The kernel's architecture is centralized (Figure 11). The three main architecture modules (allocation, instantiation and evolution engines) implement the basic concepts of the WWHT model. A multimodal presentation manager completes the architecture by centralizing the resources and the communications between the different modules. The ELOQUENCE platform is described in more details in (Rousseau, 2006).

In the second research work, we will see that a distributed agent architecture is more adequate for the multimodal presentation of information in an ambient environment.

SECOND FRAMEWORK: OPPORTUNISTIC PRESENTATION OF INFORMATION

The second research work is about the opportunistic and multimodal presentation of information to mobile users in an ambient environment (Jacquet, 2005). The purpose here is to provide the mobile users with information through either private

Figure 9. The specification tool. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

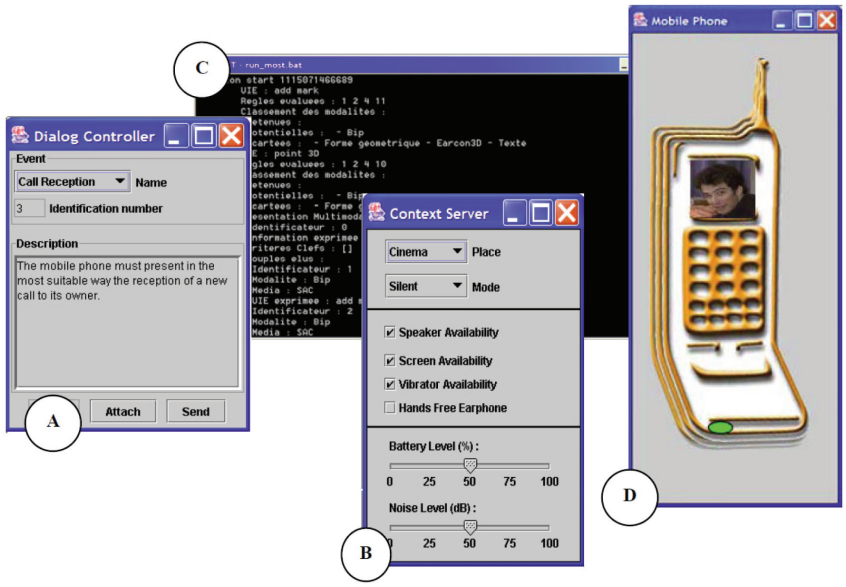


devices that they can carry with them (PDA, mobile phone, portable media player, etc.), or public devices that they may stumble upon while moving around (public display screens, loudspeakers, etc.). Let us underline three key points:

- Information relevant to users depends on the physical space where users are located. For instance, someone who gets inside a restaurant is most probably interested in the restaurant's menu; however a traveler who gets inside an airport is more likely to need to know which are his/her check-in desk and boarding gate.
- Information should be targeted to a given group of people. Indeed, it does not make sense to display an information item no one is interested in. It only confuses people and increases information lookup time. This is especially true for public displays, such as those found in airports, which are often overloaded with information (Figure 12).
- We consider information providing and information presentation to be two distinct processes. This means that a user can (conceptually) receive information as he/she moves, but be able to *look at* (or *listen to*) it only when he/she comes close to a suitable presentation device. Information can be temporarily stored by a digital representation of the user when the person is moving.

In this way, a user can gather information items in an opportunistic way when he/she encounters them, and even if there is no suitable presentation device at that moment. Information presentation can take place at a later time, in an opportunistic fashion too, when the user is close to a suitable device. This introduces a decoupling between

Figure 10. The simulation tool. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



two phases (providing and presenting information), which is necessary to both opportunistic behaviors. For this decoupling to be effective, the functional core must not be linked directly with the interface, from a software architecture point of view. It is thus necessary to introduce an intermediate entity between the interface and the functional core; otherwise information providing and presentation would be linked. For this reason, we propose the KUP model, in which the aforementioned decoupling occurs through a *user entity*.

The KUP Model

In an ambient intelligence system, the co-existence of both physical and digital entities prompts for new conceptual models for interactions. We in-

Figure 11. Runtime kernel architecture. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

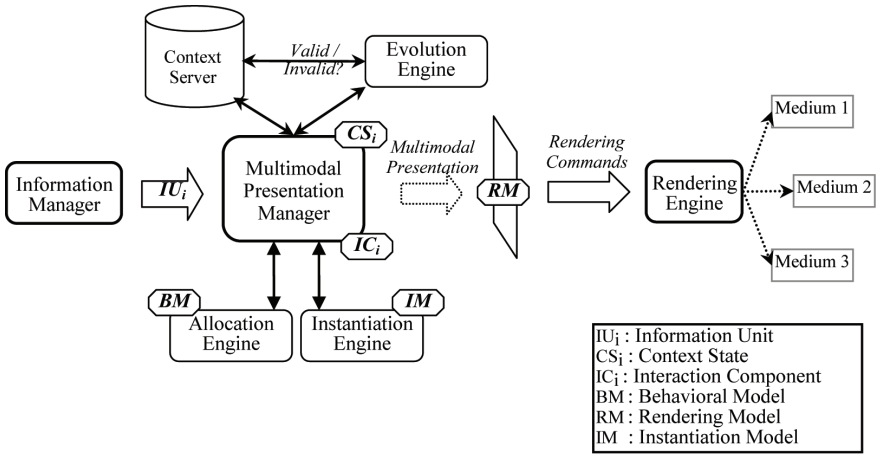


Figure 12. At Paris-Charles-de-Gaulle airport, this set of screens is permanently displaying a list of 160 flights, even when only three travelers are looking for information. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

troduce a model called KUP (Knowledge, User, Presentation) which is composed of three main entities, each of them having a physical facet and a digital (software) facet:

- K is the entity that represents information (or knowledge) sources. We call *semantic units* the information items produced by this entity (a semantic unit can be, for instance, the boarding gate of a given traveler; this notion is equivalent to the notion of "elementary information unit" in the WWHT model). The software facet of the K entity corresponds to the semantic component of classical architecture models (*functional core* in Seeheim (Pfaff, 1983) and ARCH (Bass, 1992), *abstraction facet* in PAC (Coutaz, 1987), *model* in MVC (Krasner, 1988), etc.).
- U is the user entity. Its physical facet corresponds to the human user. Its digital facet is active and is not limited to representing user attributes. For instance, it can store information for the user and negotiate the presentation of information items with devices.
- P is an entity responsible for presenting information to the user. Its digital facet corresponds to the *interface* of the classical architecture models. As we only consider outputs here, the interface is therefore limited to information presentation. Its physical counterpart corresponds to the presentation device.

KUP's architecture model introduces two original features:

- it includes an active software representation of the user (U), whereas it is generally omitted or very basic in the classical models. This software representation is more than a mere description of the users with a profile or preferences;
- this software entity associated with the user lies at the center of the model, which gives an utmost importance to the user, especially because all communications

within the model are handled by this entity. Ultimately this "user" entity is responsible for decoupling information providing and information presentation.

KUP's architecture model distinguishes itself from the classical architecture models (Seeheim, ARCH, PAC, MVC, etc.) because in the latter the user is always outside the system; it is never explicitly represented by an active entity (Figure 13). Conversely, in KUP (Figure 14) the digital entity representing the user is considered as the core of the model, which allows to decouple the process of providing information (performed by K entities) from the process of presenting information (performed by P entities).

Physical space plays a significant role in an ambient intelligence system. In particular, the system is supposed to react to certain user movements in the physical space. To model these interactions and ultimately build systems able to react to the corresponding events, we introduce two concepts that we think are of highest importance in an ambient intelligence system: the *perceptual space* of an entity and the converse notion, the *radiance space* of an entity. Roughly speaking, they respectively correspond to what an entity can perceive, and to which positions it can be perceived from. Let us now define these notions properly.

First and foremost, let us define what *perception* means for the different kinds of entities considered. For a user, perception corresponds to *sensory perception*: hearing, seeing, or touching another entity. For non-human entities, perception corresponds to the *detection* of other entities. For instance, a screen or an information source can be able to detect nearby users and non-human entities. From a technological point of view, this can be achieved by a variety of means, for instance using an RFID reader.

Perceptual Space

Intuitively we wish to define the perceptual space of an entity as the set of positions in the physical space that it can *perceive*. However, *perception* depends on the input modalities that *e* uses, each of which has a given perceptual field. For instance, for the user entity, the visual field and the auditory field of a human being are not the same: a text displayed on a screen located at position *P* two meters on the back of the user cannot be seen, whereas a sound emitted by a loudspeaker located at the same place can be heard without any problems. Does *P* belong to the perceptual space of the user? The answer to this question depends not only on the position of *P*, but also on the modality considered and even on attribute values. For example, a text displayed two meters ahead of the user can be read with a point size of 72[7], but not with a point size of 8. Therefore the perceptual space depends on the physical space, on the modality space, and on the attribute modality space.

We define the notion of *multimodal space* or *m-space*, as the Cartesian product of the physical space *E*, the space *M* of available modalities, and the space of modality attributes *A*. (More precisely, it is the union of Cartesian products, because the space of modality attributes depends on the modality considered.) A point in the m-space is defined by a tuple containing the physical coordinates *c* of the point, a given modality *m*, and an instantiation *i* of this modality (set of values for the attributes of *m*). One can then define the perceptual space of an entity *e* as the set of points *X(c, m, i)* of the m-space such that if another entity is located at the physical coordinates *c* and uses the modality *m* with the instantiation *i*, then it is perceived by *e*. Indeed, the definition of the perceptual space of an entity *e* includes spatial positions, but these positions are conditioned by the modalities used. Note that not all points in the m-space make sense, but this is not a problem as a perceptual space is always a *subset* of the m-space.

Figure 13. Classical HCI model: the user has no explicit representation in the interactive system. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

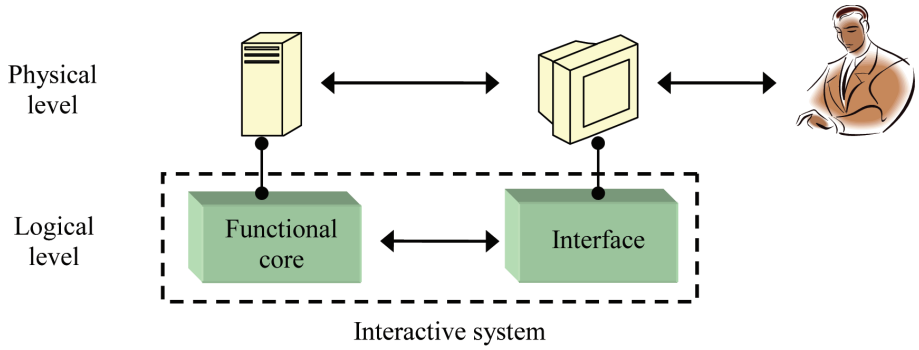
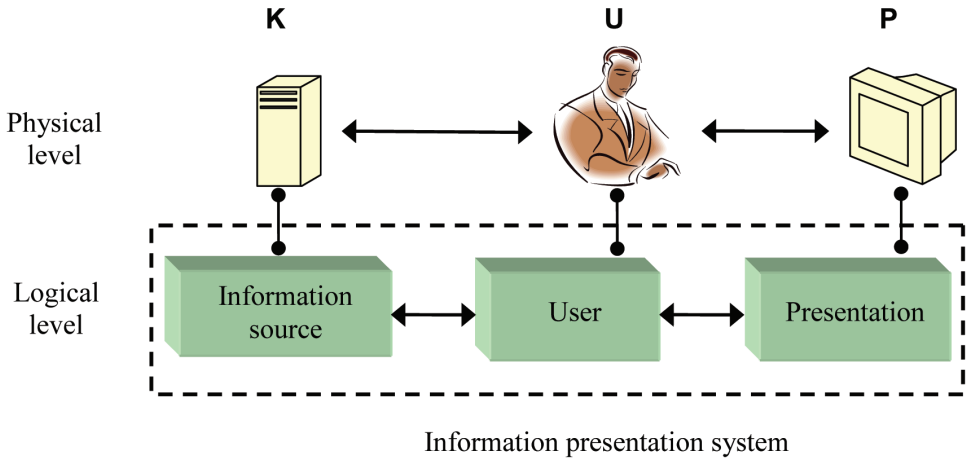


Figure 14. KUP model: the user is at the center of the information presentation system. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



Radiance Space

We define the radiance space of an entity *e*, using a modality *m* with an instantiation *i*, towards an entity *e'*, as the set of points *x* of the physical space *E* such that if *e'* is located in *x*, then *e* belongs to the perceptual space of *e'*. In other words, it is the set of points in space from which *e'* can perceive *e*. Let us notice that the radiance space of an entity *e* is always defined with respect to another entity *e'*. Indeed, the radiance space not only depends on the "emitting capabilities" of *e*, but also on "receiving capabilities" of *e'*. Thus at a given point in space *x*, an entity e_1' may perceive *e*, whereas another entity e_2' may not. For instance, the radiance space of a loudspeaker that sends out a message at a given sound level will depend on the receiving entity (deaf user, or user with no auditory problem). Therefore it is not possible to define a radiance space in absolute terms. The situation is analogous to that of satellite telecommunication. The coverage area of a satellite (which can be termed as its radiance space), is the set of points of Earth's surface where it is possible to receive signals from the satellite, *with a dish antenna of a given diameter*. The cover-

age area cannot be defined independently of the receiving antenna.

The concepts of perceptual space and radiance space are respectively reminiscent of those of *nimbus* and *aura* introduced by Benford and Fahlen (Fahlen, 1992; Benford, 1993). However, the nimbus of an entity represents what is perceived by this entity, whereas the perceptual space is the spatio-modal dimension in which the nimbus can build up. Likewise, the aura of an entity represents the set of manifestations of this entity, whereas the radiance space is the spatio-modal dimension in which the aura can express itself.

Sensorial Proximity: Originating Event

In the KUP model, all interactions between entities happen following a particular event: *sensorial proximity*. This kind of event arises when an entity e_1 enters or leaves the perceptual space of another entity e_2[8]. With the above definitions for radiance and perceptual spaces, we must underline the fact that the concept of sensorial proximity spans two aspects. On the one hand, it includes spatial proximity that refers to the distance between the two entities, and their respective orientations. On the other hand it also includes the input/output capabilities of the entities[9]. For instance, a blind user coming very close to a screen will trigger no sensorial proximity event. It is the same for a sighted user coming at the same place but with his/her back towards the screen.

As our work focuses on information presentation, and therefore on *output*, we consider that changes in sensorial proximity are the only type of input events in the system. They trigger all of the system's reactions, in particular the outputs produced by the system. Up to now we do not have studied other types of input, for instance explicit user input, but this could be the subject of further investigation.

An Opportunistic Model for the Presentation of Information

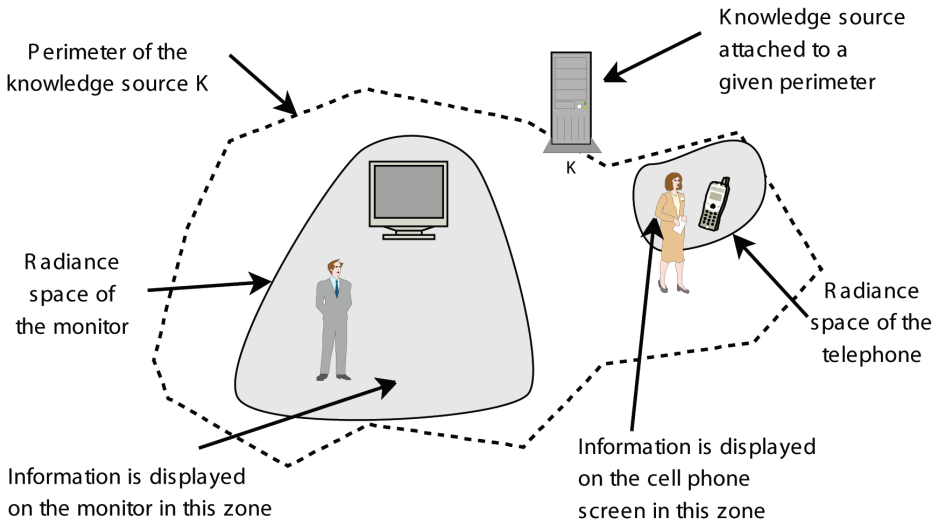
Using the KUP model, one can clearly separate the process of providing information, from the process of presenting information. When a user (U) enters the radiance space of a knowledge source (K), the latter provides his/her logical entity with one or several relevant semantic units. At the moment when the user receives these semantic units, it is possible that no presentation device (P) is within his/her perceptual space. However, since users are supposedly mobile, it is possible that a presentation device enters the user's perceptual space at a later time. This will then trigger a sensorial proximity event which will initiate the process of presenting the user's semantic units on the device[10]. Figure 15 summarizes the interaction between knowledge sources, users and presentation devices.

Agent Architecture

A configuration like the one presented on Figure 18 could be hardwired, but this would not take full advantage of the very modular structure of the KUP model. For instance, if the staff changes the location of knowledge sources and presentation devices, or brings in new devices in case of a particular event, it would be cumbersome if a hardwired configuration had to be changed by hand. However, as the KUP model entirely relies on perception relationships between entities, it is possible to design implementations that autoconfigure and adapt to changes without human intervention. We propose a decentralized architecture based on agents, in which each entity of the model has an agent counterpart:

- user agents (U) are an active software representation of human users,
- knowledge agents (K) provide user agents with information,
- presenter agents (P) are the software interface of the physical presentation devices.

Figure 15. A semantic unit provided by a knowledge source can be presented by several presentation devices. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



They can evaluate the cost of the presentation of an information item on a device, and perform the presentation.

This world of agents is a "mirror" of the real world, at least as far as our three types of entities of interest are concerned.

We assume that all agents can communicate with one another. Sensorial proximity relations, originating in the physical world, are mirrored in the world of agents. For instance, if a user *a* perceives a presentation device *b*, then the same relation exists between the agent counterparts. From a technical point of view, communication can use now-ubiquitous wireless networks such as Wi-Fi or mobile phone networks; sensorial proximity can be detected by various means, such as RFID tag detection, feature recognition in images, etc. Although agents have a notion of *having a location*, this does not mean that the corresponding processes have to run at this location. It is possible to have agents run on delocalized servers while retaining the flexibility of the architecture.

Agents are reactive: they are idling most of the time, and react when particular events occur. In practice, a given agent *a* can react to three types of events:

- another agent *b* has just come close to *a*[11],
- an agent *b*, that previously was close to *a*, has just gone away,
- *a* has just received a message through the network, from an agent *c*, which is not necessarily close to *a*.

Therefore, if there were only agents in the system, nothing would ever happen. Indeed, agents have reactive behaviors when the associated physical entities move. It means that the proactive properties of the system fully come from physical entities and human users: the latter usually move, and hence trigger cascades of reactions in the system.

Allocation and Instantiation in KUP

In KUP, allocating and instantiating modalities happen in a decentralized fashion. In the first framework, as interaction involved only a unique user on a unique workstation, we had preferred a

centralized approach. Conversely, in this second framework, entities involved are disseminated in space, so it makes senses to resort to a decentralized architecture that matches the agent architecture mentioned above. In this way, when a U entity enters the radiance space of a P entity, the two associated agents will negotiate the most suitable modality[12] (and its instantiation) to present U's semantic units on the presentation device. This negotiation process relies on the concept of *profile*. A profile is a set of weights given to the modalities and their instances. Profiles are defined with respect to a tree-like taxonomy of modalities, common to the three types of entities.

Figure 16 gives an example of a partial taxonomic tree for output modalities.

Each entity defines a weighting tree which is superposed to the taxonomic modality tree[13]. The goal of a weighting tree is to add weights to a taxonomic tree, in order to express capabilities, preferences and constraints of users, devices and semantic units. A weight is a real number between 0 (included) and 1 (included). It can be located at two different places:

- **at node level:** the weight applies to the whole sub-tree rooted at this node. A weight of 1 means that the modalities of the sub-tree may be used, whereas a weight of 0 means that the modalities may not be used. Values in between allow one to introduce subtle variations in how much a modality is accepted or refused. This is used to express preference levels,
- **at attribute level:** the weight is a function that maps every possible attribute value to a number in the interval [0, 1]. This function indicates the weight of every possible value for the attribute. The meaning of the weights is the same as above: attributes values whose weight is close to 1 are acceptable; values whose weight is close to 0 are not.

A profile is defined as a weighting tree that spans the whole taxonomic tree. Figure 17 gives an example of a partial profile. It could correspond to an American, visually impaired user, who by far prefers auditory modalities over visual ones. The weights are given inside black ovals, just next to the nodes. Weight functions are given for some attributes. Depending on whether the attributes are of continuous or discrete nature, the functions are either continuous or discrete.

Given a user *u*, a presentation device *d* and a semantic unit *s*, the most suitable modality (along with its instantiation) to present *s* to *u* on *d* is determined by considering the *intersection* of the

Figure 16. Example of a partial taxonomy for output modalities. In this basic example we consider two kinds of output modalities, visual ones (example: text) and auditory ones (example: computer-generated spoken dialogue). © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.

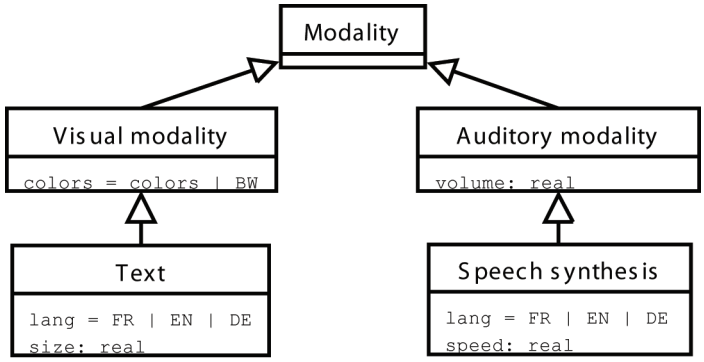
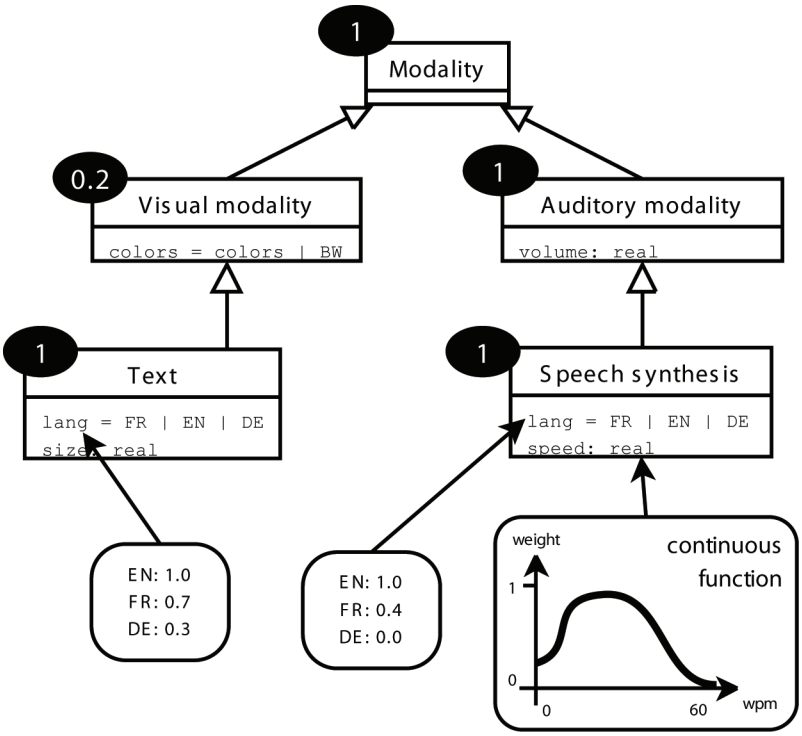


Figure 17. Example of a partial profile (weighting tree). © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



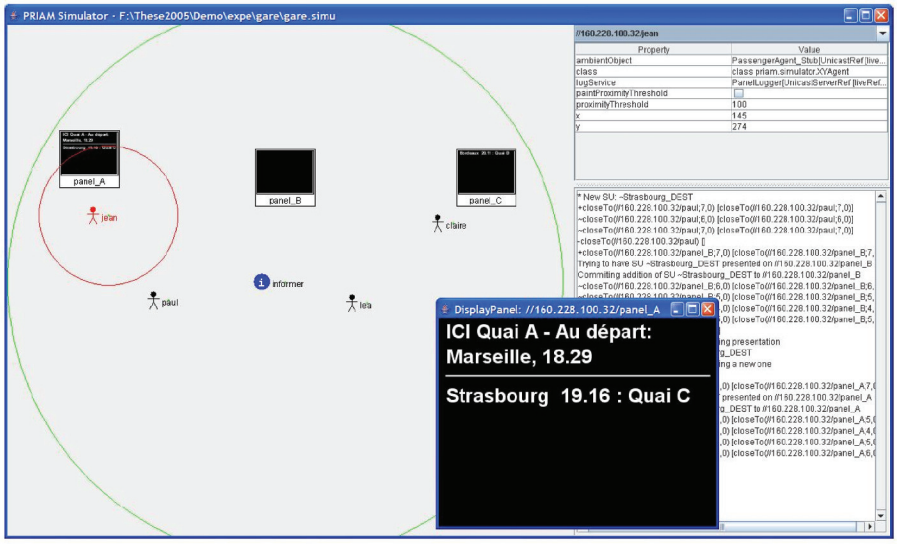
three weighting trees. This intersection method eventually produces a resulting weighting tree whose leaves are candidate modalities. Then, all that the system has to do is choose the modality with the highest weight, and instantiate it using the attribute values with the highest weights. Actually, this is the simplest of the situations, where only one semantic unit is to be presented to one user, using one presentation device. In the more general case where there are several users close to a device, or conversely if there are several devices close to one (or several) devices, more complex algorithms have been designed to have several devices collaborate with one another. They ensure a global consistency of information presentation while guaranteeing a minimal satisfaction level to each user. These algorithms are thoroughly described in (Jacquet, 2006; Jacquet, 2007).

The PRIAM Platform

In order to implement and validate the concepts introduced by the KUP model we have developed an agent platform called PRIAM (PResentation of Information in AMbient environments). Real-scale experiments being quite complex and costly to carry out, this platform includes a simulator that enables researchers to test every component of an ambient application, without being obliged to deploy it in real-scale (Figure 18). This simulator has also enabled us to validate the behavior of allocation and instantiation algorithms prior to real experiments. Every kind of situation, either simple or very complex, can be tested in this way, with the required number of presentation devices, users and knowledge sources.

We have nonetheless gone beyond the mere simulation stage. Three pseudo-real-scale experiments have been conducted. The objective

Figure 18. Simulator of the PRIAM platform. On this screenshot, three display screens and four users where simulated. The screens display information about departing trains. Screen contents appear on the simulator window, but they can also be "popped-out" like the one in the foreground. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



of these experiments was to prove that using a dynamic display that only displays information relevant to the people located at proximity, it was in average shorter and easier to find one's item of interest than using a long static list. Dynamic display was implemented on a computer, with an infrared detection system to detect the proximity of users. Static lists were either sheets of paper or static images on screen.

The first experiment displayed examination results for students. The second one was based on the same setting, but displayed airport information (boarding gate number). In these experiments we compared the times needed to find one's information item with static and dynamic displays. The purpose of the third experiment was to help train passengers to find a transfer, without having to walk too much. In this one, we compared the number of elementary moves needed to reach one's platform, when using static or dynamic displays. These experiments enabled us to test the platform in pseudo-real scale, and to demonstrate the benefits of displaying only information relevant to users located in the vicinity of a presentation device. Indeed, in this case the device is far less overloaded with irrelevant items and users can lookup the items of interest more quickly. The results show that dynamic displays are superior in all cases: lookup time is respectively 50% and 25% faster in the first and second experiments, the number of elementary moves was divided by 2,4 in the third experiment (Jacquet, 2007).

The experiments have underlined some issues too, especially those related to privacy. For instance, when a passenger is alone in front of a screen, it is easy for an ill-intentioned person to know his/her destination. In one possible solution the system would introduce a few irrelevant items in the presentation so as to add noise and thus prevent third parties to gain access to private information.

In this second work, we have explored the problem of presenting multimodal information in an ambient setting. The distributed nature of information systems within the physical environment of users has led us to choose a multi-agent

model in which agents representing users lie at the core. It is to be noted that the world of agents is only a mirror of the real world: the system is not proactive by itself; instead it reacts to changes and moves originating in the physical world. This means eventually that we do not really build a *world of agents*, but rather that we *agentify the real world*. This takes us back to the vision of ambient intelligence in which computerized systems monitor the actions of human beings in an unobtrusive way, so as to trigger actions when it is really relevant, and without disturbing normal user actions.

COMPARISON

Both research works presented in this chapter explored the benefits of using several output modalities to improve the way information is presented to the user. They address two different situations: a "classical" interaction situation where the user is fixed with respect to a unique interactive system and an ambient interaction situation in which the user moves and is likely to use several interactive systems.

In both cases we have proposed models for adaptive output multimodal systems. However, to not face several difficulties at the same time, we divided the issues between the two frameworks. For instance, in the second one is focused on the new constraints induced by the ambient environment. Hence, we have adopted a distributed agent architecture while the first frameworks relies on a centralized architecture. We have also insisted in the second project on the need to have an active representation of the user. However we have limited ourselves on some aspects that have already been explored in the first project. Thus we have used only exclusive[14] multimodality (while complementary or redundant multimodality was supported in the first project) and we did not handle presentation evolution since it was already explored in the first project.

Regarding allocation and instantiation problems, the algorithms used in both frameworks are quite different. Algorithms defined in the second one are more powerful. For instance, the allocation process in the first framework is directive while it is cooperative in the second one. In the same way the instantiation process in the first framework is local while it is global in the second one.

Finally there are still some problems where we adopted the same options in both projects because these problems are still open and constitute some of our future research directions. For instance, the semantic fission process is manual in both frameworks, the content of modalities is predefined and not automatically generated in both frameworks, and the instantiation process is homogeneous in both frameworks (we will detail this problem in the next section).

Table 2 synthesizes the comparison between both frameworks.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Thanks to the interaction richness it can offer, multimodality represents an interesting solution to the problems induced by an ever more variable interaction context. It is no longer reasonable today to continue to propose static and rigid interfaces while users, systems and environments are more and more diversified. To the dynamic character of the interaction context, the interface must also respond by a dynamic adaptation. The two frameworks described above constitute a first answer to the problem of adaptive multimodal presentation of information. However, they have also revealed some new problems that have not been addressed yet and which represent our future research directions. We summarize them in the following sections.

Table 2. Comparison between both frameworks. The last column show what an ideal adaptive output multimodal system should be. © Yacine Bellik. Used with permission.

Criterion	First framework	Second framework	Ideal system
Type of architecture	Centralized architecture	Distributed architecture	Depends
Support of CARE properties	Redundancy/Complementarity supported	Exclusive multimodality	Redundancy/Complementarity supported because a system which supports CARE properties is also capable of supporting exclusive multimodality
Content generation automaticity	No (predefined modality content)	No (predefined modality content)	Yes (generated modality content) because this will reduce the design costs.
Semantic fission automaticity	No (manual semantic fission)	No (manual semantic fission)	Yes (automatic semantic fission) because this will reduce the design costs.
Allocation strategy	Directive allocation	Cooperative allocation	Cooperative allocation because it allows to optimize the use of modalities and medias resources
Instantiation strategy	Local instantiation	Global instantiation	Global instantiation because it allows to optimize the use of modalities and medias resources
Type of instantiation	Homogeneous instantiation	Homogeneous instantiation	Heterogeneous instantiation because heterogeneous instantiation is more powerful. A system capable of heterogeneous instantiation is also capable of homogeneous instantiation

Heterogeneous instantiation

In our models we associate a unique instantiation to each elementary information unit. For instance, the string "Gate n° 15" which represent the concrete content of the elementary information unit indicating the boarding gate for a given flight, could be displayed using Arial font, 72 dots size; and white color. Hence, the instantiation of the morphological attributes is homogeneous and is applied to all elements of the modality content. However, sometimes it could be interesting to apply a particular instantiation to a part of the content. For instance, in the previous example, the number "15" could be displayed with a different color and a blinking bold style. It is not possible to specify it easily in our current models. One possible solution is to decompose again this elementary information unit into two others elementary information units, so we can instantiate each information unit independently. However this solution is not intuitive and will make the behavioral model complex. A more interesting solution could be to define a new modality called, for instance, "2Texts" which will gather a content composed by two strings and 2 sets of morphological attributes (one set for each string). This way it becomes possible to associate to each content part a different instantiation.

Fusion in Output

Usually, fusion is a concept which is associated to input multimodality. However this concept can also be relevant to output multimodality, depending on the global software architecture (centralized/distributed) and the inconsistency detection strategy (early/late strategy). Let us take again the example seen in the first project and which is about the receipt of a phone call. We have seen that this semantic information can be decomposed into two elementary information units: the phone call event and the caller identity. Let us suppose that for each of these information units the allocation process chooses the "Ring"

modality: a generic ring for the call event and a custom ring for the caller identity. This will induce inconsistence within the whole multimodal presentation and thus a back-tracking to the allocation process for a new allocation request. A possible solution could be to exploit the time factor to play both rings in a sequential way. However, if the system plays the custom ring first then the generic ring becomes useless. And if the generic ring is played first then it is likely that the user would have already answered the call before the system would have played the custom ring. A better solution would be to merge both modality contents and finally keep only the custom ring. Indeed custom ring is capable of expressing both information units, while the generic ring can only express the call event. This example shows that sometimes, a given instantiation may be attached to two different information units. This kind of relation is not yet supported in our models. We will try in future work to add this kind of relation to our models so we will be able to define output fusion algorithms.

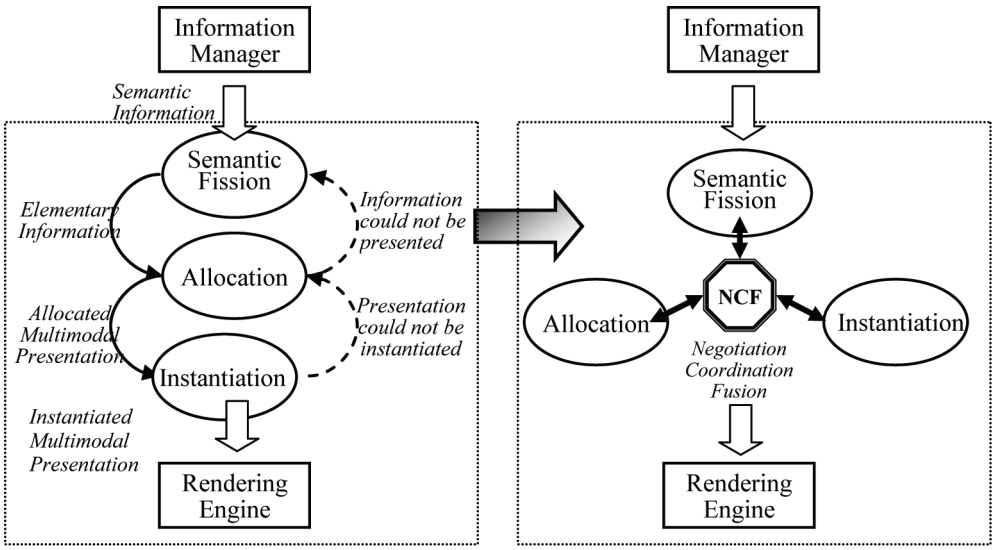
Negotiation-Based Approach

Our current models apply their different phases in a sequential way. In the case of a blocking situation in a given phase, a back-tracking is done to the previous phase. It could be interesting to explore another approach based on a true negotiation between the different modules involved during the different phases (Figure 19). However the mechanisms of this negotiation process are still to be defined.

Influence of Inputs on Outputs

Inputs and outputs in an interactive system can be considered as dynamic interdependent flows of information. Thus, system outputs have to remain consistent with user inputs to ensure good interaction continuity. This continuity cannot be achieved unless inputs and outputs are incorporated inside the same design process. The modalities and media used in input may influence the choice of output modalities in particular in the case of lexical feedbacks. For instance, text entered on a keyboard will generally induce a visual feedback, while it could

Figure 19. Toward a negotiation-based approach. © Y. Bellik, C. Jacquet, C. Rousseau. Used with permission.



sometimes be preferable in the case of a speech command to produce a speech feedback.

As regards the semantic feedbacks, the second application developed using the first framework (air traffic control simulator) showed that the output part of the system needs to maintain an internal representation of the multimodal presentations it has provided. This allows the output part to answer requests coming from other modules, for instance about the pointed objects: when the user clicks on an (X,Y) position on the screen, only the output part of the system knows which object is located at this position since it is the output part of the system which knows which modalities and which modality attributes have been used to present application objects. This second application showed also that the output part of the system needs to know which input interaction means have been used so as to provide consistent output presentations.

Even though this application allowed us to start exploring some solutions, there are still some questions which need to be investigated, in particular those about the overall software architecture of a bidirectional (input and output) multimodal system.

REFERENCES

André, E. (2000). The generation of multimedia presentations. In R. Dale, H. Moisl & H. Somers (Eds.), *A Handbook of Natural Language Processing* (pp. 305–327). New York: CRC.

André, E. (2003). Natural language in multimedia/multimodal systems. In R. Mitkov (Ed.), *Computational Linguistics* (pp. 650-669). New York: Oxford University Press.

André, E., Finkler, W., Graf, W., Rist, T., Schauder, A., & Wahlster, W. (1993). The automatic synthesis of multimodal presentations. In M. T. Maybury, (Ed.), *Intelligent Multimedia Interfaces* (pp. 75-93). Menlo Park, CA: AAAI Press.

Arens, Y., & Hovy, E. H. (1995). The design of a model-based multimedia interaction manager. *Artificial Intelligence*, *9*(3), 167–188. doi:10.1016/0954-1810(94)00014-V

Balme, L., Demeure, A., Barralon, N., Coutaz, J., & Calvary, G. (2004). CAMELEON-RT: A software architecture reference model for distributed, migratable, and plastic user interfaces. In *Proc. of EUSAI 2004, European symposium on ambient intelligence, No. 2, vol. 3295 of Lecture Notes in Computer Science* (pp. 291-302). Eindhoven, The Netherlands: Springer.

Bass, L., Faneuf, R., Little, R., Mayer, N., Pellegrino, B., & Reed, S. (1992). A metamodel for the runtime architecture of an interactive system. *SIGCHI Bulletin*, *24*(1), 32–37. doi:10.1145/142394.142401

Bellik, Y. (1995), *Interfaces multimodales: Concepts, modèles et architectures*. Unpublished doctoral dissertation, Université Paris XI, Orsay, France.

Benford, S., & Fahlen, L. (1993). A spatial model of interaction in virtual environments. In *Proceedings of the Third European Conference on Computer Supported Cooperative Work (ECSCW'93)*.

Bernsen, N. O. (1994). Foundations of multimodal representations: a taxonomy of representational modalities. *Interacting with Computers*, *6*(4). doi:10.1016/0953-5438(94)90008-6

Bohn, J., & Mattern, F. (2004), Super-distributed RFID tag infrastructures. In *Proceedings of the 2nd European Symposium on Ambient Intelligence (EUSAI 2004)* (pp. 1–12). Berlin, Germany: Springer-Verlag.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics*, *14*(3), 262–270. doi:10.1145/965105.807503

Bordegoni, M., Faconti, G., Maybury, M. T., Rist, T., Ruggieri, S., Trahanias, P., & Wilson, M. (1997). A standard reference model for intelligent multimedia presentation systems. *Computer Standards & Interfaces*, *18*(6-7), 477–496. doi:10.1016/S0920-5489(97)00013-5

Braffort, A., Choisier, A., Collet, C., Dalle, P., Gianni, F., Lenseigne, B., & Segouat, J. (2004). Toward an annotation software for video of sign language, including image processing tools and signing space modeling. In *Proceedings of the Language Resources and Evaluation Conference (LREC'04)*.

Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Souchon, N., Bouillon, L., et al. (2002). Plasticity of user interfaces: A revised reference framework. In *TAMODIA '02: Proceedings of the First International Workshop on Task Models and Diagrams for User Interface Design* (127–134). Bucharest, Romania: INFOREC Publishing House Bucharest.

Coutaz, J. (1987). PAC, an object-oriented model for dialog design. In H.-J. Bullinger, B. Shackel (Eds.), *Proceedings of the 2nd IFIP International Conference on Human-Computer Interaction (INTERACT 87)* (pp. 431-436). Amsterdam: North-Holland.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., & Young, R. M. (1995). Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT'95)*.

Coutaz, J., & Rey, G. (2002). Foundation for a theory of contextors. In [New York: ACM Press.]. *Proceedings of CADUI*, *02*, 283–302.

Dalal, M., Feiner, S., McKeown, K., Pan, S., Zhou, M., Höllerer, T., et al. (1996). Negotiation for automated generation of temporal multimedia presentations. In *Proceedings of ACM Multimedia'96* (pp. 55-64).

Demeure, A., & Calvary, G. (2003). Plasticity of user interfaces: towards an evolution model based on conceptual graphs. In *Proceedings of the 15th French-speaking Conference on Human-Computer Interaction*, Caen, France, (pp. 80-87).

Dey, A. K. (2000). *Providing architectural support for building context-aware applications*. Unpublished doctorial dissertation, Georgia Institute of Technology, College of Computing.

Duarte, C., & Carriço, L. (2006), A Conceptual Framework for Developing Adaptive Multimodal Applications. In *Proceedings of Intelligence User Interfaces (IUI'06)* (pp. 132-139).

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., & Burgelman, J.-C. (2001). *Scenarios for Ambient Intelligence in 2010, Final report*. Information Society Technologies Advisory Group (ISTAG), European Commission.

Elting, Ch., & Michelitsch, G. (2001). A multimodal presentation planner for a home entertainment environment. In *Proceedings of Perceptual User Interfaces (PUI) 2001*, Orlando, Florida.

Elting, C., Rapp, S., Möhler, G., & Strube, M. (2003). Architecture and Implementation of Multimodal Plug and Play. In *ICMI-PUI `03 Fifth International Conference on Multimodal Interfaces*, Vancouver, Canada.

Elting, Ch., Zwickel, J., & Malaka, R. (2002). Device-dependent modality selection for user-interfaces - an empirical study. In *International Conference on Intelligent User Interfaces IUI 2002*, San Francisco, CA.

Fahlen, L., & Brown, C. (1992). The use of a 3D aura metaphor for computer based conferencing and teleworking. In *Proceedings of the 4th Multi-G workshop* (pp. 69-74).

Fasciano, M., & Lapalme, G. (1996). PosGraphe: a system for the generation of statistical graphics and text. In *Proceedings of the 8th International Workshop on Natural Language Generation* (pp. 51-60).

Feiner, S. K., & McKeown, K. R. (1993). Automating the generation of coordinated multimedia explanations. In M. T. Maybury, (Ed.), *Intelligent Multimedia Interfaces* (pp. 117-139). Menlo Park, CA: AAAI Press.

Frohlich, D. M. (1991). The design space of interfaces. In L. Kjelldahl, (Ed.), *Multimedia Principles, Systems and Applications* (pp. 69-74). Berlin, Germany: Springer-Verlag.

Gellersen, H., Kortuem, G., Schmidt, A., & Beigl, M. (2004). Physical prototyping with smart-its. *IEEE Pervasive Computing / IEEE Computer Society [and] IEEE Communications Society*, *3*(3), 74–82. doi:10.1109/MPRV.2004.1321032

Grolaux, D., Van Roy, P., & Vanderdonckt, J. (2002). FlexClock: A plastic clock written in Oz with the QTk toolkit. In *Proceedings of the Workshop on Task Models and Diagrams for User Interface Design (TAMODIA 2002)*.

Jacquet, C. (2007). KUP: a model for the multimodal presentation of information in ambient intelligence. In *Proceedings of Intelligent Environments 2007 (IE 07)* (pp. 432-439). Herts, UK: The IET.

Jacquet, C., Bellik, Y., & Bourda, Y. (2005). An architecture for ambient computing. In H. Hagras, V. Callaghan (Eds.), *Proceedings of the IEE International Workshop on Intelligent Environments* (pp. 47-54).

Jacquet, C., Bellik, Y., & Bourda, Y. (2006), Dynamic Cooperative Information Display in Mobile Environments. In B. Gabrys, R. J. Howlett, L. C. Jain (Eds), *Proceedings of KES 2006, Knowledge-Based Intelligent Information and Engineering Systems*, (pp. 154-161). Springer.

Johnston, M., & Bangalore, S. (2005). Finite-state multimodal integration and understanding. *Natural Language Engineering*, *11*(2), 159–187. doi:10.1017/S1351324904003572

Kerpedjiev, S., Carenini, G., Roth, S. F., & Moore, J. D. (1997). Integrating planning and task-based design for multimedia presentation. In *Proceedings of the International Conference on Intelligent User Interfaces* (pp. 145-152).

Krasner, G. E., & Pope, S. T. (1988). A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. *Journal of Object Oriented Programming*, *1*(3), 26–49.

Lafuente-Rojo, A., Abascal-González, J., & Cai, Y. (2007). Ambient intelligence: Chronicle of an announced technological revolution. *CEPIS Upgrade*, *8*(4), 8–12.

Martin, J. C. (1998). TYCOON: Theoretical framework and software tools for multimodal interfaces. In J. Lee, (Ed.), *Intelligence and Multimodality in Multimedia Interfaces*. Menlo Park, CA: AAAI Press.

Navarre, D., Palanque, P., Bastide, R., Schyn, A., Winckler, M. A., Nedel, L., & Freitas, C. (2005). A formal description of multimodal interaction techniques for immersive virtual reality applications. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT'05)*.

Nigay, L., & Coutaz, J. (1993). Espace problème, fusion et parallélisme dans les interfaces multimodales. In *Proc. of InforMatique'93*, Montpellier (pp.67-76).

Nigay, L., & Coutaz, J. (1995). A generic platform for addressing the multimodal challenge. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'95)* (pp. 98-105).

Pfaff, G. (1983). User interface management systems. In *Proceedings of the Workshop on User Interface Management Systems*.

Rist, T. (2005). Supporting mobile users through adaptive information presentation. In O. Stock and M. Zancanaro (Eds.), *Multimodal Intelligent Information Presentation* (pp. 113–141). Amsterdam: Kluwer Academic Publishers.

Rousseau, C. (2006). *Présentation multimodale et contextuelle de l'information*. Unpublished doctoral dissertation, Paris-Sud XI University, Orsay, France.

Stephanidis, C., Karagiannidis, C., & Koumpis, A. (1997). Decision making in intelligent user interfaces. In *Proceedings of Intelligent User Interfaces (IUI'97)* (pp. 195-202).

Stephanidis, C., & Savidis, A. (2001). Universal access in the information society: Methods, tools, and interaction technologies. *UAIS Journal*, *1*(1), 40–55.

Stock, O., & the ALFRESCO Project Team. (1993). ALFRESCO: Enjoying the combination of natural language processing and hypermedia for information exploration. In M. T. Maybury (Ed.), *Intelligent Multimedia Interfaces* (pp. 197-224). Menlo Park, CA: AAAI Press.

Teil, D., & Bellik, Y. (2000). Multimodal interaction interface using voice and gesture. In M. M. Taylor, F. Néel & D. G. Bouwhuis (Eds.), *The Structure of Multimodal Dialog II* (pp. 349-366).

Thevenin, D., & Coutaz, J. (1999). Plasticity of user interfaces: Framework and research agenda. In *Proceedings of the 7th IFIP Conference on Human-Computer Interaction, INTERACT'99,* Edinburgh, Scotland (pp.110-117).

Wahlster, W. (2003). Towards symmetric multimodality: Fusion and fission of speech, gesture and facial expression. In Günter, A., Kruse, R., Neumann, B. (eds.), *KI 2003: Advances in Artificial Intelligence, Proceedings of the 26th German Conference on Artificial Intelligence* (pp. 1-18). Hamburg, Germany: Springer.

Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Communications of the ACM*, *36*(7), 75–84. doi:10.1145/159544.159617

ENDNOTES

1 Of course, in the case of a multimodal presentation, several modes may be used.

2 For instance a scanned text saved by the system as a picture will be perceived by the user as a text and not as a picture.

3 Radio-Frequency Identification

4 This process can be done recursively as in (Wahlster, 2003) where a presentation planner recursively decomposes the presentation goal into primitive presentation tasks.

5 The morphological attributes refer to the attributes that affect the form of a modality. For instance, font size for a visual text modality or volume for a spoken message.

6 We observe also a refinement with respect to the internal rectangle size and the text position in steps.

7 For a user with a normal visual acuity.

8 This is the same as saying that entity e_2 enters or leaves the perceptual space of entity e_1.

9 Thus the notion of perceptual proximity is not commutative in the general case.

10 The presentation can happen as long as the semantic units are not outdated. A semantic unit can become outdated for two reasons. *Spatial outdating* may happen when the user leaves the radiance space of the knowledge source that has provided the semantic unit (but this is not always the case). *Temporal outdating* is controlled by a metadata element associated with the semantic unit.

11 The notion of *closeness* refers to *sensorial proximity.*

12 To make a clear distinction between problems, we have decided to restrict the second

project to *exclusive* multimodality, because complementarity and redundancy of modalities has already been studied in the first project. We have instead focussed on constraints specific to ambient environments.

[13] Except K entities that define a weighting tree for each semantic unit that they produce. Indeed, each semantic unit is supposed to be able to express itself into its own set of modalities. In consequence, weighting trees are attached to the produced semantic units, not to the K entities themselves.

[14] Exclusive multimodality allows the use of different modalities, but not in a combined way.