A Taxonomy for
Multimedia and Multimodal User Interfaces

*Joëlle Coutaz *, Jean Caelen***

* Laboratoire de Génie Informatique (IMAG)
BP 53 X, 38041 Grenoble Cedex, France
Tel: +33 76 51 48 54
Fax: +33 76 44 66 75
e-mail: joelle@imag.fr

** Institut de la Communication Parlée (INPG)
46 av. Félix Viallet, 38031 Grenoble Cedex, France
Tel: +33 76 57 45 36
Fax: +76 57 47 10
e-mail: jcaelen@icp.imag.fr

Abstract

This paper presents one French research effort in the domain of multimodal interactive systems: the Pole Interface Homme-Machine Multimodale. It aims at clarifying the distinction between multimodal and multimedia systems and suggests a classification illustrated with current interactive systems.

1. Introduction

Graphical user interfaces (GUI) are now common practice. Although not fully satisfactory, concepts in GUI are well understood and software tools such as interaction toolkits and UIMS technology, are widely available. Parallel to the development of graphical user interfaces, natural language processing, computer vision, and gesture analysis [12] have made significant progress. Artificial and virtual realities are good examples of systems based on the usage of multiple modalities and medias of communication.

As noted by Krueger in his latest book "Artificial Reality II", multimodality and multimedia open a complete new world of experience [13]. Clearly, the potential for this type of system is high but our current understanding on how to design and build such systems is very primitive.

This paper presents one French research effort in the domain of multimodal interactive systems: the Pole Interface Homme-Machine Multimodale. It aims at clarifying the distinction between multimodal and multimedia systems and suggests a classification illustrated with current interactive systems.

2. Presentation of the Pole IHMM

The Pole IHMM (Interface Homme-Machine Multimodale, i.e. Multimodal Man-Machine Interface) is one of the research streams of PRC-CHM (Communication Homme-Machine, i.e. Man-Machine Communication) [11]. PRC-CHM is supported by the French government to stimulate scientific communication across French research laboratories in the domain of Man-Machine Interface. PRC-CHM is comprised of four "poles" (i.e research streams): speech recognition, natural language, computer vision, and, since fall 1990, multimodal man-machine interfaces.

Pole IHMM is concerned with the integration of multiple modalities such as speech, natural language, computer vision, and graphics [IHMM91]. The goal is two-fold:

- understand the adequacy of multimodality in terms of cognitive psychology and human factors principles and theory,

- identify software concepts, architecture, and tools for the development of multimodal user interfaces.

In order to focus research efforts on realistic goals, experimental multimodal platforms will be developed by interdisciplinary teams. Six such teams are currently designing a multimodal platform in Grenoble, Lyon, Nancy, Paris, and Toulouse:

- Grenoble: Multimodal user interface for a mobile robot.
- Lyon: Multimodal interaction and education.
- Nancy: Multimodal, multimedia workstations : application to the processing of composite documents.
- Paris: Creating and manipulating icons with a multimodal workstation; Speech recognition and talking head.
- Toulouse: Distributed multimodal system.

3. Multimedia User Interfaces

3.1. Definition

A media is a technical means which allows written, visual, or sonic information to be communicated among humans.

By extension, a multimedia system is a computer system able to acquire, deliver, memorize, and organize written, visual, and sonic information. In the domain of computer science:

- written material is not restricted to physical hard copies. It is extended to textual and static graphical information which is visually perceivable on a screen;
- visual material is usually associated with full motion video, more rarely with realistic graphical animations such as those produced in image synthesis;
- sonic information includes vocal or musical pre-recorded messages as well as messages produced with a sound or voice synthetizer.

3.2. Classification

Multimedia systems may be classified in two categories: first generation multimedia systems and full-fledged multimedia systems.

First generation multimedia systems are characterized by "internally produced" multimedia information. All of the information is made available from the standard hardware such as bitmap screen, sound synthetizer, keyboard and mouse. Such basic hardware has led to the development of a large number of tools such as user interface toolkits and user interface generators. With some rare exceptions such as Muse [10] and the Olivetti attempt [3], all of the development tools have put the emphasis on the graphical media. Apart the SonicFinder, a Macintosh finder which uses auditory icons [7], computer games have been the only applications to take advantage of non speech audio information.

Full-fledged multimedia systems are able to acquire non digitized information. The basic apparatus of first generation systems is now extended with microphones and CD technology. Fast compression/decompression algorithms such as JPEG [17] make it possible to memorize multimedia information. While multimedia technology is making significant progress, user interface toolkits and user interface generators keep struggling in the first generation area. Since the basic user interface software is unable to support the new technology, multimedia applications are developped on a case per case basis. Multimedia electronic mail is made available from Xerox PARC, NeXT and MicroSoft: a message may include text, graphics as well as voice annotations. FreeStyle from Wang, allows the user to insert gestural annotations which can be replayed at will. Authoring systems such as Guide, HyperCard and Authorware allow for the rapid prototyping of multimedia applications. Hypermedia systems are becoming common practice although navigation is still an unsolved problem.

To summarize, a multimedia computer system includes multimedia hardware to acquire, memorize and organize multimedia information. From the point of view of the user, a multimedia computer system is a sophisticated repository for multimedia information. At the opposite of multimodal computer systems, it ignores the semantics of the information it handles.

4. Multimodal User Interfaces

4.1. Definition

A modality may be the particular form used for rendering a thought, the way an action is performed. In linguistics, one makes a distinction between the content and the attitude of the locutor with regard to the content. For example, the content "workshop, interesting" may be expressed using different modalities such as: "I wish the workshop were interesting"; "The workshop must be interesting"; "The workshop will be interesting". In addition to these linguistic modalities one must consider the important role played by intonation and gesture. Thus human to human communication is naturally multimodal.

By extension, a computer system is multimodal if it is able to support human modalities such as gesture, written or spoken natural language. As a result:

- a multimodal system must be equipped with hardware to acquire and render multimodal expressions in "real time", that is, with a response time compatible with the user's expectations,

- it must be able to choose the appropriate modality for outputs,

- it must be able to understand multimodal input expressions.

4.2. Classification

Current practices in multimodal user interfaces lead to the following taxonomy: exclusive, and synergic multimodal user interfaces. In addition to the modality per se, we need to consider the effect of concurrency.

A user interface is *exclusive mudimodal* if:
- multiple modalities are available to the user, and
- an input (or output) expression is built up from one modality only.

An input expression is a "sentence"
- produced by the user through physical input devices, and
- meaningful for the system. In particular, a command is a sentence.

As an example of exclusive multimodal user interface, we can imagine the situation where, to open a window, the user can choose among double-clicking an icon, using a keyboard shortcut, or say "open window". One can observe the redundancy of the means for specifying input expressions but, at a given time, an input expression uses one modality only.

Xspeak [16] extends the usual mouse-keyboard facilities with voice recognition. Vocal input expressions are automatically translated into the formalism used by X window [15]. Xspeak is an exclusive multimodal system: the user can choose one and only one modality among the mouse, keyboard and speech to formulate a command. In Grenoble, we have used Voice Navigator [Articulate 90] to extend the Macintosh Finder to an exclusive multimodal Finder. Similarly, Glove-Talk [6] is able to translate gesture acquired with a data glove into speech (synthesis). Eye trackers are also used to acquire eye movements and interpret them as commands. Although spectacular, these systems are by no means exclusive multimodal only.

A user interface is *synergic multimodal* if:
- multiple modalities are available to the user, and
- an input (or output) expression is built up from multiple modalities.

For example, the user of a graphics editor such as ICP-Draw [18] and Talk and Draw [14], can say *put that there* while pointing at the object to be moved and showing the location of the destination with the mouse or a data glove. In this formulation, the input expression involves the synergy of two modalities. Speech events, such as *that* and *there*, call for complementary input events, such as mouse clicks and/or data glove events, interpretable as pointing commands.

Clearly, multimodal events must be linked through temporal relationships. For example, in Talk and Draw, the speech recognizer sends an ASCII text string to Gerbal, the graphical-verbal manager. The graphics handler time-stamps high level graphics events (e.g. the identification of selected objects along with domain dependent attributes), and registers them into a blackboard. On receipt of a message from the speech recognizer, Gerbal waits for a small period of time (roughly one-half second), then asks the blackboard for the graphical events that occurred after the speech utterance has completed. Graphical events that do not pertain to a window of time are discarded. It results from this observation that windowing systems which do not time-stamp events are the wrong candidates for implementing synergic multimodal platforms.

One important feature in user interface design is concurrency. Concurrency makes it possible for the user to perform multiple physical actions simultaneously, to carry multiple tasks in parallel (multithread dialogues), to allow the functional core and the user interface to perform computations asynchronously. In our case of interest:

- *concurrency in exclusive multimodal* user interfaces allows the user to produce multiple input expressions simultaneously, each expression being built from one modality only. For example, it would be possible for the user to say "open window" while closing another one with the mouse;

- *concurrency is necessary to synergic multimodal* user interfaces since, by definition, the user may use multiple channels of communication simultaneously. The absence of concurency would result in a strict ordering with conscious pauses when switching between modalities. For example, the specification of the expression *put that there* would require the user to say *put that*, then *click*, then utter *there*, then *click*.

4.3. Voice-Paint, synergic multimodal system

We have developed Voice-Paint [8], a first experience in integrating voice and graphics modalities based on our multiagent architecture, PAC [5, 2]. Conceptually, it is a very simple extension of events as managed by windowing systems. Agents, which used to express their interest in graphics events only, can now express their interest in voice events. As graphics events are typed, so are voice events. Events are dispatched to agents according to their interest.

We have applied this very simple model to the implementation of a Voice-Paint editor on the Macintosh using Voice Navigator [1], a word-based speech recognizer board: as the user draws a picture with the mouse, the system can be told to change the attributes of the graphics context (e.g. change the foreground or background colors, change the thickness of the pen or the filling pattern, etc.). Our toy example is similar in spirit to the graphics editor used by Ralph Hill to demonstrate how Sassafras is able to support concurrency for direct manipulation user interfaces [9].

Voice-Paint illustrates a rather limited case of multimodal user interface: concurrency at the input level. This is facilitated by Voice Navigator whose unit of communication is a "word". From the user's point of view, a word may be any sentence. For Voice Navigator, pre-recorded sentences are gathered into a data base of patterns. At run time, these patterns are matched which the user's utterances. The combination of Voice Navigator and graphics events into high level abstractions (such as a command) does not require a complex model of the dialogue. Thus, Voice-Paint does not demonstrate the integration of multiple modalities at the higher level of abstractions. This work is precisely the research topic of Pole IHMM.

5. Summary

Multimedia and multimodal user interfaces use similar physical input and ouput devices. Both acquire, maintain and deliver visual and sonic information. Although similar at the surface level, they serve distinct purposes:

- a multimedia system is a repository of information produced by multiple communication techniques (the medias). It is an information manager which provides the user with an environment for organizing, creating and manipulating multimedia information. As such, it has no semantic knowledge of the information it handles. Instead, data is encapsulated into typed chunks which constitute the units of manipulation (e.g. creation, deletion, and, in the particular case of hypermedia systems [4], linkage between chunks). Chunk contents are ignored by the system;

- a multimodal system is supposed to have the competence of a human interlocutor. At the opposite of multimedia systems, a multimodal system analyzes the content of the chunks produced by the environment in order to discover a meaning. Conversely, it is able to produce multimodal output expressions that are meaningful to the user.

In the current state of the art, one can identify another distinctive feature between multimedia and multimodal systems: multimedia information is the subject of the task (it is manipulated by the user) whereas multimodal information is used to control the task. With the progress of concepts and techniques, this distinctive usage will grow blurred over time.

So far, we have tried to clarify the distinction between multimedia and multimodal systems, and we have proposed a classification for multimodal user interfaces. We need now to analyze the implication of multimodality on software architectures.

6. Conclusion

This article mentions a first step experience aimed at the implementation of synergic multimodal user interfaces. It does not claim any ready-for-use solutions. Instead, it presents a possible framework for bundling multiple modalities into a consistent organization. Our first experimental results encourage us to extend our expertise in multiagent architectures for GUI's to multimodal user interfaces.

7. References

[1] Articulate systems inc.: The Voice Navigator Developer Toolkit; Articulate Systems Inc., 99 Erie Street Cambridge, Massachusetts, USA, 1990.

[2] L. Bass, J. Coutaz: Developing Software for the User Interface; Addison Wesley, 1991.

[3] C. Binding, S. Schmandt, K. Lantz, M. Arons: Workstation audio and window based graphics: similarities and differences; Proceedings of the 2nd Working Conference IFIP WG2.7, Napa Valley, 1989, pp. 120-132.

[4] J. Conklin: Hypertext, an Introduction and Survey; IEEE Computer, 20(9), September, 1987, 17-41.

[5] J. Coutaz: PAC, an Implemention Model for Dialog Design; Interact'87, Stuttgart, September, 1987, pp. 431-436.

[6] S.S. Fels: Building Adaptative Interfaces with Neural Networks: the Glove-Talk Pilot Study; University of Toronto, Technical Report, CRG-TR-90-1, February, 1990.

[7] W. W. Gaver: Auditory Icons: Using Sound in Computer Interfaces; Human Computer Interaction, Lauwrence Erlbaum Ass. Publ. , Vol. 2, 1986, 167-177.

[8] A. Gourdol: Architecture des Interfaces Homme-Machine Multimodales; DEA informatique, Université Joseph Fourier, Grenoble, June, 1991.

[9] R.D. Hill: Supporting Concurrency, Communication and Synchronization dans Human-Computer Interaction-The Sassafras UIMS; ACM Transactions on Graphics 5(2), April, 1986, pp. 179-210.

[10] M. E. Hodges, R.M. Sasnett, M.S. Ackerman: A Construction Set for Multimedia Applications; IEEE Software, January, 1989, pp. 37-43.

[11] Pôle Interface Homme-Machine Multimodale du PRC Communication Homme-Machine, J. Caelen, J. Coutaz eds., January,1991.

[12] M. W. Krueger, T. Gionffrido, K. HINRICHSEN: Videoplace, An Artificial Reality; CHI'85 Proceedings, ACM publ., April, 1985, 35-40.

[13] M. W. Krueger: Artificial Reality II; Addison-Wesley Publ., 1990.

[14] M. W. Salisbury, J. H. Hendrickson, T. L. Lammers, C. Fu, S. A. Moody: Talk and Draw: Bundling Speech and Graphics; IEEE Computer, 23(8), August, 1990, 59-65.

[15] R.W. Scheifler, J. Gettys: The X Window System; ACM Transaction on Graphics, 5(2), April, 1986, 79-109.

[16] C. Schmandt, M. S. Ackerman, D. Hndus: Augmenting a Window System with Speech Input; IEEE Computer, 23(8), August, 1990, 50-58.

[17] G.K. Wallace: The JPEG Still Picture Compression Standard for Multimedia Applications; CACM, Vol. 34, No.4, April, 1991, pp. 30-44.

[18] J. Wretö, J. Caelen: ICP-DRAW, rapport final du projet ESPRIT MULTIWORKS no 2105.