Multimedia and Multimodal User Interfaces: A Taxonomy for Software Engineering Research Issues

JOËLLE COUTAZ

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

Abstract: This article aims at clarifying the distinction between multimodal and multimedia computer systems. A dimension space is proposed that accounts for a classification of such systems as well as for identifying the implications from the software architecture point of view. The discussion is illustrated with the analysis of current multimedia and multimodal systems and points out some useful areas for future research such as the fusion of modalities at multiple levels of abstraction.

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 [1] 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 [2]. 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 our early analysis and experience with the integration of multiple modalities of communication between a user and an interactive system. Next section discusses the notions of media and modality in their broad sense. In section 3, a dimension space is proposed that accounts for a classification of multimedia and multimodal computer systems. The discussion is illustrated with a number of current such systems. In section 4, the dimension space serves as a basis for analyzing the implications of multimodality on software design.

2. Media and Modality

Before considering the notions of multimedia and multimodality, we need to clarify the terms media and modality as used in their broad sense. We will then consider these terms in the domain of computer human interaction.

2.1. Media

A media is a technical means that allows the expression and the communication of information [3, vol.10, pp.6788]. It is a vehicle that conveys specific types of information which in turn, require specific types of communication channels. A communication channel is defined by the effective connection of a transmitter with a receiver both dealing with compatible types of information. These types determine the type of the communication channel. When considering the human being, receivers are assimilated to the sensory system whereas transmitters are implemented by the effector system.

Information types conveyed by artificial media such as computers, include text, graphics, still images, music, voice, video, etc. To be communicated successfully, text, graphics, and still images require a visual channel, that is the connection of a visual transmitter (such as the screen of the computer) with a visual receiver (such as the vision system of a human being); music and voice involve audio channels, while video may require both audio and visual apparatus. From the perspective of the human being, the first group of information types are classified under the more general category: visual information, whereas the second one is known as sonic information. They respectively require a visual and an audio communication channel.

The video example indicates that a useful distinction should be made between information that use a single communication channel from those which require multiple channels. In addition, one channel may carry mixed information types. For example, a soundless video can show the still image of a text mixed with a graphical representation such as a pie chart. Alternatively, it may contain a live image shot from the real world in which graphical information may be cropped. In these two examples, information types are mixed but they are transmitted through a unique channel: the visual communication channel. Figure 1 summarizes our analysis of the notion of media.

As an example, we may consider the results of the XVIth Winter Olympic Games in Albertville. These have been delivered using multiple medias[1] such as television and newspapers, carrying different types of information such as live videos, textual comments, illustrative pictures, and statistical graphics. Each of them was used to express the same informational content (i.e., the results of the olympic games) triggering different communication channels (e.g., the audio and visual channels), acting for the olympic games observers as different, possibly complementary, instruments for reasoning about the substance of the information.

Although the notion of "communication channel" provides a coarse dimension space for characterizing medias, it is sufficient for the purpose of our discussion. As a refinement exercise, one could follow Gaver's suggestion and consider medias in terms of the affordances they make available [4]. An affordance expresses the directness between the intrinsic, possibly perceivable properties of an artifact and the actions it makes possible and obvious [5].

Let's consider the 5 general categories of information that can be received and transmitted by a human being:
- V = {g, t, p}, the set of visual types: g (for graphics), t (for text), p (for pictures)
- S = {m, v ...}, the set of sonic types, e.g., music, voice
- T = set of tactile types
- O = set of olfactive types
- G = set of gesture types
- V* = {g, t, p, g+t, g+p, t+p, g+t+p}, the set of combinations of types in V, (+ denotes type mixing)
- S* = set of combinations of types in S
- T* = set of combinations of types in T
- O* = set of combinations of types in O
- G* = set of combinations of types in G

A media M is defined as a non empty set of communication channels C_i, each channel being defined as a triple:
$M = \{C_i\}, 1 \leq i \leq 5$
$C_i = <t_i, Tr_{ti}, Rc_{ti}>$
$t_i \in V^* \cup S^* \cup T^* \cup O^* \cup G^*$ ($\cup$ denotes the union operator) and $\forall\ C_i, C_j \in M, t_i \neq t_j$
Tr_{ti}, Rc_{ti} are respectively the transmitter and receiver specialized in processing information of type t_i.

Figure 1. An attempt to formalize the notion of media.

Whereas "media" has a rather well-targetted meaning as a vehicle for conveying information, the term "modality" covers multiple notions. I have selected two of them that can be conveniently related to the notion of media: the psychological perspective and the general sense.

[1] In French, media is a neutral masculine noun (média) whose plural is médias. Although the American Heritage Dictionary of the English Language published in 1970 explicitly comdemns the usage of medias, I will follow the French convention.

2.2. Modality

In psychology, the notion of modality refers to the sensory categories such as vision, audition, olfaction, touch [3, vol.10, pp.7002]. In the theoretical framework of the Model Human Processor [6] as well as in ICS [7], these categories are modelled as specialized processors: the sensory subsystems process incoming sense data in their domain (e.g., vision, acoustics) whereas effector subsystems control motor output (e.g., articulation, limbs).

This definition of modality reveals a straightforward relationship with the notion of media: a modality denotes a type of human communication channel. For example, when an operator looks at a computer screen, the visual communication channel is activated. The user receives visual information through a receiver of the visual subsystem. He is also said to use the vision modality.

Modality considered in its broad sense is the way an idea is expressed or perceived, or the manner an action is performed. It may be viewed as an information modifier or an amplifier that denotes the engagement or the position of the issuer with regard to the informational content. For example, the content "olympic games results - availability", may be expressed using different linguistic modalities such as: "I wish the olympic games results were available!", or "the olympic games results will be available".

Modality is made explicit in many ways such as markers and features. For example, interjection and verb tense (e.g., conditional, present) are typical linguistic markers. In oral verbalisation, intonation has a linguistic modal function as in “it snows?” (interrogative intoneme) and “it snows!” (assertive intoneme). In computer vision, perceptual features such as highlight, shading, motion and texture are considered as modalities. This means that there exists a processor specialized for each kind of perceptual feature, each one operating on the same raw signal (i.e., the image). In hand-written text, pen pressure and character shapes are modalities which, as well as linguistic modalities, capture the writer’s engagement.

It results from our analysis that modality refers to the manner a human being communicates or acquires information. This manner concerns the type of communication channels used and affects the expression of the information tranferred over these channels.

Having clarified the notions of media and modality, we need now to discuss multimedia and multimodality in computer systems.

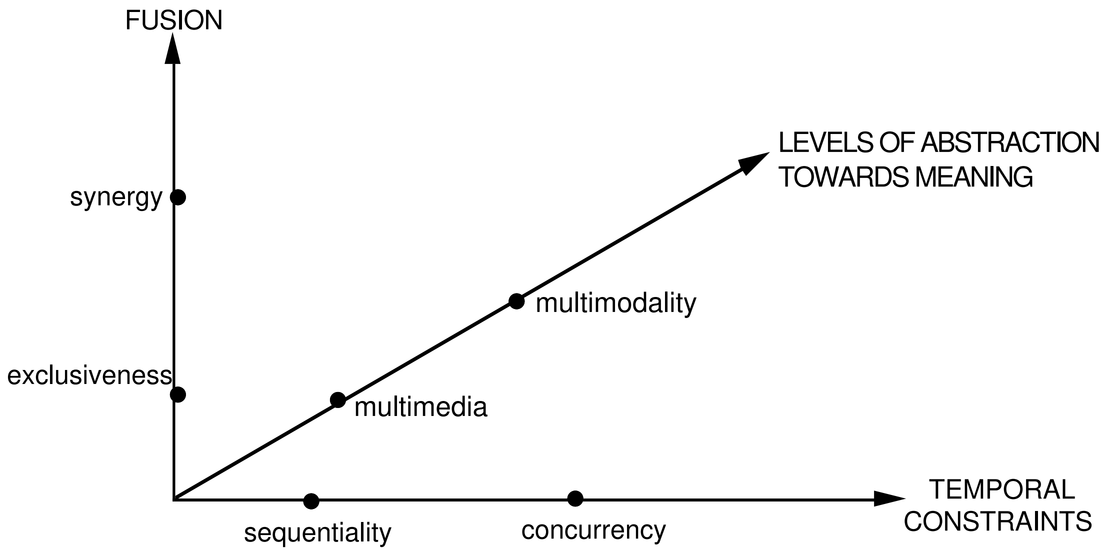

Figure 2. The Dimension Space.

3. Multimedia and Multimodal Computer Systems: a Dimension Space

To a first approximation, a multimedia computer system can be defined as a system capable of supporting multiple medias. Similarly, a multimodal system is able to communicate with a user along multiple modalities, media and modality covering the notions discussed above. Although straighforward, these definitions do not capture the subtleties of the domain. Therefore I propose a dimension space useful for classifying current and future multimedia and multimodal systems as well as for making more precise the problem space for software designers.

The dimension space includes three axes: levels of abstraction, fusion, and time constraints. Figure 2 illustrates the possible values along each dimension.

3.1. Levels of abstraction

Incoming data on a receiver may be processed at multiple levels of abstraction. As an illustration, consider the layers of abstraction involved in a computer vision system [8] (Figure 3).

At the lowest, most basic level we find the image or signal, as generated by the acquisition process. This information is then described at a second level as geometric information expressed in a coordinate space which is inherent to the signal. The third level in this hierarchy is typically concerned with object geometry. Although geometric in nature, the coordinate system for this level switches from 2-D image coordinates to a 3-D object based coordinates. An object based coordinate space makes it possible to integrate information from different viewing positions. Viewpoint invariance then leads to such fundamental concepts as "object constancy". It becomes possible to "see beyond" the visual stimulus to an imagined (internal) world of individual objects. However, to associate the perception with known objects, the geometric information in both the image description and in the object description must be grouped to form cues. The idea of individual objects leads to the final, most abstract level. The interpretation level expresses perceptions as symbolic information. Symbols permit perceptions to be associated to previous experiences. Association with known "objects" leads to expectations concerning nearby objects as well as expectations of behaviors.

Speech recognition as well as speech synthesis systems operate along similar principles. The important point is that information is represented and processed at multiple levels of abstraction. This transformation process makes possible the extraction of meaning from raw data and conversely the production of perceivable information from symbolic abstract representations.

Multimedia systems are limited to the lowest level of abstraction. Although pictures are modified by compression algorithms, the basic concept is still a pixel. Raw data is not processed to obtain meaning. At the most, it is encapsulated as chunks of information (e.g., a sequence of still images). In hypermedia systems, meaning is modelled as an add-on, a meta-information sitting on top of the chunks. Meaning is not automatically built from raw data. Similar observations hold for sound and text.

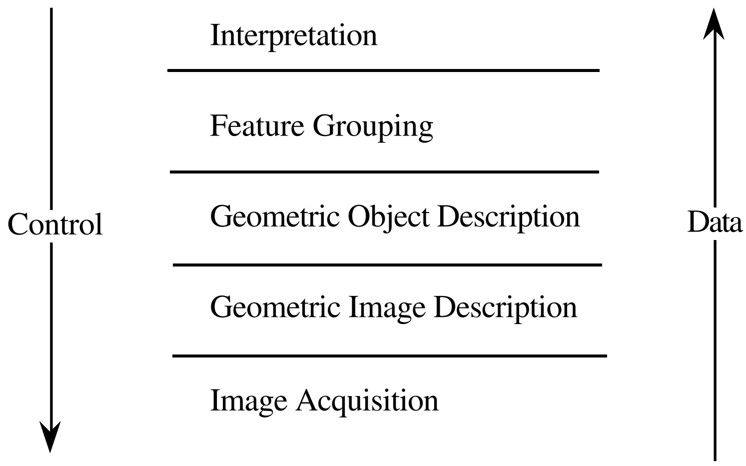

Figure 3. Levels of abstractions in a computer vision system.

At the opposite, multimodal systems strive for meaning. They include all of the levels of abstraction from raw data to symbolic representations on which reasoning may be performed to augment the success of the cooperation between the user and the system. Coming back to the computer vision example, a high level representation of the face and the hands will allow the system to recognize the user, interpret hand gestures or face expressions. Without falling in the anthropomorphism pit, one may find an analogy between the levels of abstractions in a multimodal system and the ICS theoretical psychological model.

In ICS, the human information processing system is subdivided into a set of specialized subsystems. The sensory subsystems transform sense data into specific mental codes that represent the structure and content of the incoming data. These representations are then handled by subsystems that are specialized in the processing of higher-level representations: the morphonolexical subsystem for processing the surface structure of language, the object subsystem for processing visuospatial structures, and the propositionnal and implicational subsystems for more abstract and conceptual representations. The ouput of these higher-level subsystems are directed to the effector subsystems (articulatory and limb).

In summary, multimedia systems do not detect meaning automatically whereas multimodal system do. In a multimedia system: (a) conceptual representations are encoded as meta information, (b) the information per se is encapsulated and is not processed. In a multimodal system, (a) abstract concepts are elaborated automatically from raw information and vice versa, and (b) abstract concepts are maintained dynamically during the communication process. In the following sections, we call:

- "information", unprocessed data exchanged between a system and a user, and
- "expression", data exchanged between a system and a user for which the system maintains a meaning.

3.2. Fusion

Fusion covers the combination of different types of information or different types of modalities. The absence of fusion is called *exclusiveness*, while the existence of fusion forms a *synergy*.

For a multimedia system, fusion is mixing types of information in the same communication channel. For example, cropping live images with graphics is a fusion of information types over the visual channel.

For a multimodal system, fusion concerns mixing modalities to build an input or an output expression. Thus, a multimodal user interface is

- *exclusive* if input (or output) expressions are built up from one modality only,
- *synergic* if input (or output) expressions are built up from multiple modalities.

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 ways for specifying input expressions but, at a given time, an input expression uses one modality only.

As an example of synergic multimodal system, the user of a graphics editor 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.

3.3. Time constraint

Time constraint expresses the absence or presence of parallelism at the interface. The absence is referred to as *sequentiality*, while the presence is called *concurrency*.

For a multimedia system, time constraints may be analyzed at multiple grains, namely: between communication channels, between chunks of information, and inside a chunk. Concurrency at the communication channel level expresses the possibility for the user and the system to communicate simultaneously through multiple channels (for example, hearing music and reading or typing text or watching a sound video). Sequentiality imposes one active channel at a time. Time constraints between chunks of information covers the synchronization between, for example a sequence of still images and a text: as the user scrolls the text, the image sequence is updated accordingly. Time constraints inside a chunk of information is used to express temporal relationships as in the animation of graphics objects.

In a multimodal system, time constraints express the possibility for the user to build exclusive or synergic expressions sequentially or concurrently. Sequentiality or concurrency may be studied at the input/output expression level as well as at finer grain such as the physical actions involved in the specification of expressions. Sequential exclusiveness implies that the user can build or receive at most one input or output expression at a time. Concurrent exclusiveness means that the user is able to build or receive multiple input/ouput expressions, each expression being expressed in a unique, possibly different, modality. As an example of concurrent exclusive multimodal user interface, the user may say "open file FooBar" while moving a file icon in the trash with the mouse.

Sequentiality in synergic multimodality means that the user builds/receives input/output expressions using multiple modalities but only one modality is used at a time: modalities are interleaved during the construction of an input/output expression. Although not desirable, sequential synergism may be imposed for technical reasons. If we consider the "put that there" example, sequentiality would require the user to say "put that" followed by a mouse click to denote "that". He would then say "there" and click a second time to indicate the destination. In a concurrent synergic multimodal system, the user would utter the sentence and perform the mouse clicks in a "natural" way, that is without necessarily respecting the interleaving between modalities. However, from the point of view of the software designer, modal events must be linked by some temporal relationship such as being part of the same "temporal window". This notion will be discussed further in section 4.

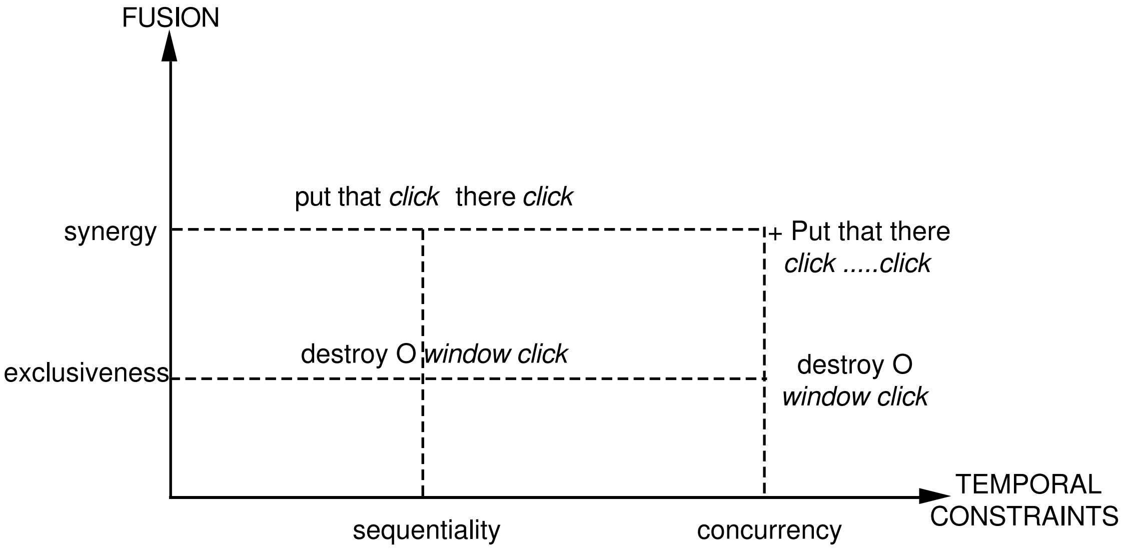

Figure 4. Examples of input expressions in multimodal user interfaces. Text appearing on one line denotes sequentiality, while text appearing on 2 lines denotes concurrency. Italic expresses mouse gesture. Normal text indicates spoken words.

Figure 4 shows examples of input expressions for multimodal user interfaces along the fusion and temporal dimensions.

3.4. Examples of Multimedia and Multimodal Computer Systems

This section provides an overview of the state of the art in the domain of multimedia and multimodal computer systems. It is intended to provide the reader with examples as elements of comparison. It does not claim to be exhaustive.

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 [9] and the Olivetti attempt [10], all of the development tools have put the emphasis on the graphics type of information. Apart the SonicFinder, a Macintosh finder which uses auditory icons [11], 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 and MPEG [12] make it possible to memorize sound, images. 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. Note that these systems are not multimodal: voice and gesture annotations are recorded but not processed to discover meaning. Authoring systems such as Guide, HyperCard and Authorware allow for the rapid prototyping of multimedia applications. Hypermedia systems are becoming common practice [13] although software architecture issues are still unclear.

Xspeak [14] 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. Concurrency is supported by the underlying platform: X window/Unix. In Grenoble, we have used Voice Navigator [16] to extend the Macintosh Finder to an exclusive multimodal Finder. Similarly, Glove-Talk [17] 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.

As examples of synergic multimodal systems, ICP-Draw [18] and Talk and Draw [19] are graphics editors that support the “put that there” paradigm. Although synergic they do not fully support concurrency. In particular, fusion in Talk and Draw is speech driven: deictic mouse events must happen after the utterance of the sentence. We will come back to this example in Section 4. CUBRICON [20] on the other hand, as well as HITS [21], support concurrent synergic multimodal interaction.

CUBRICON accepts coordinated simultaneous natural language and pointing via a mouse device. The user can input natural language via the speech device and/or the keyboard. Speech recognition is handled by a Dragon System VoiceScribe which supports discrete speech only. Although non continuous speech is unnatural to the user, it greatly simplifies the problem of fusion. In order to solve multimodal references, CUBRICON includes three types of knowledge: (1) the domain knowledge modelled as a semantic network of the concepts manipulated in the task (i.e., military tactical air control), (2) the “dual-media” language defined by a grammar which specifies where pointing gestures can be used in place of noun and locative adverbial phrases, (3) the discourse model which maintains a representation of the “focus space”, that is an history of entities and propositions expressed by the user. Thus, for example, when the user asks “What is the name of this *icon click* airbase?”, the object corresponding to the icon is retrieved from the focus space. The name of the particular instance can then be retrieved from the domain knowledge.

HITS is a development environment, i.e., an integrated set of tools to support the construction and run-time execution of multimodal computer systems [21]. The run-time kernel is based on a blackboard architecture as well as on an integrated knowledge-base, e.g., knowledge about the modalities supported in the user interface, domain knowledge, etc. The HITS blackboard permits the fusion of multiple modalities as well as the maintenance of the history.

Having set up a clear distinction between multimedia and multimodal user interfaces, we are now able to discuss the implications for software architecture.

4. Software Architecture Issues

Despite the qualificative “graphical” in the expression “Graphical User Interface”, knowledge in GUI technology, driven by direct manipulation user interfaces, covers multimedia interaction. Although there are still technical performance problems due to the volume of data to be exchanged, software architecture for multimedia systems can be designed or assessed by using models from the GUI technology. Thus, in this section the discussion focuses on how to extend GUI models to cope with multimodal interaction.

4.1. GUI's and other modalities

Computer vision systems as well as natural language (NL) processing systems are all based on a framework similar to that of multi-agent architectures used in GUI's. As shown in Figure 5, stimuli (or events) are received through physical input devices. The functional core, which implements domain dependent concepts, sits at the other end of the spectrum. Between these two extremes, a number of layers of abstractions transform input stimuli into semantic abstractions understood by the functional core. These layers are animated by cooperating agents such as in the PAC model [22] and each layer has a specific role as indentified in the Arch Model [23] or the PAC-AMODEUS model [24].

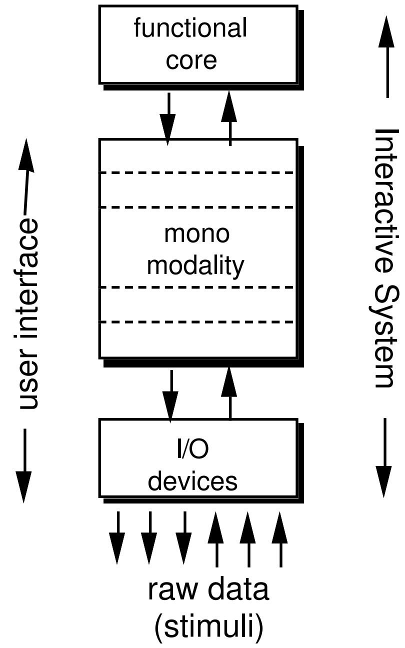

Figure 5. Framework for monomodal interaction.

Although one can observe similitudes in the way problems are tackled in the various disciplines, layers in NL and computer vision involve sophisticated modelling techniques which, in GUI's, are either unnecessary or avoided. Indeed, understanding NL expressions requires the computer system to perform two types of complementary activities:

- translate input expressions into an internal abstract model which represents the semantics of the expressions,

- maintain a dynamic model of the interaction including domain modelling, task modelling, and user modelling. This model sets the foundations for reasoning about the most recent input expressions. One class of such reasoning deals with uncertainty and ambiguity such as those inherent to anaphoras and ellipses.

An anaphoric reference is a partial description (denotation) of a concept which has been previously mentioned in the discourse. For example, in the following conversation: "Do you have a message from FooBar? Delete it", *it*, which refers to the message from FooBar, is a anaphoric phenomenon. An ellipsis is an omission of a syntactic construct without necessarily loosing the meaning of the sentence. For example, in "Do you have a message from Foo? And from Bar?" *and from Bar*, from which the syntactic construct *do you have a message from* has been eliminated, is an ellipsis.

Clearly, anaphoras and ellipses, which both refer to past expressions, cannot be solved without a dynamic model of the interaction. For example, in Partner, a speech understanding system, a history mechanism is used to process anaphoras, and the user's goals are traced to formulate hypotheses for solving ellipses [25]. At the opposite of natural language user interfaces, traditional GUIs do not maintain any dynamic model of the interaction, or if they do, the model is poor and implicit. For example, the notions of "current selected object" and "current cut buffer", which avoid the user to specify an operand to a command, support limited cases of ellipses. Similarly, the notion of "currently available commands", which structures the task space, supports context modelling in a very crude way.

In addition to linguistic problems, "human modalities" techniques are unreliable: raw data are noisy and recognition is flawful. In order to enhance recognition, models are maintained at multiple levels of abstractions, each model being used to update confidence factors in adjacent layers. These techniques can be observed in speech recognition as well as in computer vision. In GUI's, on the other hand, raw data, such as mouse clicks and keyboard events, are deterministic. When ambiguity arises as for overlapping objects in graphical editors, an ordering policy is imposed on the user.

4.2. Software Framework for Multimodality

Although similar principles of software decomposition are exploited in all of the HCI disciplines, each modality has its own system of representation. This is all right for exclusive multimodal user interfaces provided that each representation technique is able to communicate with the functional core in the formalism of that core. On the other hand, the implementation of synergic multimodal user interfaces requires fusion, thus the integration of the various modelling techniques into a uniform framework. This combination may be performed along three axes:

- keep the current modelling techniques as they are and define appropriate gateways between them;
- define a unique abstract representation onto which different modelling techniques can be hooked;
- use a medium range approach which keeps some elements of the current modelling techniques and introduce layers for integration.

In the short term, we think that the third approach is a reasonable way to go. As illustrated in Figure 6, the two extreme layers, the interface with the functional core and the event layers, are good candidates for integration. The intermediate layers, which implement highly specialized concepts and know-how, should be kept unchanged. However, a common model of the dialogue (e.g., history, user model, task model) should be shared by all of the modalities through specific gateways.

Modelling the dialogue should be performed at a high level of abstractions, i.e., at the same level as the interface with the functional core. As discussed above, dialogue modelling in GUI is very poor or inexistent whereas dialogue modelling in human modalities is unavoidable. However, the representation techniques are too much biased by the modality they serve. We need to jump one step further and model the dialogue at a more psychological level.

At the other extreme, 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 the time window are discarded.

It results from this observation that windowing systems that do not time-stamp events are the wrong candidates for implementing synergic multimodal user interfaces. Similarly, speech systems that do not provide any facility for setting up the granularity of speech events (e.g., word level rather than sentence level) may impede the fusion of events at low level of abstraction such as the morphological level [26].

We have developed a first experience in integrating voice and graphics modalities at the event level. 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, 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.) [27]. Our toy example is similar in spirit to the graphics editor used by Ralph Hill [28] to demonstrate how Sassafras is able to support concurrency for direct manipulation user interfaces.

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 topic will be developed in the framework of a new project launched in the French scientific community (i.e., Pôle Interface Homme-Machine Multimodale of PRC Communication Homme-Machine [29] and will be formalized and refined in the continuing ESPRIT BRA project AMODEUS2.

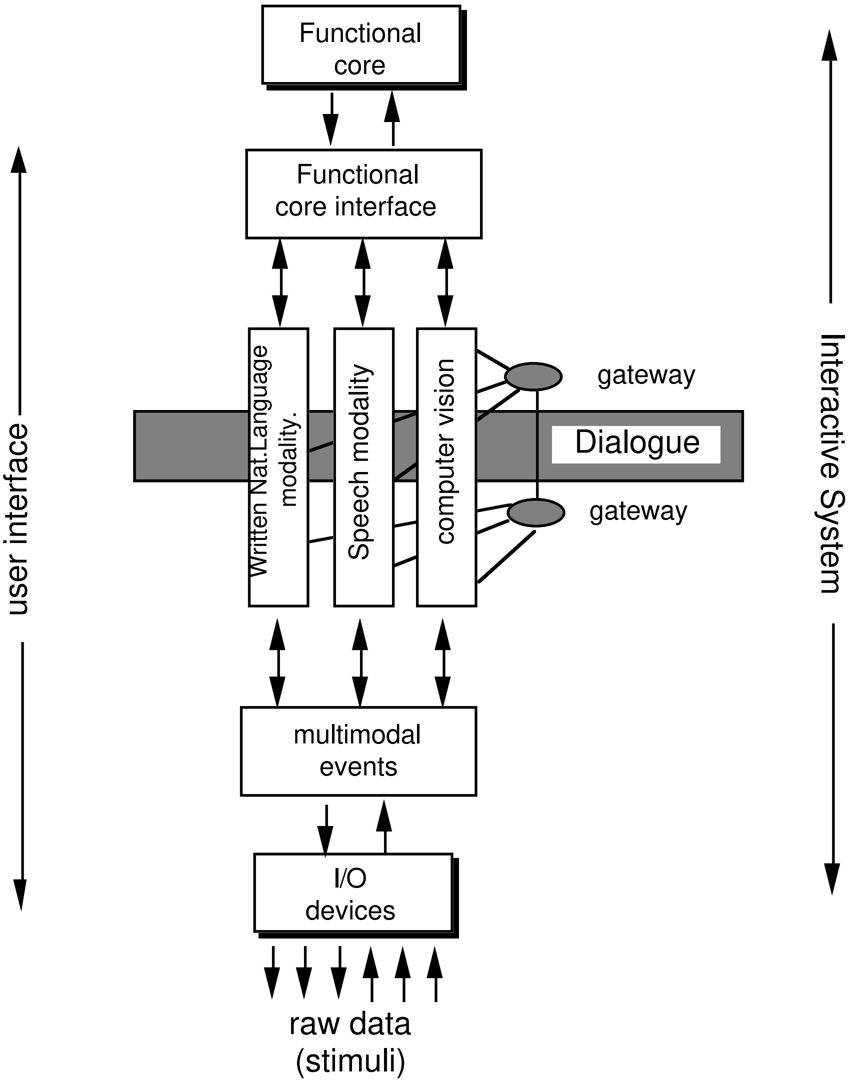

Figure 6. A framework for integrating multiple modalities of interaction.

5. Summary

In summary, a media is defined as a non empty set of typed communication channels, each channel being essentially defined by the category of information it transfers (e.g., visual, sonic, gestural, etc.). While a media is a technical vehicle, modality denotes the manner a human being communicates or deliver information. This manner concerns the type of communication channels used and affects the expression of the information transferred over these channels.

Multimedia and multimodal computer systems both support several types of communication channels. Both support information fusion and both of them are concerned with temporal constraints. They differ however by their capacity for abstraction: whereas multimodal systems automatically extract meaning from raw information, multimedia systems encapsulate raw data into chunks of information on top of which meaning is modelled by hand. This difference in power for abstraction has direct consequences for the end-user as well as for the software designer. For the software designer, a dynamic model of the cooperation must be modelled based on a representation of meaning that is common to all modalities. Although this fundamental issue has been tackled in HITS and MMI2 [30], it is still an opened question.

Acknowledgement: This article was influenced by stimulating discussions with my PhD students, Laurence Nigay, Arno Gourdol, Daniel Salber, as well as with my colleagues of the pole IHM-Multimodal such as Jean Caelen and Marie-Luce Bourguet.

Reference

[1] M. W. Krueger, T. Gionffrido, K. Hinrichsen, “Videoplace, An Artificial Reality”, CHI'85 Proceedings, ACM publ., April, 1985, 35-40.

[2] M. W. Krueger, “Artificial Reality II", Addison-Wesley Publ., 1990.

[3] Grand Dictionnaire Encyclopédique, Larousse, Volume 10, 1986.

[4] W. W. Gaver, “Technology Affordances", Human Factors in Computing Systems, CHI'91 Proceedings, ACM Press publ., April, 1991, 79-84.

[5] D. A. Norman, “The psychology of everyday things", Basic Books, New York, 1988.

[6] S. Card, T. Moran, A. Newell, “The Psychology of Human Computer Interaction", Lawrence Erlbaum Ass. Publ., Hillsdale, N.J., 1983.

[7] P. Barnard, “Cognitive Resources and the Learning of Computer Dialogs", in Interfacing Thought, Cognitive aspects of Human Computer Interaction, J.M. Carroll Ed., MIT Press Publ., pp. 112-158.

[8] J.L. Crowley, "Knowledge, Symbolic Reasoning and Perception", Intelligent Autonomous Systems II, Amsterdam, December, 1989.

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

[10] 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.

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

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

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

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

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

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

[17] 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.

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

[19] 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.

[20] J. Neal, C. Thielman, K. Bettinger, J. Byoun, "Multi-modal References in Human-Computer Dialogue", Proceedings of AAAI-88, 1988, pp. 819-823.

[21] J. Hollan, E. Rich, W. Hill, D. Wroblewski, W. Wilner, K. Wittenburg, J. Grudin, "An Introduction to HITS: Human Interface Tool Suite. In Intelligent User Interfaces, J. W. Sullivan and S.W. Tyler Eds., ACM Press, 1991, pp. 293-337.

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

[23] L. Bass, R. Little, R. Pellegrino, S. Reed, R. Seacord, S. Sheppard, M. Szezur, "The Arch Model: Seeheim Revisited", User Interface Developers,Workshop, 1991.

[24] L. Nigay, J. Coutaz, "Building user interfaces : Organizing software agents", Conference ESPRIT'91, Bruxelles, november, 1991.

[25] J.-M. Pierrel, J. Caelen, "Compréhension de la parole: Méthodes et Applications", 12èmes journées francophones sur l'informatique, La Machine Perceptive, EC2 Ed., Grenoble, Janvier, 1990, pp. 53-103.

[26] M.L. Bourguet, J. Caelen, "Multimodal Human-Computer Interfaces: event management and interpretation based on inferred user's intentions", submitted to IFIP WG2.7 Working Conference, Engineering for Human-Computer Interaction, Finland, August, 1992.

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

[28] 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.

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

[30] M. Wilson, The first MMI2 Demonstrator, "A Multimodal Interface for Man Machine Interaction with Knowledge Based Systems", Deliverable D7, ESPRIT project 2474 MMI2, Tech. report Rutherford Appleton Laboratory, Chilton Didcot Oxon OX11 0QX, RAL-91-093, 1991.