# Modeling Multimodal Human-Computer Interaction

**Incorporating the well-known Unified Modeling Language into a generic modeling framework makes research on multimodal human-computer interaction accessible to a wide range of software engineers.**

*Zeljko Obrenovic*
*Dusan Starcevic*
University of Belgrade

Multimodal interaction is part of everyday human discourse: We speak, move, gesture, and shift our gaze in an effective flow of communication. Recent initiatives such as perceptual[1] and attentive user interfaces[2] put these natural human behaviors in the center of the human-computer interaction (HCI). By letting our highly skilled and coordinated human communicative behavior control interactions with a system more transparently than ever before, these interfaces improve accessibility for diverse users and usage contexts, advancing the performance stability, robustness, expressive power, and efficiency of communication.

To improve coverage, reliability, and usability, researchers are designing new multimodal interfaces that automatically learn and adapt to important user, task, and environmental parameters.[3] Designing such interfaces is a challenging task, however. As the "Multimodal Interaction" sidebar describes, although many approaches exist, sound and practical solutions for developing multimodal systems are still lacking. However, trends and industry standards in the software engineering community introduce new possibilities for improving the analysis, design, and implementation of multimodal systems.

We've designed a generic modeling framework for specifying multimodal HCI using the Object Management Group's Unified Modeling Language.[4] Because it's a well-known and widely supported standard—computer science departments typically cover it in undergraduate courses, and many books, training courses, and tools support it—UML makes it easier for software engineers unfamiliar with multimodal research to apply HCI knowledge, resulting in broader and more practical effects. Standardization provides a significant driving force for further progress because it codifies best practices, enables and encourages reuse, and facilitates interworking between complementary tools.[5]

## MULTIMODAL HCI MODELING FRAMEWORK

Model-driven software development, which emphasizes models rather than computer programs,[6] inspired our approach. Following a model-driven approach lets us use concepts that are much less bound to the underlying implementation technology and much closer to the problem domain than conventional programming environments.[5] Modeling at different levels of abstraction clearly benefits multimodal systems, as implementation platforms typically use primitives that are far removed from concepts such as modality or multimodal integration.

Rather than specifying an interaction modality such as speech, gestures, or graphics, our framework defines a generic approach for modeling such modalities. The model, therefore, focuses on the notion of an abstract modality, which defines the common characteristics of HCI modalities regardless of their specific manifestations.

To create a generic framework for modeling multimodal HCI, we explored two problems:

0018-9162/04/$20.00 © 2004 IEEE

## Multimodal Interaction

As they emerge, multimodal interfaces are moving the balance of interaction closer to the human and offering expressive, transparent, efficient, and robust human-computer interaction.

In computer sciences, the meaning of the term "modality" is ambiguous. In human-computer interaction, the term usually refers to the human senses—vision, hearing, touch, smell, and taste—but many researchers distinguish between computing modalities and the sensory modalities of psychology.[1]

Sharon Oviatt offered a more practical definition, saying that multimodal systems coordinate the processing of combined natural input modalities—such as speech, touch, hand gestures, eye gaze, and head and body movements—with multimedia system output.[2] Matthew Turk and George Robertson further refined the difference between multimedia and multimodal systems, saying that multimedia research focuses on the media, while multimodal research focuses on human perceptual channels.[3] They added that multimodal output uses different modalities, such as visual display, audio, and tactile feedback, to engage human perceptual, cognitive, and communication skills in understanding what is being presented. Multimodal interaction systems can use various modalities independently, simultaneously, or by tightly coupling them.

Starting with Richard Bolt's early work,[4] developers have introduced many practical multimodal systems. Although these solutions have demonstrated multimodal interaction's efficacy, the results have not been widely used. Few proposed solutions are easily generalized for use in other contexts.

We've also witnessed numerous attempts to create theoretical frameworks for multimodal human-computer interaction. For example, in his modality theory, Niels Ole Bernsen introduced a generative approach to analyzing modality types and their combinations based on his taxonomy of generic unimodal representation modalities.[5]

Most existing theoretical approaches have little practical value, however, and applying them in common software design processes is difficult.

### References

1. M.M. Blattner and E.P. Glinter, "Multimodal Integration," *IEEE MultiMedia*, Oct.-Dec. 1996, pp. 14-24.
2. S.L. Oviatt, "Ten Myths of Multimodal Interaction," *Comm. ACM*, Nov. 1999, pp. 74-81.
3. M. Turk and G. Robertson, "Perceptual User Interfaces (Introduction)," *Comm. ACM*, Mar. 2000, pp. 33-35.
4. R.A. Bolt, "Put That There: Voice and Gesture at the Graphics Interface," *Proc. Siggraph*, ACM Press, 1980, pp. 262-270.
5. N.O. Bernsen, "Foundations of Multimodal Representations: A Taxonomy of Representational Modalities," *Interacting with Computers*, vol. 6, 1994, pp. 347-371.

- defining the modality concept precisely, and
- identifying a UML extension for modeling multimodal interaction.

To define the modality concept, we created a metamodel representing an abstract, higher-level view of various aspects of multimodal interaction. We then introduced UML extensions for modeling basic modalities and describing complex multimodal systems.

## Metamodel

Defining multimodal user-interface models requires a vocabulary of modeling primitives. Our metamodel therefore formally describes basic multimodal interaction concepts. The metamodel's main concept is that an HCI modality engages human capabilities to produce an effect on users.

Figure 1 shows a simplified HCI modality model. HCI modalities can be simple or complex. A simple HCI modality represents a primitive form of interaction; a complex HCI modality integrates other modalities and uses them simultaneously.

We defined input and output types of a simple HCI modality using the computer as a reference point. Our input and output modalities are therefore not symmetric with human input and output modalities. They represent a computer viewpoint in which code, not neural circuitry, controls the interaction with users.

Input modalities are event-based or streaming-based and require a user device to transfer human output into a form suitable for computer processing. Event-based input modalities—such as input via a keyboard or mouse—react to user actions by producing discrete events.

Streaming-based modalities sample input signals with some resolution and frequency, producing a time-stamped array of sampled values. For example, a computer detects a user's voice or psychological signals by sampling input signals with sensors such as a microphone or electrode.

Applications can use sampled values directly, but additional computing modules often further process the values before sending them to the application. For example, speech and handwriting recognition platforms generate tokens based on a complex analysis of sampled data. With our framework, we can model recognition-based streaming modalities that add a pattern-searching process over streaming data. All recognition-based modalities are probabilistic in nature and usually used in noisy environments so they often introduce some recognition error.

Output modalities present either static or dynamic data to the user. Some modalities, such as speech, are inherently dynamic, but many dynamic presentations are simply animations of some static modality. A movie, for example, represents animated static pictures. We can describe this kind of dynamic presentation using a time scale to determine the level of human processing needed to produce a desired effect.

HCI researchers agree that there are three important levels of human interactive response:[7]

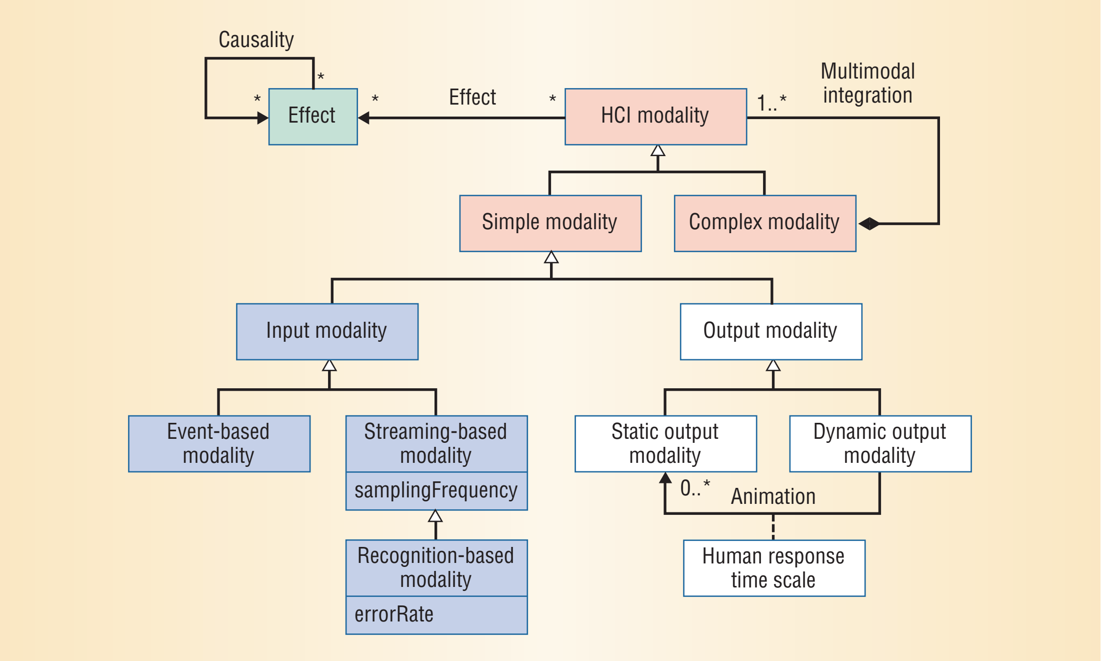


***Figure 1. Simplified human-computer interface modalities model. HCI modalities can be simple or complex. A complex HCI modality integrates two or more modalities to use them simultaneously.***

- perceptual processing,
- immediate response, and
- unit task.

*Perceptual processing time* (about 0.1 second) is the amount of time the human perceptual system spends integrating and processing signals. Two stimuli within this time seem fused, and responses feel instantaneous. Movies, for example, show 10 or more frames per second, creating a sensation of continuity for average users.

*Immediate response time* (about 1 second) is the minimal amount of time a user requires to react to a new situation—for example, the appearance of a new form on the screen. If presentation changes occur faster than this, users don't feel like they're waiting. Short animations usually exploit this effect.

*Unit task time* (about 10 seconds) represents a time scale of the simplest tasks the user wants to perform. Eliciting a more complex reaction from users requires presenting the data at a slower rate.

Each modality engages human capabilities, producing some effect on the user. Table 1 classifies these effects into four main categories:

- *Sensory* effects describe the human sensory apparatus's processing of stimuli.
- *Perceptual* effects result from the human perceptual system's analysis of sensor data.
- *Motor* effects describe human mechanical actions, such as head movement or pressure.
- *Cognitive* effects occur at higher levels of human information processing and include memory, attention, and curiosity processes.

In our metamodel, these concepts are subclasses of the Effect class in Figure 1.

**Table 1. Simplified classification of human concepts used for defining effects of multimodal interaction.**

| Classification | Concepts |
|---|---|
| Sensory | Stimulus: light, sound, vibration |
| | Sensory excitation |
| | Sensory processing: color, sharpness, peripheral vision |
| Perceptual | Pattern recognition |
| | Grouping: similarity, proximity, or voice color or timber |
| | Highlighting: color, polarity, or intensity |
| | 3D cue such as stereo vision or interaural time difference |
| | Illusion |
| Motor | Movement: translation or rotation |
| | Force: pressure or twisting |
| | Hand or head movement |
| | Degree of freedom |
| Cognitive | Short- or long-term memory and memory processes such as remembering and forgetting |
| | Attention: focus and context |
| | Reasoning: deductive, inductive, and abductive |
| | Problem solving: Gestalt, problem space, and analogical mapping |
| | Analogy |
| | Skill acquisition: skill level, proceduralization, and generalization |
| | Linguistics: speech, listening, reading, and writing |
| | Curiosity |

Table 2. UML stereotypes for multimodal user-interface modeling.

| Type | Description |
|---|---|
| *Class stereotypes* | |
| Input modality | Captures some human output, such as movement or speech |
| Static output modality | Statically presents data—for example, pictures or graphics |
| Dynamic output modality | Dynamically presents data—for example, movies or 3D animation |
| Human interactive response | Defines a human interactive response time scale |
| Complex modality | Integrates two or more modalities |
| Sensory effects and parameters | Visual, audio, or haptic stimuli produced by output devices |
| Human movement | Human motor effect of movement |
| Perception, 3D cues, and perceptual parameters | Visual, audio, and haptic perceptual effects produced by the user interface |
| Cognitive and linguistic effects; analogy | Cognitive and linguistic effects produced by the user interface |
| *Association stereotypes* | |
| Integration | Connects a complex modality with a simple or other integrated complex modality |
| Effect | Connects a modality class with a sensory, perceptual, or cognitive effect the modality produced |
| Comparison | Connects perceptual parameters with objects being compared; perceptual effects are always based on comparing some basic stimulus |
| Rendering | Connects output modalities with an output device |
| Capturing | Connects input modality with the human output that modality captures |

Effects are often interconnected. For example, all perceptual effects result from sensory effects. These relations help designers determine the result of using some effects.

## UML extensions

Although our metamodel is independent of a specific modeling language, we used it to define UML extensions. Incorporating a generic HCI framework into UML provides a standard way to produce quantifiable and analyzable models. We used UML v. 1.5, which is widely supported by existing computer-aided software engineering (CASE) tools.

UML is a general-purpose modeling language that includes built-in facilities for customizing—or profiling—a particular domain. Profiles fully conform to general UML semantics but specify additional constraints on selected general concepts to capture domain-specific forms and abstractions. Our formal extension mechanism lets practitioners extend UML semantics to include

- *stereotypes*—adornments that give new semantic meanings to modeling elements,
- *tagged values*—key value pairs that we can associate with modeling elements, and
- *constraints*—rules defining the models' well-formedness.

We defined a new profile that introduces several UML extensions based on the proposed metamodel. We've focused primarily on extending class diagram artifacts as we model modalities as a specific style or class of interaction. A concrete user interface can be viewed as an instance of these models.

We can use several other UML modeling elements and models directly. For example, sequence and collaboration diagrams describe interaction among objects in a system, and we can declare these objects as instances of stereotyped classes. Similarly, messages among objects go over links that are instances of class diagram associations.

Our extensions can describe a multimodal interaction at different levels of abstraction with various levels of sensory, perceptual, and cognitive detail. Table 2 shows some of our UML class and association stereotypes.

# MODELING INTERACTION MODALITIES

To illustrate some uses of our proposed modeling framework, we've applied it in both basic modality models, such as textual or tabular presentation and aimed hand movement, and higher-level models of complex multimodal user interfaces.

## Basic modalities

In addition to being valuable in analysis and content repurposing,[6] basic modality models are useful for educational purposes because they explicitly describe effects of modalities that are typically used intuitively.[8]

Figure 2 is a UML class diagram, created with defined UML extensions, describing the effects of graphical textual presentation.

A screen's basic presentation modality is a pixel, rendered by a raster screen device. Pixels form letters—complex modalities that add the perceptual effect of shape recognition based on the user's knowledge of the alphabet. Words integrate letters, adding the perceptual effect of grouping by proximity.

Text lines integrate words to add the perceptual effect of grouping by good continuation. Text lines are grouped into paragraphs, which enrich presentation with several perceptual effects: Paragraphs group text lines by proximity; alignment changes the shape of the entire paragraph; and indentation highlights the first line because it is usually shorter than the other lines of text.

Figure 3 illustrates a table as a presentation modality. Table cells are the basic presentation

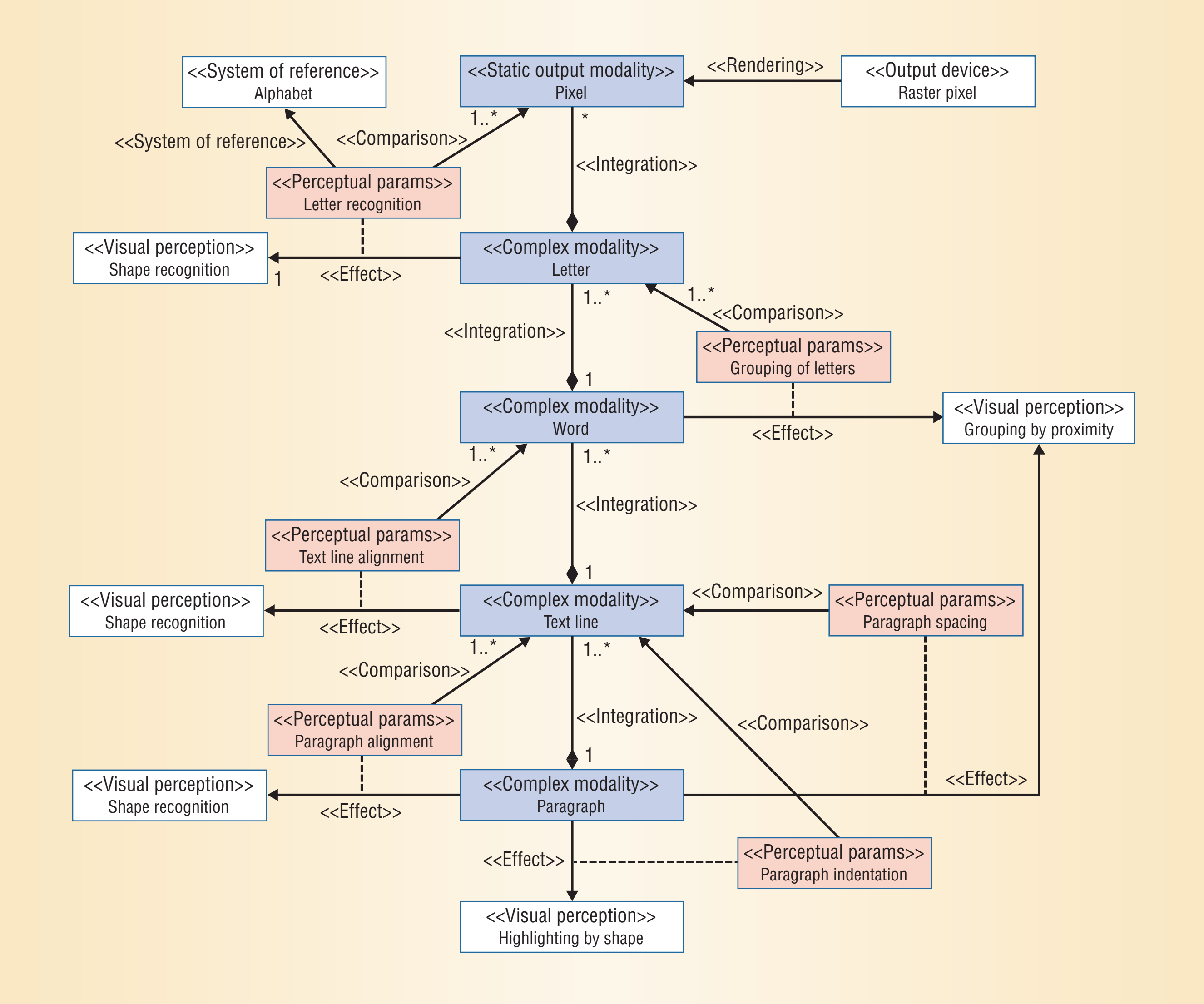


***Figure 2. Description of textual presentation modality. Text is a complex modality that produces various visual perceptual effects.***

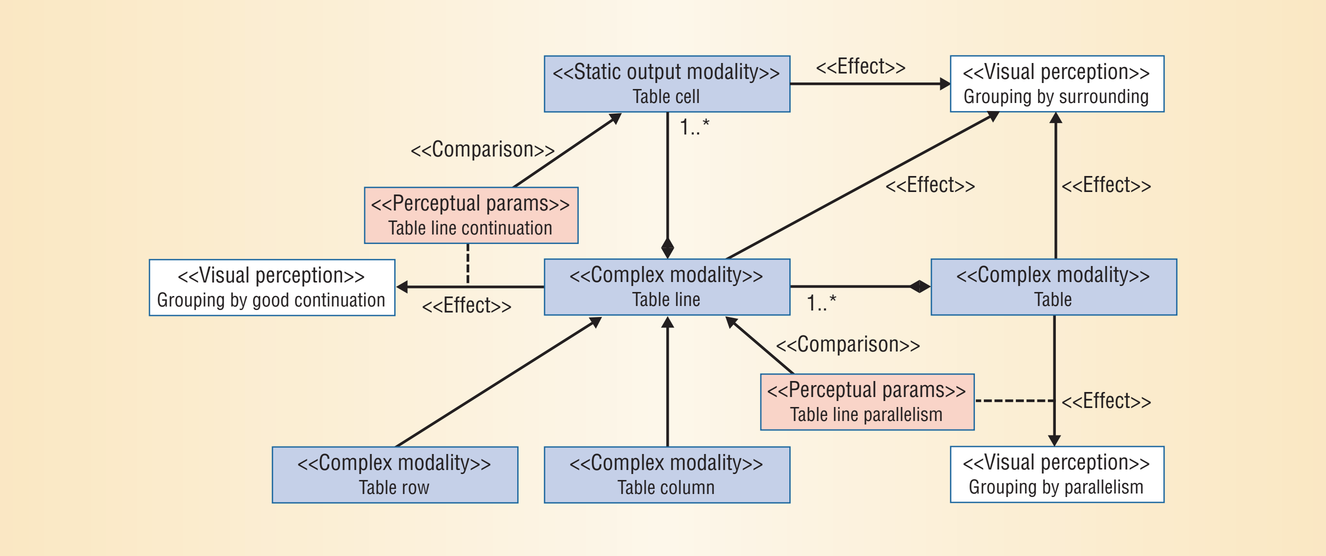
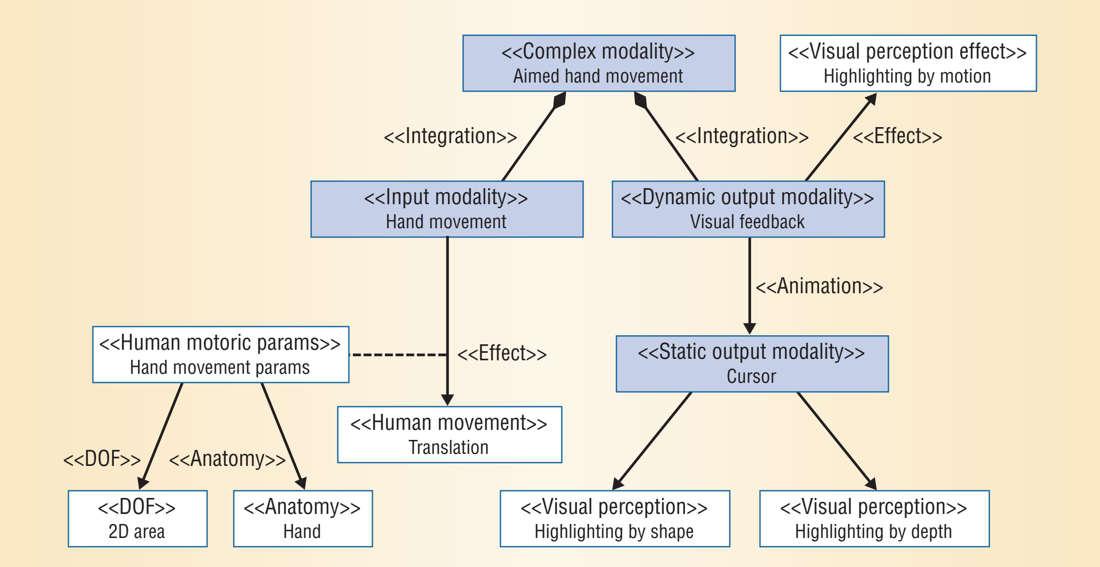


***Figure 3. Description of tabular presentation modality. A table is a complex modality that visually organizes a presentation using perceptual effects such as grouping by closure or good continuation.***

***Figure 4. Description of an aimed hand movement modality. An aimed hand movement is a complex modality that integrates hand movement input with graphical feedback. DOF = degrees of freedom.***



modality, introducing the visual perceptual effect of grouping by surrounding. Table cells are grouped into table lines (rows or columns), adding the perceptual effects of grouping by good continuation and, optionally, grouping by surrounding (row or column borders). A table integrates lines, bringing in the perceptual effects of grouping by parallelism and surrounding (table border).

Figure 4 illustrates the aimed hand movement often used in WIMP (windows, icon, menu, pointer) interfaces. Aimed hand movement is a complex modality integrating hand movement input (that is, the motions of a user's hand on a flat surface) and visual feedback. The visual feedback is a dynamic presentation modality animating the static presentation of a cursor, usually in the shape of an arrow. The static cursor introduces the perceptual effect of highlighting by shape and sometimes by depth (shadow), while dynamic visual feedback adds the perceptual effect of highlighting by motion.

## Complex multimodal user interfaces

We can view user interfaces as one-shot, higher-order messages sent from designers to users.[9] A user-interface designer defines an interactive language that determines which messages and levels the interaction will include.

Multimodal user interfaces, however, typically use commercially available implementation platforms, which don't offer modality and multimodal integration concepts. Consequently, determining the designer's original intent, which can be important when analyzing and reusing parts of the user interface, is sometimes impossible.

Higher-level multimodal interface models can help to better track the original developer's aims. Developers could create these models even before design and implementation, using them to consult HCI experts or to work with analysis tools to evaluate general decisions. Analyzing highly abstract and incomplete models early in the development cycle is critical because software designers make most fundamental design decisions during this stage.[5]

We used our framework to describe a multimodal presentation of brain electrical activity. The environment, the mmViewer, uses various visualization and sonification modalities to efficiently perceptualize biomedical data.[10]

Visualization in the mmViewer is based on animated topographic maps projected onto the scalp of a 3D head model using several graphical modalities, including 3D presentation, animation, and color. Sonification in mmViewer is the modulation of natural sound patterns to reflect certain features of processed data, emphasizing the temporal dimension of the selected visualized scores.

Because the topographic map itself represents a large amount of visual information, sonification covers the presentation of global parameters of brain electrical activity, such as the global index of left/right hemisphere symmetry. Changing the sound source's position in the 3D world sonifies this parameter. Therefore, the physician could perceive the activation of a hemisphere as the movement of a sound source toward that hemisphere.

Figure 5 is a simplified UML class diagram of perceptual and cognitive effects the designer wants the environment to produce. Multimodal presentation of electroencephalogram (EEG) activity is a complex modality integrating 3D visualization and sonification. 3D visualization integrates a 3D head model with an animated color map. By letting users freely explore the model, 3D visualization adds a motion parallax visual cue. Shadow and lighting let users recognize the 3D cues in the head model.

Animation dynamically changes the colors in a static color map based on brain electric activity values. This animation is smooth, occurring fast enough to activate users' visual perceptual processing.

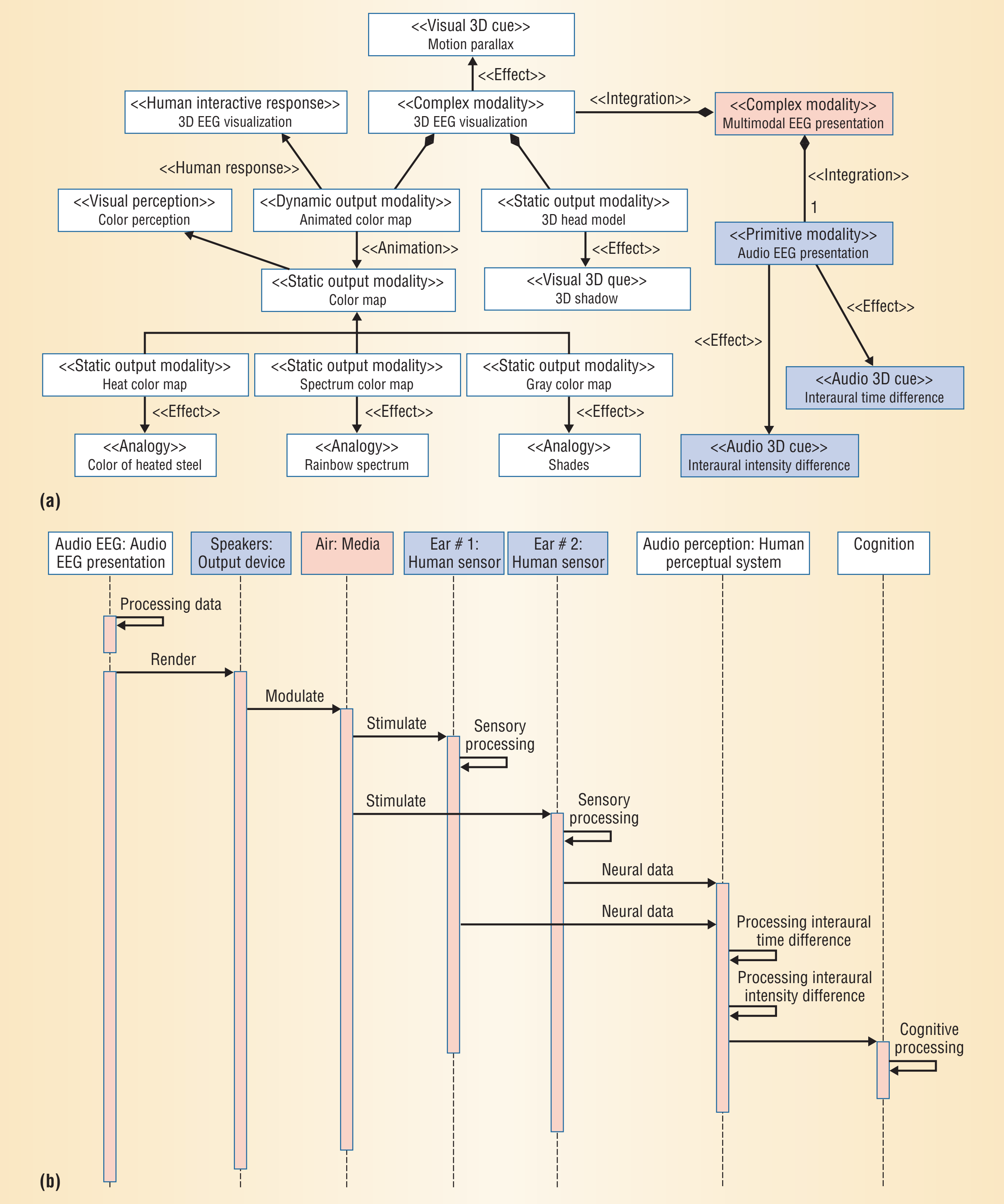


***Figure 5. UML class diagram. (a) Effects produced by the environment for the 3D presentation of electroencephalogram (EEG) signals; (b) simplified interaction sequence diagram of audio presentation modality with the environment.***

We used three types of color maps:

- heat, which maps brain electrical activity values to colors analogous to the colors of heated steel (black: cool; red: hot; white: extremely hot);
- spectrum, which uses colors analogous to the familiar rainbow spectrum; and

- gray, which uses different shades of gray, from black to white.

The interaural time and the sound intensity difference produce a stereo effect that determines sonification. UML sequence and collaboration diagrams can describe interaction among user-interface objects. They can also describe the environment's interaction dynamic. For example, the sequence diagram in Figure 5b describes interaction between the sonification modality, the presentation device, and the human sensory, perceptual, and cognitive systems.

UML models provide a metadescription of a multimodal system, but researchers can use automation to create tools for analyzing and transforming these models. For example, developers can apply user-interface models at various abstraction levels to develop efficient tools for repurposing existing user interfaces to other platforms.[6] Developers can also use models to automate some design phases for new multimodal interfaces. For example, we've developed tools for Java-based multimodal interface frameworks.[8]

Developers can use a standard means for representing multimodal interaction to seamlessly transfer UML interface models between design and specialized analysis tools. Many existing tools automatically support interchanging UML models, and their compatibility has increased with the introduction of standards such as the Exchangeable Metadata Interface (XMI).

Our multimodal interaction metamodel could provide the context for concepts in which humans perceive subtle relations. Developers can use UML's semantic extensions to provide formal descriptions of multimodal interfaces at various levels of abstraction.

We developed customized extensions of Rational Rose, one of the most widely used UML CASE tools, and used them to create the presented models. In future work, we plan to extend existing software development processes, such as the Rational Unified Process, with primitives for better description of multimodal systems and to integrate our solutions into various existing CASE tools. ■

## Acknowledgments

We thank Bran Selic, who provided useful feedback on the work described here. We also thank the anonymous referees whose excellent suggestions significantly improved this article.

## References

1. M. Turk and G. Robertson, "Perceptual User Interfaces (Introduction)," *Comm. ACM*, Mar. 2000, pp. 33-35.
2. R. Vertegaal, "Attentive User Interfaces," *Comm. ACM*, Mar. 2003, pp. 31-33.
3. S. Oviatt, T. Darrell, and M. Flickner, "Multimodal Interfaces That Flex, Adapt, and Persist," *Comm. ACM*, Jan. 2004, pp. 30-33.
4. Object Management Group, "The Unified Modeling Language Specification, v. 1.5," OMG, Mar. 2003, www.omg.org/technology/documents/formal/uml.htm.
5. B. Selic, "The Pragmatics of Model-Driven Development," *IEEE Software*, Sept./Oct. 2003, pp. 19-25.
6. Z. Obrenovic, D. Starcevic, and B. Selic, "A Model-Driven Approach to Content Repurposing," *IEEE MultiMedia*, Jan.-Mar. 2004, pp. 62-71.
7. A. Newell, *Unified Theories of Cognition*, Harvard Univ. Press, 1990.
8. Z. Obrenovic, D. Starcevic, and V. Devedzic, "Using Ontologies in the Design of Multimodal User Interfaces," Human-Computer Interaction, *Proc. IFIP Interact 03 Conf.*, IOS Press, 2003, pp. 535-542.
9. R. Prates, C. de Souza, and S. Barbosa, "A Method for Evaluating the Communicability of User Interfaces," *Interactions*, Jan./Feb. 2000, pp. 31-38.
10. E. Jovanov et al., "EEG Analysis in a Telemedical Virtual World," *Future Generation Computer Systems*, vol. 15, 1999, pp. 255-263.

***Zeljko Obrenovic** is a researcher at the Center for Command Information Systems and is a lecturer at the University of Belgrade and at the Military Academy of Serbia and Montenegro. His research interests include human-computer interaction as well as the development of advanced user interfaces in education and medicine. He received a PhD in computer science from the University of Belgrade, Serbia and Montenegro. Contact him at obren@fon.bg.ac.yu.*

***Dusan Starcevic** is a professor at the School of Business Administration, University of Belgrade, and a visiting professor at the university's School of Electrical Engineering. He is also chair of the Department of Information Systems at the School of Business Administration. His main research interests include distributed information systems, multimedia, and computer graphics. He received a PhD in information systems from the University of Belgrade. Contact him at starcev@fon.bg.ac.yu.*