Interactive Systems & New Interface Technologies

Engineering for Multimodal Human-Computer Interaction

Laurence Nigay[1] and Phil Gray[2]
1 CLIPS-IIHM Lab., University of Grenoble, France
2 Dept. of Computing Science, University of Glasgow, UK

1 Introduction

Engineering for Human-Computer Interaction (EHCI) focuses on system development models, methods, processes and tools that enable teams to create effective user interfaces on-time and on-budget. Software engineering typically identifies four main classes of activity within a system development project: Planning, Analysis, Design, Evaluation. For the design activity, architectural modeling (an organization of computational elements and the description of their interactions) is becoming a central challenge for large, complex interactive systems. Indeed, with the advent of new interaction technologies and user-centered concerns, the user interface portion of interactive systems is becoming ever more large and complex. This increasing complexity and size of software systems requires sound engineering principles and frameworks to formally structure the design activity into multiple but consistent perspectives. Moreover, software tools for the construction of user interfaces will not eliminate architectural issues as long as the construction of user interfaces requires programming. Clearly, developers and maintainers of interactive systems need to rely on canonical models for identifying software components, for organizing their interconnections, for reasoning about them and for maintaining them in a productive way.

It is impossible to discuss the engineering of multimodal Human-Computer Interaction (HCI) and, in particular, software architecture models without first focusing on the user interface changes that will require new results in EHCI. Indeed, innovation in EHCI follows innovation in UI design; developments in EHCI make sense only when you know the types of interfaces for which you are building methods and tools. Therefore we start by identifying issues for emerging Multimodal User Interfaces (MUI) before discussing implications for EHCI.

2 Emerging interaction modality and multimodality

The use of multiple modalities such as speech, non-speech sound, gesture, touch and both static and dynamic graphics opens a vast world of possibilities for HCI. By extending the sensory-motor capabilities of computer systems to better match the natural communication means of human beings, multimodal interfaces enhance interaction between users and systems in several ways:

- Information bandwidth (that is the amount of information being communicated) is increased;
- Signal-to-noise ratio of conveyed information (that is the rate of information useful for the task being performed) is also increased;
- The resulting interaction is more efficient, robust, comfortable and natural for users.

2.1 A vast world of modalities defined by the recent interaction paradigms

Recent progress achieved in wireless networks, the miniaturization of microprocessors and in related software (e.g., ad hoc peer-to-peer networks, “tiny” operating systems) make it possible to foresee the disappearance of the "grey box" that is the personal computer, or at least to understand that it is no longer the only location for interaction between people and the digital world. This development is largely driven by the emerging concept of Ubiquitous Computing, as predicted by Mark Weiser [23], and from the evolution occurring in the field of interaction modeling. Research in this area is now becoming directed towards models of interaction in which data-processing resources are distributed in a multitude of everyday objects with which users interact in explicit (active modalities) and implicit ways (passive modalities). This has given rise to several recent interaction paradigms including Mobile Computing, Augmented Reality [24], Ubiquitous/Pervasive Computing [23], Tangible Interfaces [15] [9] and Embodied Interfaces [8]. All of these developments separately and, more powerfully, when combined, increase the set of possibilities for and, indeed, the requirement to use, multimodal interaction. An example of a recent type of modality is provided by “phicons” [15] (Physical Icons) that define new input modalities based on the manipulation of physical objects or physical surfaces such as a table or a wall that can be used for displaying information (output modality) in an ubiquitous computing scenario. More speculatively, we can envisage further additions to the repertoire of modalities in the future, including smell [4] and direct brain input.

2.2 Passive/Active modalities: A unifying point of view on recent interaction paradigms

Within the vast world of possibilities for input modalities (from the user to the system) as well as for outputs (from the system to the user), we distinguish two types of modalities: active and the passive [3]. For inputs, active modalities enable a user to issue a command or send data to the computer (e.g., a voice command or a gesture recognized by a camera). Passive modalities refer to information that is not explicitly

expressed by the user, but rather is automatically captured for enhancing the execution of a task, as in perceptual user interfaces. For example, in the "Put that there" seminal multimodal demonstrator of R. Bolt, eye tracking was used for detecting which object, the user was looking at, on screen.
We identify two forms of multimodality based on the combination of passive/active modalities:

- Active modalities augmented by passive modalities for *making active modalities more robust*.
- Passive and active modalities integrated for *obtaining the user's intention*. Passive and active modalities are modeled as two modalities of equal importance.

2.3 Sensing and recognition-based modalities

As pointed out in [19], interaction must move beyond the desktop and beyond direct manipulation. We can expect less traditional workstation-oriented WIMP interfaces in the future and, as a consequence of substantially more computing power, recognition-based modalities including handwriting, vision-based gesture, and object recognizers will be more common [17] [20]. As pointed out in [14], such modalities are already actively studied and we need to envision systems that will take advantage of the computing power to use several sources of sensing data in parallel that will be combined (complementarily or redundantly) to improve recognition performance, such as the use of "20 cameras with narrow but overlapping field of view to cover a single user" [14].

2.4 Mixed initiative interactive system: Combining conversational and direct manipulation paradigms

Moving beyond the desktop implies that the WIMP interface will no longer be the standard as it has been for nearly 20 years (starting in 1984 with the Macintosh). We can foresee a huge diversity in interaction modalities as explained in section 2.1 as well as in the interaction paradigms built with these modalities. In particular, we can envision a combination of conversational (computer-as-partner [1]) and direct manipulation (computer-as-tools [1]) paradigms. Indeed with ubiquitous systems, interaction computing systems will more closely resemble other forms of human interaction with, and in, the world, by speaking, gesturing and using tools; these natural actions will be both explicit and implicit, in other words will lead to active and passive modalities (section 2.2). Such combinations of conversational and direct manipulation paradigms, also called mixed initiative interactive systems, are a promising avenue since in the multimodal community such paradigms have been extensively studied, albeit in isolation by different research communities. For example, many studies have focused on multimodal conversational interfaces based on speech as a dominant modality combined with gestures, while several other studies focus on multimodal interaction enhancing the sensory-motor capabilities of a WIMP interface by enriching it with innovative modalities such as vision-based head tracking or tilting a PDA for scrolling a text as well as combined modalities based on a fusion mechanism to obtain complete commands.

2.5 Huge variability in available and used interaction modalities

Early multimodal systems involved the designer making the selection or combination prior to use, which meant that the designer had to prejudge which modalities and combinations of modalities would best suit the user's context and activity. Later systems were adaptable, in that they allowed users to explicitly choose from the designer's palette of modalities at run time, but this involves cognitive effort and distraction from user tasks. While explicit control by users must be available if demanded (for example for explicitly migrating a part of a user interface from one display to another one), implicit interaction (passive modalities) and context capture can potentially be used to adaptively select and combine modalities, in the form of adaptive multimodal interaction. Such implicit support for the dynamics of user activity is a central aspect of pervasive/ubiquitous computing, where the aim is to let the user focus on his task not his tools, and where a good tool is considered to be one that the user acts through rather than on. The goal is to let users act through multimodal interaction devices rather than on them. As part of ubiquitous computing, adaptive interfaces to varying input and output capabilities, such as a graphical user interface that must run on PC, on PDA and on cell phones, have been a subject of several studies. Moreover research on adaptive multimodal user interface is one facet of the research axis on plasticity as coined by Thevenin & Coutaz: plasticity is the capacity of a user interface to withstand variations of both the system physical characteristics and the environment while preserving usability.[1]

The extent to which interaction techniques and modalities can be successfully selected automatically remains the subject of debate within the HCI research community [5]. It is therefore a prime candidate for a "grand challenge" in the area of multimodal systems.

[1] Note that a plastic user interface can also be monomodal.

3 Engineering for multimodal human-computer interaction

The issues discussed above impact on the models, methods, processes and tools for multimodal Human-Computer Interaction. In this section we discuss some of these issues from the engineering point of view: the transition from WIMP interfaces to ubiquitous interactive systems requires new interaction models and corresponding tools for design and development.

3.1 Innovation towards an interaction model for multimodality

As explained in section 2.1, recent interaction paradigms such as tangible UIs have opened up an enormous, and little understood, world of possibilities for interaction, including modalities based on the manipulation of physical objects (such as a bottle) and modalities based on the manipulation of a PDA, etc. In the face of such an increasing variety of interaction modalities we can no longer expect to model each input and output modality in all their diversity at the concrete level. For example, because of the overwhelming number of recommendations that would be generated, the W3C cannot possibly provide recommendations for each new interaction technique. The time has come to reason about modalities at a higher level of abstraction. A core model must be defined for identifying and describing the generic building blocks for pure and combined modalities. Such a core model for modality integration will greatly help designers and programmers by allowing them to reason at a higher level of abstraction than the level of a particular modality. This conceptual result will be a crucial tool in facing the increasingly large variety of modalities.

3.2 Innovation towards a multimodal software architecture model

We identify several points for a new multimodal software architecture model. Designers and developers of interactive systems need to rely on software architectural models:

- for specifying software components,
- for organizing their interconnections,
- for reasoning about components and interconnections,
- for verifying ergonomic and software properties,
- for modifying and maintaining them in a productive way.

Combining conversational and direct manipulation paradigms: while a software decomposition of conversational interfaces mainly relies on cooperative agents and focuses on the Dialog Controller for interpreting and generating multimodal communicative acts, several software architecture models for direct manipulation and multimodal commands have been driven by software engineering properties such as portability and modifiability. A new software architecture model for multimodal interfaces must be defined for mixed initiative interactive systems.

Based on the interaction model for multimodality (section 3.1), the software elements of a passive/active pure/combined modalities must be identified. Such building blocks must be included in the multimodal software architecture model.

Generic services such as fusion mechanisms must be identified and included in the multimodal software architecture model. For example generic services for dealing with uncertainty and errors linked to sensing and recognition-based modalities (section 2.3) must be defined and integrated within the software architecture model.

To scope the huge variability in interaction modalities described in section 2.5, adaptation is a central point that has impact on software architecture. For designing adaptable multimodal interfaces (choices made by the user) a “meta” user interface (a kind of end-user programming interface) and its corresponding software structure must be defined and related to the multimodal software architecture model. For defining adaptive multimodal interfaces (choices made by the system), the implementational software architecture must be distributed, mobile, able to dynamically discover new modalities (resource awareness) and reconfigure them (adaptability) [13]. It must also be able to handle different distributed system paradigms (i.e., both “always connected” client-server and intermittently connected ad hoc peer-to-peer networks) and provide appropriate security for both passive and active i/o throughout data lifetime, from sensors to application. As pointed out in [6], the goal of the new architecture for multimodal interfaces is “to provide the infrastructure that allows components to be coordinated to support a user in a task, regardless of environment or locality”.

Finally, the model must be designed to be open, with well-defined interfaces and data interchange schema definitions plus metamodels to support adaptation. As Chris Mairs expressed eloquently in the 2006 IEE/BCS Turing Lecture, the emergence of open systems has already been instrumental in making effective multimodal systems available as assistive technologies for the disabled.

3.3 Innovation towards Model-Driven Engineering for multimodal HCI

Going one step further with software architecture models, the huge variability of modalities and the need for adaptation naturally lead to an emerging area of software engineering (Figure 1): MDA (Model Driven Architecture) and MDE (Model Driven Engineering). Indeed MDA/MDE promotes the separation between domain and technological concerns by the definition of platform-independent and platform-specific models (PIM/PSM) in the engineering process, such as the classical Y one. Such separation is strongly related to early work in HCI on device-independent user interface specification for automatically generating the concrete interfaces. The huge variability and the adaptability required for multimodal interfaces creates a major research challenge: how can we achieve the design and engineering advantages of device-independent user interface models (i.e., the abstract user interface) while incorporating the pure/combined modality model based on the multimodal interaction model of section 3.1 and respecting the differences in form of interaction embedded at the concrete level?

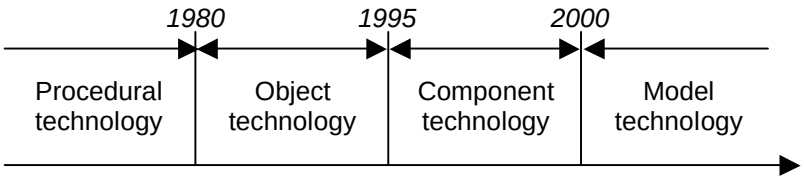

Figure 1: Evolution in development technologies.

3.4 Innovation towards tools for rapid prototyping / development

As pointed out by B. Gaines' model [11] (Figure 2) on how science technology develops over time, it is now timely to make a step change in the domain of multimodal interaction. As shown in Figure 2, after the initial breakthrough phase [2] we are now at the stage of replication. Although several multimodal systems have been built, their development still remains a long and difficult task. Indeed, the flexibility and robustness of multimodal systems give results of an increased complexity of the software that current design/development tools do not address appropriately. The existing frameworks dedicated to multimodal interaction are currently few and limited in scope. Either they address a specific technical problem including the fusion mechanism [10] [18], the composition of several devices [7] and mutual disambiguation [20] [10], or they are dedicated to specific modalities such as gesture recognition [25], speech recognition [12] or the combined usage of speech and gesture [16].

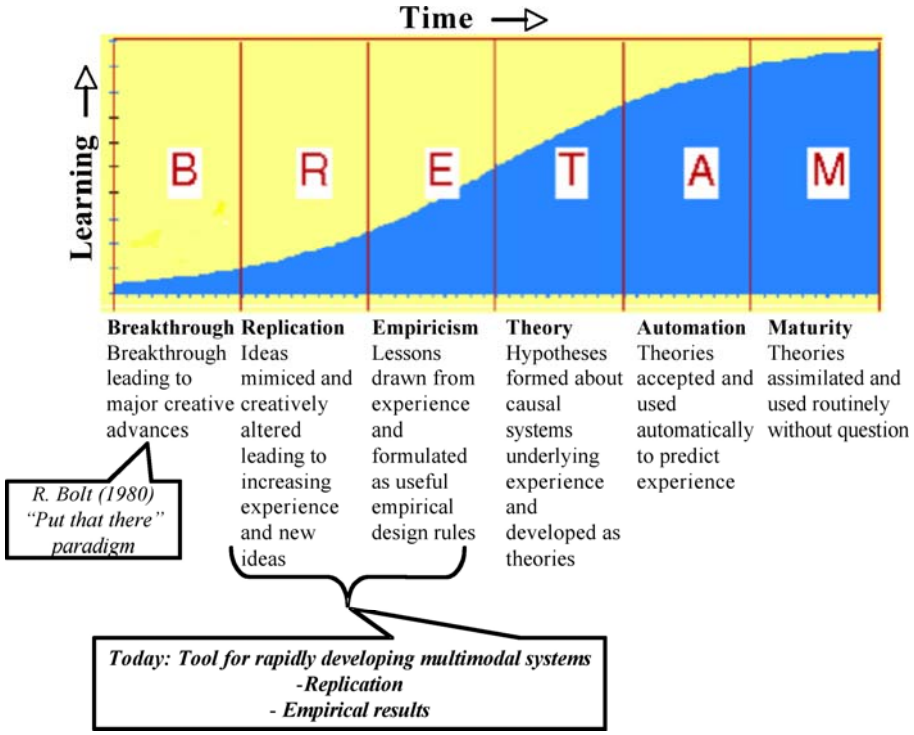

Figure 1. B. Gaines's model on how science technology develops over time, applied to multimodal interaction. [Gaines 91]

Such tools dedicated to multimodal interfaces are crucial for moving away from WIMP interfaces. They will allow multimodal user interfaces to be created more quickly and to be able to be integrated cost-effectively

into new ubiquitous and mobile environments. Such tools should enable rapid prototyping and therefore more iterations as part of an iterative design method for achieving usable multimodal user interfaces [17] as well as providing a means of enabling end-users and other stakeholders to become directly involved in the development and configuration of systems over which they can take genuine ownership and control.

Moreover, several modalities are based on the manipulation of physical objects. As a consequence, tools for rapidly prototyping multimodal interfaces must include the physical part of the interaction; simulation on screen will no longer be sufficient [17].

As defined in [17], in the general context of user interface software tools, such tools for multimodal interfaces must aim to have a *low threshold* (easy to use) while providing a *high ceiling* (how much can be done with the tool). Additionally, in order to take account of the ever-widening world of modalities, the tools must be easily extendable. The framework + plugin approach of IDEs like Eclipse provide a potential model.

4 References

[1] Beaudouin-Lafon, M. Designing Interaction, not Interfaces. Proceedings of AVI '04, ACM, 2004, pp. 15-22.

[2] Bolt, R. A. "Put-that-there": Voice and gesture at the graphics interface. Proceedings of SIGGRAPH'80, 14, 3,1980, pp. 262–270.

[3] Bouchet, J., Nigay, L. ICARE: A Component-Based Approach for the Design and Development of Multimodal Interfaces. Extended Abstracts ACM CHI'04, 2004, pp. 1325-1328.

[4] Brewster, S.A., McGookin, D.K. and Miller, C.A. *Olfoto: Designing a smell-based interaction.* To appear in Proceedings of ACM CHI 2006.

[5] Chalmers, M, Galani, A. Seamful Interweaving: Heterogeneity in the Theory and Design of Interactive Systems. Proc. ACM DIS 2004, pp. 243-252, August 2004.

[6] Cheng, S., Garlan, D., Schmerl, B., Sousa, J., Spitznagel, B., Steenkiste, P., Hu. N. Software Architecture-based Adaptation for Pervasive Systems. International Conference on Architecture of Computing Systems Trends in Network and Pervasive Computing, Springer Verlag, LNCS 2299, 2002.

[7] Dragevic, P., Fekete, J.-D. ICON: Input Device Selection and Interaction Configuration, Demonstration. UIST 2002 Companion, 2002, pp. 47-48.

[8] Fishkin, K.P., Gujar, A., Harrison, B. L., Moran, T. P., Want, R.. Embodied User Interfaces for Really Direct Manipulation. Communications of the ACM, 43, 9, 2000, pp. 74-80.

[9] Fishkin, K.P. A Taxonomy for and Analysis of Tangible Interfaces. Journal of Personal and Ubiquitous Computing, Springer-Verlag, 8, 5, 2004, pp. 347-358.

[10] Flippo, F., Krebs, A., Marsic, I. A Framework for Rapid Development of Multimodal Interfaces. Proceedings of ICMI'03, ACM, 2003, pp. 109-116.

[11] Gaines, B.R. Modeling and Forecasting the Information Sciences. Information Sciences 57-58,1991, pp. 3-22.

[12] Glass et al. A Framework for Developing Conversational User Interfaces, Proceedings of CADUI'2004, Springer-Verlag, 2004, pp. 354-365.

[13] Gray, P, Sage, M. Dynamic Links for Mobile Connected Context-Aware Systems. Proc EHCI 2001, 281-298.

[14] Hudson, S. E. Leveraging 1,000 and 10,000-Fold Increases: Considering the Implications of Moore's law on Future UI Tools Research. ACM CHI 2005 Workshop, 2005. Available via: http://hci.stanford.edu/srk/chi05-ui-tools/

[15] Ishii, H., Ullmer, B. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. Proceedings of ACM CHI'97, 1997, pp. 234-241.

[16] Krahnstoever et al. A real-time framework for natural multimodal interaction with large screen displays. Proceedings of ICMI'02, IEEE, 2002, pp. 349- 354.

[17] Myers, B. Hudson, S.E. and Pausch, R. Past, Present, and Future of User Interface Software Tools. ACM Transactions on Computer-Human Interaction, 7, 1, 2000, pp. 3-28.

[18] Nigay, L., Coutaz, J. A Generic Platform for Addressing the Multimodal Challenge. Proceedings of CHI'95, ACM, 1995, pp. 98-105.

[19] Olsen, D. R. and Klemmer, S. R. The Future of User Interface Design Tools. ACM CHI 2005 Workshop, 2005. Available via: http://hci.stanford.edu/srk/chi05-ui-tools/

[20] Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. & Ferro, D. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. Human Computer Interaction, Lawrence Erlbaum Publ., 15, 4, 2000, pp. 263-322.

[21] Thevenin, D. and Coutaz, J. Plasticity of User Interfaces: Framework and Research Agenda. Proceedings of 7th IFIP International Conference on Human-Computer Interaction Interact'99, Chapman & Hall, 1999, pp. 110-117.

[22] Turk, M., Robertson, G. Eds, Perceptual user Interfaces. Special issue. Communications of the ACM, 43, 3, 2000, pp. 32-70.

[23] Weiser, M. Some computer science issues in ubiquitous computing. Communications of the ACM, 36, 7, 1993, pp. 75-84.

[24] Wellner, P., MacKay, W., Gold, R. Eds. Special Issue of Communications of the ACM, 36, 7, 1993.

[25] Westeyn, T. et al. Georgia Tech Gesture Toolkit: Supporting Experiments in Gesture Recognition. Proceedings of ICMI'03, ACM, 2003, pp. 85–92.