Multimedia and multimodal systems: commonalities and differences

S. Anastopoulou, C. Baber, M. Sharples
anasto@eee-fs7.bham.ac.uk, c.baber@bham.ac.uk, m.sharples@bham.ac.uk
Educational Technology Group, School of Engineering, University of Birmingham, UK

1. Introduction

This paper focuses on the differences between multimodal and multimedia systems as well as some assumptions of multimodal interaction. It is argued that effectiveness of human-computer interaction can be maximised when the interactive experience is unified.

2. Definitions

Due to differences on terminology usage in the literature the terms modality, medium and representation need to be defined. In a communication act, such as learning in a classroom, modality refers to the sensory or perceptual experience (e.g. visual, tactile, etc.) and is closely related to the individual. Medium is a means of conveying a representation (to a human), e.g. a diagram or a text. Representation sketches or stores information, e.g. semantic net, English language. Consider, for example, a classroom where pupils are taught about gravity. They listen and look while the teacher explains by speaking and gesturing (perceptual experience). The teacher has written an equation of gravity (w=m*g) on the board as well as a diagram. These are different types of representation. Some of the artefacts that the teacher or the pupils use carry several representations, e.g. the board contains text, form or diagram written on it. Imagine now that pupils use the pen as an artefact to experiment with the law of gravity. Modality refers to the use of visual, auditory or tactile cues that pupils use to see the representations, hear the teacher's presentation or handle the pen to understand by doing (sensory or perceptual experience).

Additionally, the concept of multimodality needs also to be introduced. Multimodality is based on the use of sensory modalities by which humans receive information. These modalities could be tactile, visual, auditory, etc. It also requests the use of at least two response modalities to present information (e.g. verbal, manual activity) (Baber 2001). So, for example, in a multimodal interaction a user may receive information by vision and sound and respond by voice and touch. Multimodality could be compared with 'unimodality', which would be based on the use of one modality only to receive or present information (e.g. watching a multimedia presentation and responding by pressing keys).

3. Interactive systems

Multiple representations or multimedia systems share a common aim with multimodal systems: the effective interaction with the user. Effective interaction is considered regarding a system that is not only easy-to-use but also able to support the user in performing a task. Independently of the technological differences in the implementation of those systems, they aim to support their users while they perform particular tasks.

However, multimedia and multimodal systems have important differences. Lee (1996) identified that multimedia systems deal with the presentation of information. Multimodal systems interpret and regenerate information presented in different media (Lee 1996). Turk (2000) transfers the comparison to the user interface. According to him, the distinction between multimedia and multimodal user interfaces is based on the system's input and output capabilities. Thus, a multimodal user interface supports multiple computer input and output, e.g. using speech together with pen-based gestures. A multimedia user interface supports multiple outputs only, e.g. text with audio or tactile information provided to the user. As a result, multimedia research is a subset of multimodal research (Turk 2000). Baber (2001) argues that multimodal human-

computer interaction can have two perspectives: the human-centred and the technology-centred. According to the human-centred perspective, multimodal systems should support more than one sensory and response modality of the users. The technology-centred approach defines a multimodal system to be one that supports concurrent combination of (input) modes. Alternatively, it could at least specify which mode is operational on each situation (Baber 2001).

An alternative difference between multimodal and multimedia systems can be based on the perspective of the interactive experience. From the system's point of view, a multimedia system is also multimodal because it provides, via different media, the user with multimodal output, i.e. audio and visual information, and multimodal input, e.g. typing with the keyboard, clicking the mouse. From the user's point of view, a multimedia system makes users receive multimodal information. However, they can respond by using specific media, e.g. keyboard and mouse, which are not adaptable to different users or contexts of use. Additionally, while interacting with a multimodal system, users receive multimodal input and are able to respond by using those modalities which provide convenient means of interaction. While in multimedia systems the user has to adapt to the system's perceptual capabilities, in multimodal systems the system adapts to the preferences and needs of the user.

This argument, however, aims to highlight the importance of the interactive experience and not the importance of the individual per se. If the distinction is based only on the individual, a system could be multimodal for one user and multimedia to another.

3.1 Assumptions in multimodal interaction

In multimodal systems research, it is often assumed that human-human communication is 'maximally multimodal and multimedia' (Bunt 1998). The 'added-value' of multimodal systems is taken for granted and there is a lack of research about *why* we need to develop them. As Bunt (1998) stated: "in natural communication, all the modalities and media that are available in the communicative situation, are used by participants" (p. 3). But this is not always the case.

Furthermore, current research on multimodal interfaces is based on the naturalness of communication between the user and the system (Marsic 2000). Naturalness refers to a human-computer communication that would be like human-human communication. Thus, researchers are focused on technological achievement by generating recognition techniques of natural language, gestures, etc (Waibel 1995; Cohen 1997; Julia 1998; Oviatt 2000). Current research is mainly focused on the integration and synchronisation of different modalities to enhance communication (Bellik 1997; Oviatt 1997b). The main aim is to provide users with a system that is able to emulate how humans interact with each other. It would take advantage of human and machine sensing and perceiving capabilities to present information and context in meaningful and natural ways (Turk 2000).

However, there are differences between human-human and human-computer interaction. In human-human interaction, for example, there is available a quite sophisticated system (human's mind), which indicates which modality to be used and when. Current multimodal research often assumes that technology supported modalities are useful while performing particular tasks without questioning why.

Additionally, while humans interact with the computer, they need to transform their conceptions of activities into systematic and procedural functions. When the interaction is completed, humans need to interpret the interaction in order to make sense of it. For example, to transform feet into centimetres with a calculator, users need to think of procedures to figure out what calculations they need to do. When the calculation is completed, they need to interpret the result to make it useful, e.g. imagine the result in length.

While interacting with the calculator, humans need to know much more than how to use the calculator. They need to know how to transform feet into inches, what is the relation between

inches and centimetres, etc. The procedures, however, have been internalised and are considered as one (Collins 1990). The experience has been unified and it is perceived as a whole. Another example of a unified experience would be how to drive a car. At the beginning, drivers need to think each procedure, e.g. to change the gear. As they gain expertise, the task become internalised and unified. Drivers then can do other things while driving a car, e.g. discuss.

To summarise, to what extent multimodal systems research should focus on supporting natural interaction as opposed to *effective* interaction is under question; where effective interaction is defined in relation to some task, e.g. the learning outcome (Lee 1996). A successful interaction with a multimodal system would be one that provides the user with procedures unified into an integrated experience. In the case of educational technology, a successful multimodal interaction would be one where users could overcome the difficulties they have while interacting with technology and are able to concentrate on the content of the information provided. In such an occasion the technology would fulfil its main aim to become the artefact that provides information/knowledge to the user. From the users' perspective, users could unify their experience of interacting with technology into an integrated one that would focus on learning.

4. Conclusions

This paper is focusing on differences between multimodal and multimedia systems. Multimedia systems refer to users' adaptation of a system's perceptual capabilities. Multimodal systems support users multiple ways of response according to their preferences and needs. Furthermore, assumptions of multimodal interaction are discussed to reveal shortcomings of current research. Initially the 'maximum' use of multimedia and multimodal communication is discussed, to expose the concept of the 'added-value' of technology. Subsequently, the concept of naturalness of communication is compared with the concept of the unified experience. It is argued that an experience that can be unified with expertise would lead to more effective human-computer interaction.

5. References

1. Baber, C., Mellor, B. (2001). "Using critical path analysis to model multimodal human-computer interaction." International Journal of Human Computer studies **54**: pp.613-636.
2. Bellik, Y. (1997). Media Integration in Multimodal Interfaces. IEEE Workshop on Multimedia Signal Processing, Princeton, New Jersey.
3. Bunt, H. C. (1998). "Issues on multimodal human-computer communication." Lecture notes in computer science **1374**: pp.1-12.
4. Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications. Proceedings of the Fifth International Multimedia Conference (Multimedia '97), Seattle, USA, ACM Press.
5. Collins, H.M. (1990). Artificial Experts: social knowledge and intelligent machines. USA, MIT Press.
6. Julia, A. C. L. (1998). MVIEWS: Multimodal Tools for the Video Analyst. Proceedings of IUI '98: International Conference on Intelligent User Interfaces, San Fransisco, CA.
7. Lee, J. (1996). Introduction. Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. J. e. Lee. Menlo Park, CA, AAAI Press.
8. Marsic, I., Medl, A., Flanagan, J. (2000). "Natural Communication with Information Systems." Proceedings of the IEEE **88**(8): pp.1354-1366.
9. Oviatt, S., Cohen, P., (2000). "Multimodal Interfaces that process what comes naturally." Communications of the ACM **43**(3).
10. Oviatt, S., DeAngeli, A., Kuhn, K. (1997b). Integration and synchronization of input modes during multimodal human-computer interaction. CHI'97, Atlanta, USA, ACM Press.
11. Turk, M., Robertson, G. (2000). "Perceptual User Interfaces." Communications of the ACM **43**(3): pp.33-34.
12. Waibel, A., Vo, M.,T., Duchnowski, P., Manke, S. (1995). "Multimodal interfaces." Artificial Intelligence Review **10**(3-4): pp.299-319.