Metaphors and Multimodal Interaction

Dorothy Rachovides, Zoë Swiderski, Alan P. Parkes
Computing Department,
Lancaster University
Lancaster. LA1 4YR. U.K.
+44 1524 592326
{dot | zoe | app}@comp.lancs.ac.uk

ABSTRACT
Multimodality has the potential to facilitate richer interaction styles in both information retrieval and learning environments. However, its true potential will not be realised unless consideration is given to the application of *combined* modalities. This paper asserts that multimedia output from a system actually *requires* multimodality on the part of the user in order to ensure that the effectiveness of the communication or information is not lost. The notion of a "multi-modelling" approach to interaction along with the use of gestures and metaphors have been examined and two systems are described which attempt to implement these approaches.

Keywords
multimodality, interaction, multimedia, metaphors

1. INTRODUCTION
Information can be conveyed between people in a number of ways. People draw on a range of materials (e.g. pen and paper) and physical abilities (e.g. gesturing) in order to express themselves. While machines have become more prevalent as providers of information, the methods used to convey and receive information have undergone some radical changes. This evolution from human-human to machine-human information sharing was merely the beginning. It was not sufficient for machines to simply replace humans as providers of information; they had to provide larger quantities of well-presented information and more possibilities of interactivity.

The traditional keyboard/mouse interaction style has long been regarded as limiting in terms of expressiveness, efficiency and how naturally it can be used, which has led to an interest in the development of alternative methods of input. Similarly, the output produced by such systems has become more dynamic, and exploits enhanced graphical interfaces to provide an enriched audio-visual experience. Much development has occurred in terms of input and output facilities. Output, in particular, can consist of a rich structured and interlinked collection of multi-media objects. However, little thought has been devoted to the expressiveness of *combined* multiple input modalities. Users are unlikely to receive the information they require, or be able to refer to it appropriately when they receive it, unless the same consideration is given to the input as is increasingly being given to the output.

This paper discusses the relationship between multimodality input and multimedia output, with specific consideration of the implications of multimedia for the form, content and meaning of multimodality input. The potential for adopting a "multi-modelling" approach to multimodality along with the use of metaphors is discussed. Finally, two systems which incorporate the use of metaphors are described.

2. THE IMPLICATIONS OF MULTIMEDIA OUTPUT FOR MULTIMODALITY INPUT
Over recent years, many systems have been developed for the dissemination of information. One example is a multimedia encyclopaedia. Another is the computer-based training system. In such circumstances, there is no longer a *human* information provider. The digital artifact assumes this role. This has consequences for the information receiver. Today, the information receiver, or user, is usually faced with *multimedia* information. This information may be delivered in a highly structured form, where the user is guided to the result of the interaction, so the interaction and control is probably limited. Alternatively, the information may be poorly structured, and therefore a high level of interaction and user control is subsequently required. Neither of these two extreme situations is optimal for the user. They may result in a confused user who feels overwhelmed or insufficiently informed, or is perhaps reduced to being little more than a "page turner" of an "electronic book". The output has a much higher level of expressiveness than that provided by the input.

The user has access to a wide range of modalities through which he or she typically interacts. Interaction should involve *bi-directional* communication. In human-computer interaction, while there is often great richness in the system's output, users' ability to make use of the modalities available to them is severely limited. Moreover, having received the multimedia output, the facilities available for the user to specify to the system what it is that he or she really requires are also typically very limited. Facilities such as pointing, clicking, and maybe simple speech and primitive

gestures are provided for, but almost always each modality is considered in isolation from any other.

To date, most research in multimodal interaction has focused on using each modality separately, or in pairs, i.e. speech and gesture, gesture and gaze. Very little research has been carried out on the basis of a detailed analysis of human-human multimodal interaction. As a result, crucial contextual metadata are overlooked, such as the effect of facial expressions, gestures, or voice intonation on the meaning of an utterance.

However, the above discussion is not meant to imply that a multimodality interface will always be used in a multimodal way by its users. Oviatt [1] points out that human–human communication is typically a mixture of unimodal and multimodal interactions. In a well-designed system, individuals would be able to choose whether to interact multimodally or not. Often, this choice would be made on the basis of the activity being carried out, or the context in which that activity takes place. Though users favour the ability to interact multimodally, they do not always choose to do so, and they usually explore the use of each modality separately and then form their own pattern of interaction.

One problem in human-computer interaction is whether designers should tailor their systems to the user, or users should tailor their interaction patterns to the system. Considering multimodality interaction, the problem is the extent to which we can assume that multimodality will be exploited in a uniform way by different users. In Oviatt's study [1], users adopted either simultaneous or sequential integration patterns when combining speech and pen input. Each user's integration pattern was established early and remained consistent, but nevertheless, each user's pattern was unique. As an aside, it is probably the case that similar individual patterns of usage also apply to users' use of multimedia *output*.

A further important factor in multimodal systems is the extent to which the integration of modalities introduces *redundancy* in the content specified by different modalities. However, redundancy is often *complementarity*. The ability to convey the same information in several different modalities does not imply that a user will use all of these modalities to interact at any one time, but rather may choose which modality or combination of modalities is suitable at the given moment, in the particular context. Likewise, if the system produces output involving multiple media types, the user will often focus on a preferred media format, which may lead to the risk of missing important information. Thus, redundancy is often a useful property of multimedia output. The implications for multimodality user input has yet to be fully explored.

3. THE RELATIONSHIP BETWEEN MODALITIES

To us, the term "multimodal input" implies the existence of simultaneous or temporally co-ordinated expressions in a variety of modalities. The two most frequently combined modalities in human-human multimodal interaction, *speech* and *gesture*, are highly interdependent and synchronized during interaction. They are not always simultaneous, as gesture can often precede speech, or complement it by conveying information that is not explicitly uttered. Such cases typically involve a quick switch from speech to gesture and back to speech. This is accomplished so quickly and blended so naturally that it is perceived as simultaneous.

The view of linguists and some computer scientists that speech is a primary input mode has biased early multimodal systems towards speech input and "point-and-speak" systems. This has rendered speech to be the primary input mode in most multimodal systems in which it is included. Unfortunately, this has led to systems that consider other modalities that are employed as secondary, thus failing to recognize information that is not present in the speech. Speech is not the exclusive carrier of information. Even in a simple "point-and-speak" interface, it is possible to imagine a scenario in which both modalities in a particular activity are an indispensable component of the meaning of the "utterance". Consider telling a system to move a previously marked block of text to a new location:

> "*move that"* [spoken, accompanied by] *pointing to block of text "to there"* [spoken, accompanied by] *pointing to target location*

As this simple example demonstrates, when users interact multimodally they selectively eliminate linguistic complexities and replace them with an interaction pattern, which involves unimodal and multimodal aspects. However, what results is a complex "linguistic" structure in which meaning depends on the temporal and significant relationship between expressions in two modalities.

Different input modalities can be used to specify different content. The different modalities found in emerging technologies that recognise speech, handwriting, manual gesturing, head movement and gaze can significantly differ in the information they specify. They can also differ in their functionality during communication, the ways in which they are integrated with each other and their suitability for incorporation into different interface styles. In some cases, a given modality can be a simple analogue of another, in the sense that there is a direct translation between one and the other. However, in many cases, modalities vary in the degree to which they represent similar information, with some groups of modalities being more similar (speech and writing) than others (speech and facial expression).

4. TOWARDS "MULTI-MODELS" OF MULTIMODALITY INPUT

Multimodality has the potential to facilitate richer interaction styles in both information retrieval and learning environments. However, its true potential will not be realised unless consideration is given to the application of *combined* modalities, both simultaneously and over time. Progress has long been made in the structural and grammatical analysis of *language*, where the term is usually meant in the *uni*modal sense, as it applies to, say spoken English, of which the structural and semantic analysis is, of course a well-established field. However, mixed modality interaction, while drawing on the various languages of speech, gesture, etc., implies that account must be taken of the relationship between the simultaneously expressed statements from each of these languages. For example, the utterance "we'll get this paper finished by this evening", when accompanied by the quickly raised eyebrows of the speaker, might mean something quite different when accompanied by the speaker's reassuring smile. For multimodality interaction, then, the corresponding "grammar" would describe the structure of mixed modality "sentences", and the lexicon would map out the meaning of mixed modality "words". The meaning of an utterance would be inextricably linked with all of the multimodality components of the "utterance" and the relationship between them. In this respect,

what is required is a "*multi-model"* of multimodality communication. Such a model would enable us to specify and interpret mixed-modality inputs, and support an expressiveness and flexibility of input to match that increasingly found in forms of output.

At the time of writing, multimodality interaction in HCI is much less sophisticated than that offered by the combination of speech, gestures and other modalities found in everyday human-human interaction. However, even the standard typing, pointing, and clicking interface offers gestural possibilities (the selection of a portion of text with a mouse is essentially gestural, after all) that have hitherto been almost exclusively applied unimodally. Thus, the central argument of this paper applies to current, as well as future, systems.

Finally, we assert that multimedia output from a system actually *requires* multimodality on the part of the user. Communicating with a system about a diagram, for example, requires more than just speech, text and simple pointing. The effectiveness of a diagram may be lost if the participants in a discussion about that diagram must constantly translate their knowledge of the diagram into an alternative form to express it to the other participants. In other words, a final requirement of the "multi-model" of modality is that it considers the role played by the media that are referred to in by the input, since, for example, even the meaning of a simple gesture such as a wave of the hand will depend partly on properties of the *referent* of that gesture. The diversity of these *properties* of multimedia information will open up new expressive possibilities for multimodal communication in human-computer interaction. The "multi-model" of multimodality communication may provide a framework in which to address such issues.

5. METAPHORS

Metaphors are not a novel feature of HCI in themselves, the *desktop metaphor* being a prime example. However, the advent of multimedia and novel interaction techniques has perhaps overshadowed the effectiveness of metaphors with more focus directed at the core input and output techniques (e.g. speech and gesture input and multimedia output) rather than the underlying mechanisms, which will support them.

The advantages of metaphor usage in interfaces is not always apparent. One path of reasoning can be found in Umberto Eco's interpretation of the words of Aristotle "… the most ingenious and vigorous of Aristotle's conclusions, [is] that the metaphor is not only a means of delight but also, and above all, a tool of cognition." Eco also points out that Aristotle describes the creation of metaphors as " 'a sign of natural disposition of the mind' because knowing how to find good metaphors means perceiving or grasping the similarity of things between each other" (*τό όμοιον θεωρείν*) (Poetics 1459 a6-8) [2]. Two projects are described in the following sections, which use metaphors extensively with a view to exploiting such similarities (both between media objects themselves as well as between real world actions and digital environments).

6. MULTIMODAL STORY CREATION - THE STORY CONDUCTOR

The story conductor under development by Dorothy Rachovides uses metaphors in two ways: (i) a series of visual metaphors are used to represent media types and (ii) the setting and the interaction style are based on metaphors of the orchestral Conductor and the theatrical stage.

The *story conductor*, i.e. the user of the system, is placed in a familiar setting to that of the orchestral conductor but in a virtual world modelled in a sense on the orchestra's stage. This stage serves as a visualisation of the context in which the conductor - user interacts.

This world has an open V shape, formed from three computer monitors. Multimedia objects feature on the two side monitors, and there is a "screen" in the centre of the stage i.e. on the central monitor. The functionality of the "screen" object is based on the context in which it operates, giving the user – conductor the sense of expectancy that all the "visual results" will appear on the screen.

The "instruments", i.e. *media objects,* manipulated by the story conductor represent a considerable range of media types, categorised as follows:

1. *A Sound Gallery*: a "jukebox" represents this sound gallery. Various sound objects, among them being music and environmental sounds, can be chosen.
2. *A Sound Effects Gallery*: a horn represents a series of sounds that can be used in combination with other sounds to emphasise points of the story.
3. *A Dialogue Gallery*: a picture of two people talking represents a series of short phrases that can be used in order to add appropriate voices to the story.
4. *A Film Gallery*: a camera represents a series of films (i.e. digitised video sequences).
5. *An Animation Gallery*: a cartoon character represents a set of characters that can be used in the story.
6. *A Photo Gallery*: A picture book represents a series of pictures that can be used in the story.
7. *Lighting Controls*: A light bulb represents the control of the lighting, for example to show the time of the day or night, or events such as sunrise, lights being switched on when entering a room, etc.
8. *Volume Controls*: A slide bar enables the user to change the volume of sound in any clip in which that sound is applicable, for example making a dialog be heard as a whisper or create a loud siren.

The above visual metaphors define the context of the conductor's world. The conductor's bimanual interaction is based on a vocabulary consisting of functional gestures, which are emblems by nature. Emblems are gestures that have standards of well formedness, a crucial language-like property that other types of gestures and pantomime lack [3]. Gestures are used in combination with eye tracking for media type selection. As implied by the conductor metaphor, the user interacts with the media objects in the same way as the conductor would interact with the musicians of an orchestra, i.e. establishing eye contact to initiate the interaction, and then using bimanual gestures to specify when and how the musicians will play. The orchestral conductor is silent throughout the interaction, but uses body language to convey information to the members of his orchestra. Based on this principle, silent interaction can be used to create multimedia stories. The user focuses on the media object to be

used and then uses gestures to select the particular clip. The story may be previewed during its creation and ultimately played in full. During the creation of the story, the conductor can choose the media that will be played, its ordering and other properties to be applied.

The goal of the user-conductor is therefore to create a story. The initial plan model is quite simple: select first media type, select first clip, select second media type select second clip, and so on. This process continues until the story is completed. However, the user–conductor may be more creative and may wish to adjust the presentation properties of the clips, the order in which they are played, and whether they are played in sequence or concurrently.

7. CONTROLLING MULTIMEDIA – THE VOLUME METAPHOR

Currently, designers must predict the types and level of information that users need unless their system can incorporate an advanced user-modelling system. The lack of a user-orientation in systems usually means that users expend considerable effort in adjusting the vast array retrieved information to suit their own requirements, both in terms of the format and level of detail.

Rogers and Scaife [4] have observed that students consistently admitted to ignoring text at the interface in favour of other media types such as diagrams and video material. Furthermore, when a variety of information is available (e.g. as is usually the case with web-based and multimedia systems), the user is left with the role of interpreting each individual representation and identifying any relationships between the different representations that may be present. Ainsworth [5] also notes that learners have difficulty translating between different representations and often fail to grasp important connections between different modes of representation.

The volume control under development by Zoë Swiderski is a mechanism that aims to address the issues outlined above by providing a basis for enabling multimedia objects to be controlled in an analogous manner to that of a volume control. A volume control (e.g. the Windows desktop volume control) is already a metaphor, which symbolises the notion of sound and provides a representation to enable users to increase and decrease its loudness. This metaphor could be effectively applied to controlling content levels of multimedia information.

Where volume levels can be ‘turned up’ or ‘turned down’, content level can also be ‘turned up’ by providing more detail (possibly by using additional distinct, but informationally related, media objects) or ‘turned down’ (using fewer media items and possibly less information). While addressing the issues of information overload by facilitating the filtering of information either by media type (images, text etc.) or detail (e.g. summary, full description, bullet points etc.), this mechanism has additional benefits. It could allow the expression of relationships between media types. Currently multimedia objects are interpreted as independent entities when it is more often the case that the collection of objects is being used to represent the same notion. Furthermore, if the combined information is examined, it can often present an alternative interpretation, which is lost when interpreting the objects individually.

8. CONCLUSION

This paper has described two systems that exploit multimodality, particularly gestural input. The story conductor features a more explicit use of gestural input, but in a sense, the “volume control” makes the control of detail a gestural activity. Furthermore, both systems exploit cross-modality and inter-medium reference, in that one form of input (gesture, adjusting a slide control, etc.) is converted into corresponding operations in an alternative media (order of shots in a video, level of detail in a text or diagram).

The two systems described reflect a “multi-modelling” approach to interaction whereby metaphors are used to represent relationships between input techniques, output presentation and the corresponding actions associated with their manipulation. This will enable users to exert full control over their digital environments.

The authors believe that the metaphors described in this paper can be seen as a step towards realising applications that are truly *multimedia* systems. They lay the foundations for exploiting not only media objects, but also the *relationships between them*. The systems described not only exploit analogous representations, but also analogous *processes (adjusting volume, conducting)* and *activities (controlling, directing).*

Despite many claims being made for the power of metaphors in human-computer interfaces, there are few examples of the creative use of non-localised metaphors apart from the famous *desktop metaphor.* In this paper we have discussed the methodological background and the realisation of two systems, which transfer real-world interactions to novel metaphors thus bridging the gap between artificial environments and user interactions.

9. ACKNOWLEDGEMENTS

The authors are supported by the Distributed Multimedia Research Group. Zoë Swiderski’s research is also supported by the Engineering and Physical Sciences Research Council (EPSRC) and British Telecom Laboratories.

10. REFERENCES

[1] Oviatt, S. (1999) ”Ten Myths of Multimodal Interaction.” Communications of the ACM 42(22), pp.74-81.

[2] Eco, U. (1984) “Semiotics and the Philosophy of Language”, Bloomington, U.S.A.: Indiana University Press.

[3] McNeil, D. (1992) “Hand and Mind: What Gestures Reveal About Thought”, University of Chicago Press, Chicago.

[4] Rogers, Y. and Scaife, M. (1998) “How Can Interactive Multimedia Facilitate Learning?”, Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. J. Lee, AAAI. Press: Menlo Park, CA.

[5] Ainsworth, S. (1999) “The Functions of Multiple Representations”, Computers & Education 33, pp.131-152.