

NIELS OLE BERNSEN

MULTIMODALITY IN LANGUAGE AND SPEECH SYSTEMS - FROM THEORY TO DESIGN SUPPORT TOOL

1. INTRODUCTION

This paper presents an approach towards achieving fundamental understanding of unimodal and multimodal output and input representations with the ultimate purpose of supporting the design of usable unimodal and multimodal human-human-system interaction (HHSI). The phrase ‘human-human-system interaction’ is preferred to the more common ‘human-computer interaction’ (HCI) because the former would appear to provide a better model of our interaction with systems in the future, involving (i) more than one user, (ii) a complex networked system rather than a (desktop) ‘computer’ which in most applications may soon be a thing of the past, and (iii) a system which increasingly behaves as an equal to the human users (Bernsen, 2000). Whereas the enabling technologies for multimodal representation and exchange of information are growing rapidly, there is a lack of theoretical understanding of how to get from the requirements specification of some application of innovative interactive technology to a selection of the input/output modalities for the application which will optimise the usability and naturalness of interaction. Modality Theory is being developed to address this, as it turns out, complex and thorny problem starting from what appears to be a simple and intuitively evident assumption. It is that, as long as we are in the dark with respect to the nature of the elementary, or unimodal, modalities of which multimodal presentations must be composed, we do not really understand what multimodality is. To achieve at least part of the understanding needed, it appears, the following objectives should be pursued, defining the research agenda of Modality Theory (Bernsen, 1993):

- (1) To establish an exhaustive taxonomy and systematic analysis of the unimodal modalities which go into the creation of multimodal *output* representations of information for HHSI.
- (2) To establish an exhaustive taxonomy and systematic analysis of the unimodal modalities which go into the creation of multimodal *input* representations of information for HHSI. Together with Step (1) above, this will provide sound foundations for describing and analysing any particular system for interactive representation and exchange of information.

- (3) To establish principles for how to legitimately combine different unimodal output modalities, input modalities, and input/output modalities for usable representation and exchange of information in HHSI.
- (4) To develop a methodology for applying the results of Steps (1) – (3) above to the early design analysis of how to map from the requirements specification of some application to a usable selection of input/output modalities.
- (5) To use results in building, possibly automated, practical interaction design support tools.

The research agenda of Modality Theory thus addresses the following general problem: given any particular set of information which needs to be exchanged between user and system during task performance in context, identify the input/output modalities which constitute an optimal solution to the representation and exchange of that information. As we shall see and as has become obvious from the literature on the subject through the 1990s, this is a hard problem, for two reasons. Firstly, already at the level of theory there are a considerable number of unimodal modalities to consider whose combinatorics, therefore, is quite staggering. Secondly, when it comes to applying the theory in development practice, the context of use of a particular application must be taken thoroughly into account in terms of task, intended user group(s), work environment, relevant performance and learning parameters, human cognitive properties, etc. A particular modality is not simply good or bad at representing a certain type of information – its aptness for a particular application very much depends on the context. This adds to the combinatorics generated by the theory an open-ended space of possibilities for consideration by the developer, a space which, furthermore, despite decades of HCI/HHSI research remains poorly mastered, primarily because such is the nature of engineering as opposed to abstract theory.

Given the many different and confusing ways in which the terms ‘media’ and ‘modality’ are being used in the literature, it should be made clear from the outset what these terms mean in Modality Theory.

A *medium* is the physical realisation of some presentation of information at the interface between human and system. Media are closely related to the classical psychological notion of the human “sensory modalities”, i.e. vision, hearing, touch, smell, taste, and balance. Thus, the graphical medium is what humans or systems see, i.e. light, the acoustic medium is what humans or systems hear, i.e. sound, and the haptic medium is what humans or systems touch. Physically speaking, graphics comes close to being photon distributions, and acoustics comes close to being sound waves. In physical terms, haptics is obviously more complex than those two and no attempt will be made here to provide a physical description of haptics beyond stating that haptics involve touching. Media are symmetrical between human and system: a human hears (output) information expressed by a system in the acoustic medium, a system sees (input) information expressed by a human in the graphical medium (in front of a camera, for instance), etc. In the foreseeable future, information systems will mainly be using the three input/output media of graphics, acoustics and haptics. These are the media addressed by Modality Theory so far. To forestall a possible

misunderstanding, the medium of graphics includes both text and “graphics” in the sense of images, diagrams, graphs etc. (see below).

The term *modality* (or *representational modality* as distinct from the sensory modalities of psychology) simply means “mode or way of exchanging information between humans or between humans and machines in some medium”. The reason why any approach to multimodality is bound to need both of the notions of media and modalities is that media only provide a very coarse-grained way of distinguishing between the many importantly different physically realised kinds of information which can be exchanged between humans and machines. For instance, a graphical output image and a typed Unix output expression are both output graphics, or an alarm beep and a synthetic spoken language instruction are both output acoustics, even though those representations have very different properties which make them suited or unsuited, as the case may be, for different tasks, users, environments, etc. It seems obvious, therefore, that we need a much more fine-grained breakdown among available representational modalities than what is offered by the distinction between different media. The notion of representational modalities just introduced is probably quite close to that intended by many authors. As early as ten years ago, Hovy & Arens (1990), observed that, e.g., tables, beeps, written and spoken natural language may all be termed ‘modalities’ in some sense.

Some additional terms are clarified briefly to avoid misunderstandings later on. *Input* means interactive information going from A to B and which has to be decoded by B. A and B may be either humans or systems. Typically in what follows, A will be a human and B will be a system. It is thus taken for granted that we all know a lot about what can take place in an interaction in which both A and B are humans, or in which several humans interact together as well as interacting with a system. *Output* means interactive information going from B (typically the machine) to A (typically a human). The term *interactive* emphasises that A and B exchange information deliberately or that they communicate. In this central sense of ‘interaction’, it is *not* interaction when, e.g., a surveillance camera tracks and records an intruder unbeknownst to that intruder. It should also be noted that Modality Theory is about (representational) modalities and not about the *devices* which machines and humans use when they exchange information, such as hands, joysticks, or sensors. The positive implication is that the world of modalities is far more stable than the world of devices and hence much more fit for stable theoretical treatment. The negative implication is that Modality Theory in itself does not address the – sometimes tricky – issues of device selection which may arise once it has been decided to use a particular set of input/output modalities for an application to be built. On a related note, the theory has nothing to say about how to do the detailed design (aesthetically or otherwise) of *good* output presentations of information using particular modalities. As the colourful field of animated interface agents illustrates at present, it is one thing to safely assume that these virtual creatures have strong potential for certain kinds of application but quite another to demonstrate that potential through successful design solutions. Finally, it should be pointed out that when we refer to the issue of which modalities to use for exchanging information of some kind, ‘information’ means information in the abstract, as in ‘medical data entry information’, information in a new interactive game to be developed, or

geographical information for the blind. Such descriptions are commonplace, and they leave more or less completely open the question of which modalities to use for the particular purpose at hand.

Modality Theory is, in fact, a century-old subject which easily antedates even the Babbage machine. People have interacted with information presentations on pyramids, in books or in magazines for a very long time. For instance, output modality analysis has a long tradition in the medium of (static) graphics. Outstanding examples are the results achieved on static graphic graphs (Bertin, 1983; Tufte, 1983, 1990). Given today's and tomorrow's input/output technologies, however, we need to address a much wider range of modalities and modality combinations. This is a truly collective endeavour. Modality Theory and the methodology for its practical application is an attempt to provide and illustrate a reasonably sound theoretical framework for integrating the thousands of existing and emerging individual contributions to our understanding of the proper use of modalities in interaction design and development.

This chapter addresses, at different levels of detail, all of the five points on the research agenda of Modality Theory described above, as follows. Section 2 presents the generation of the taxonomy for unimodal output modalities at several levels of abstraction. Section 3 proposes a draft standard representation format for modality analysis. Section 4 presents ongoing work on generating the taxonomy for input modalities. This part of the research agenda has proved to be hard and full of surprises. Section 5 presents our first full-scale application of the theory in its role as interaction design support. Finally, Section 6 concludes by discussing empirical and theoretical approaches for how to deal with the combinatorial explosion of modality combinations in multimodal systems. Due to space limitations, it has sometimes been necessary to refer to other publications for more detail.

For the obvious reason, the modality illustrations to be provided below are all presented in static graphics just like the present text itself. Current literature tends to focus on input/output modalities which are technically more difficult to produce, and which are less explored, than the static graphics modalities. It may be worthwhile to stress at this point, therefore, that all or most of the modality concept to be introduced below in fact do generalise to all possible modalities in the media of graphics, acoustics and haptics.

2. A TAXONOMY OF UNIMODAL OUTPUT

The taxonomy of unimodal output modalities to be presented is not the only one around although it appears to be the only one which has been generated from basic principles rather than being purely, or mainly, empirical in nature. In addition, its scope is as broad as that of any other attempt in the literature. A solid taxonomy based on decades of practical experience is Tufte's taxonomy of data graphics (Tufte, 1983). Twyman (1979) presents a taxonomy of static graphics representations (text, images, etc.). It is of wider scope than Tufte's taxonomy and, like the latter, based on long practical experience. Still in the static graphics domain, (Lohse et al., 1991) present a taxonomy which is based on experiments in which

they studied how subjects intuitively classify sets of static graphic representations. Of much broader scope, comparable to that Modality Theory, are the lists of modalities and modality combinations in (Benoit et al., 2000). These lists simply enumerate modalities found in a large sample of the literature on multimodality from the 1990s.

A taxonomy of representational modalities is a way of carving up the space of forms of representation of information based on the observation that different modalities have different properties which make them suitable for exchanging different types of information between humans and systems. Let us assume that modalities can be either unimodal or multimodal and that multimodal modalities are combinations of unimodal modalities, i.e. can be completely and uniquely defined in terms of unimodal modalities. These assumptions suggest that if we want to adopt a principled approach to the understanding and analysis of multimodal representations, we have to start by generating and analysing unimodal representations. Generation comes first, of course. So the crucial issue at this point is how to generate the unimodal modalities. Basically, two approaches are possible, one purely empirical, the other hypothetico-deductive, i.e. through empirical testing of a systematic theory or hypothesis. Note that both approaches are empirical ones, just in different ways. Although the purely empirical approach has a strong potential for providing relevant insights and is being used widely in the field, it appears that no stable scientific taxonomy was ever created in a purely empirical fashion from the bottom up. If, for instance, experimental subjects are asked to spontaneously cluster a more or less randomly selected set of analogue static graphic representations (Lohse et al., 1991), the subjects may classify according to different criteria, they may be unable to express the criteria they use, and in the individual subject the criteria that are being applied may be incoherent. An alternative to the purely empirical approach is to generate modalities from basic principles and then test through intuition, analysis, and experiment whether the generated modalities satisfy a number of general requirements. If not, the generative principles will have to be revised. Let us adopt the generative approach in what follows. We want to identify a set of unimodal *output* modalities which satisfies the following requirements:

- (1) *completeness*, such that any piece of, possibly multimodal, output information in the media of graphics, acoustics and haptics can be exhaustively described as consisting of one or more unimodal modalities;
- (2) *uniqueness*, such that any piece of output information in those media can be characterised in only one way in terms of unimodal modalities;
- (3) *relevance*, such that the set captures the important differences between, e.g., beeps and spoken language from the point of view of output information representation; and
- (4) *intuitiveness*, such that interaction developers recognise the set as corresponding to their intuitive notions of the modalities they need or might need. Given the practical aims of Modality Theory, it is of crucial importance to operate with intuitively easily accessible notions without sacrificing systematicity.

Multimodality in Language and Speech Systems

Granström, B.; House, D.; Karlsson, I. (Eds.)

2002, X, 243 p., Hardcover

ISBN: 978-1-4020-0635-7