Multimodality Theory

Niels Ole Bernsen

Natural Interactive Systems Lab., University of Southern Denmark

Introduction

Since the early 2000s, interactive systems designers, developers, and evaluators have increasingly been focusing on tapping into the enormous potential of multimodality and multimodal systems. In this chapter, we discuss and define the notions of modality and multimodality followed by presentation of a theory and taxonomy of all modalities of information representation in the media of graphics, acoustics, and haptics. The final part of the chapter discusses practical uses of the theory, in particular with respect to the issue of modality choice and the idea of predictive construction of multimodal representation from unimodal modalities.

What is a Multimodal System?

An Enigma

Somewhat surprisingly, given the enormous attention to multimodal systems world-wide, there is still a certain amount of confusion about what multimodality actually is. To understand multimodality, it seems important to understand why this confusion persists.

The term 'modality' itself is not overly informative. One of its relevant senses is to be a *manner of something*; another, the so-called *sensory modalities* of psychology, i.e., vision, hearing, etc., but, as we shall see, the

'modalities' of 'multimodality' cannot be reduced to the sensory modalities of psychology.

Historically, the term 'modality' already appears in something close to its present-day meaning in Bolt's early paper on advantages of using combined speech and deictic gesture [Bolt 1980]. Another early appearance is in [Hovy and Arens 1990] who mention written text and beeps as different examples of forms of representing information as output from, or input to, computer systems.

Today, the following two definitions or explanations of multimodality are perhaps among the more widespread ones.

1. A multimodal system is a system which somehow involves several modalities.

This is both trivially true and quite uninformative about what multimodality actually is. The definition does, however, put a nice focus on the question: what are *modalities*? It would seem rather evident that, if something – a system, interaction, whatever - is *multi*modal, there must be other things which are *uni*modal and which *combine* to make that something multimodal. However, another explanation often seems to be lurking in the background:

2. A multimodal system is a system which takes us beyond the ancient and soon-to-become-obsolete GUI (Graphical User Interfaces) paradigm for interactive systems. Multimodal systems represent a new and more advanced paradigm for interactive systems, or, quoting [Oviatt and Cohen 2000]: "Multimodal systems are radically different than standard GUIs".

This view probably helps explain the sense of novelty and adventure shared by many researchers and developers of multimodal systems today. However, as an explanation of multimodality, and although still near-vacuous as to what multimodality is, this one is actually false and seriously misleading. Even though there are several more informative attempts at definitions and explanations of multimodality around, these all imply that GUI systems are *multimodal* systems. Furthermore, GUI-based interaction is far from obsolete, it's a useful paradigm which is just far more familiar and better explored than most other kinds of multimodal interaction.

A Solution

To see why multimodality still has a lot of unexplored novelty to it, why the sensory modalities of psychology only far too insufficiently account for the modalities there are, and why GUIs are nevertheless multimodal, we need a theory of what's involved. The basic notions of that theory are

interaction, media, modalities, and information channels, which we will now look at in turn.

Human-Computer Interaction and Media

Human-computer 'interaction' is, in fact, *exchange of information* with computer systems, and is of many different kinds which we need not go into here (for a taxonomy, see [Bernsen and Dybkjær, in press(a)]). Exchange of information is ultimately a physical process. We never exchange information in the abstract even if we are very much used to thinking and reasoning about information in abstract terms. When humans exchange information, the information is *physically instantiated* in some way, such as in sound waves, light, or otherwise. In fact, humans are traditionally said to have five or six senses for physically capturing information, i.e., *sight*, *hearing*, *touch*, *smell*, *taste*, and, if this one is counted as well, *proprioception*. These are the sensory modalities of psychology.

Correspondingly, let us say that information, to be perceptibly communicated to humans, must be instantiated in one or more of the following six *physical media*, i.e.:

- light / vision / graphics;
- sound waves / hearing / acoustics;
- mechanical touch sensor contact / touch / haptics;
- molecular smell sensor contact / smell / olfaction;
- molecular taste sensor contact / taste / gustation; and
- proprioceptor stimulation (as when you sense that you are being turned upside down).

It is useful to define each medium through a *triplet* as above, the first element referring to the physical information carrier, the second to the perceptual sense needed for perceiving the information, and the third one to information presentation in that medium. Note that this entails a non-standard use of the term 'graphics' in English, because the graphics modalities come to include not only graphical images and the like, but also, in particular, ordinary text.

To the above we need to add a point about human *perceptual thresholds*. If a human or system is to succeed in getting dispatched physically instantiated information perceived by a *human* recipient, the instantiation must respect the limitations of the human sensors. For instance, the human eye can only perceive light within a certain band of electromagnetic frequency (approximately 380-780 Nm); the human ear can only perceive

sound within a certain Herz band (approximately 18-20.000 Herz); touch information must be above a certain mechanical force threshold to be perceived and its perception also depends on the density of touch sensors in the part of human skin exposed to the touch; etc. In other words, issuing an ultrasonic command to a soldier will have no effect because no physically instantiated information will be perceived by the soldier.

Modalities

We can now define a 'modality' in a straightforward way:

3. A modality or, more explicitly, *a modality of information representation*, is a way of representing information in some physical medium. Thus, a modality is defined by its physical medium and its particular "way" of representation.

It follows from the definition that modalities do not have to be perceptible by humans. Even media do not have to be perceptible to humans. So modalities don't even have to be represented in physical media accessible to humans, since there are physical media other than the six media listed above, all of which are partly accessible to humans. In what follows, we focus on modalities perceptible to humans unless otherwise stated.

Now we need to look at those "ways" of information representation because these are the reason for having the notions of modalities and multimodality in the first place. The simple fact is that we need those "ways" because humans use *many* and *very different* modalities for representing information instantiated in *the same* physical medium and hence perceived by *the same* human sensory system. Consider, for example, *light* as physical medium and *vision* as the corresponding human sense. We use vision to perceive *language text, image graphics, facial expression, gesture,* and much more. These are *different* modalities of information representation instantiated in *the same* physical medium.

Although the above example might be rather convincing on its own, it is useful to ask *why* the mentioned modalities are considered to be different. There are two reasons. The first one is that all modalities differ in *expressiveness,* i.e., they are suited for representing different kinds of information. A photo-realistic image, for instance, is generally far better at expressing how a particular person looks than is a linguistic description. The second reason is to do with the properties of the *recipient* of the information represented, perceptual, cognitive, and otherwise. For instance, since the blind do not have access to the medium of light, we primarily use the acoustic and haptic media to represent information to the blind. But again, since different modalities differ in expressiveness, we need all the modalities we can implement for information representation for the blind, such as

speech and Braille test for linguistic information, haptic images, etc. Even if two particular representations are completely *equivalent* in information content and instantiated in the *same* medium, the human perceptual and cognitive system works in such a way that each of the representations may be preferable to the other depending on the purpose of use. For instance, if we want a quick overview of trends in a dataset, we might use, e.g., a static graphics *bar chart*, but if we want to study the details of each data point, we might prefer to look at an informationally equivalent static graphics *table* showing each data point.

As the above examples suggest, we may actually have to reckon with a considerable number of different modalities.

Input and Output Modalities

Given the considerable number of different modalities there are, it is good practice to specify if some particular modality is an *input* modality or an *output* modality. For instance, what is a "spoken computer game"? This phrase does not reveal if the system takes speech input, produces speech output, or both, although these three possibilities are very different from one another and circumscribe very different classes of systems. By tradition, we say that, during interaction, the *user* produces *input modalities* to the system and the *system* produces *output modalities* to the user. And the point is that, for many multimodal systems, the set of input modalities is often different from the set of output modalities.

Another important point about input and output modalities is that we can form the abstract concepts of (i) the class of *all possible input modalities that can be generated by humans* and (ii) the class of *all possible output modalities that can be perceived by humans*. These two classes are *asymmetrical*. This follows from the limitations of the human sensory system as a whole both as regards the physical media it is sensitive to and its biologically determined sensory thresholds. Computers, on the other hand, can be made far more discriminative than humans on both counts. Computers can sense X-rays and other exotic "rays" (alpha, beta, gamma etc.), radar, infrared, ultraviolet, ultrasound, voltage, magnetic fields, and more; can sense mechanical impact better than humans do; and might become capable of sensing molecular-chemical stimuli better than humans do. This means that computers: (i) have more input modalities at their disposal than humans have; (ii) have or, in some cases, probably will get, far less restrictive sensory thresholds for perceiving information in some particular modalities than humans; and (iii) can output information that humans are incapable of perceiving. This is useful for interactive systems design because

it allows us to think in terms of, e.g., human interaction with a magnetic field sensing application which no human could replace.

This point about input/output modality asymmetry raises many interesting issues which, however, we shall ignore in the following. Let us simply stipulate that we will only discuss multimodal interaction in *maximum symmetrical conditions,* i.e., we will discuss multimodal input/output interaction based on the physical media humans can perceive and to the extent that humans can perceive information instantiated in those media.

Unimodal and Multimodal Interactive Systems

We can now define a multimodal interactive system.

4. A *multimodal interactive system* is a system which uses at least two different modalities for input and/or output. Thus, [IM1,OM2], [IM1, IM2, OM1] and [IM1, OM1, OM2] are some minimal examples of multimodal systems, *I* meaning input, *O* output, and *Mn* meaning a specific modality *n.*

Correspondingly,

5. A *unimodal interactive system* is a system which uses the same single modality for input and output, i.e., [IMn, OMn].

An over-the-phone spoken dialogue system is an example of a unimodal system: you speak to it, it talks back to you, and that's it. Other examples are a Braille text input/output dialogue or chat system for the blind, or a system in which an embodied agent moves as a function of the user's movements. There are lots more, of course, if we make creative use of all the modalities at our disposal. Still, the class of potential multimodal systems is exponentially larger than the class of potential unimodal systems. This is why we have to reckon with a quasi-unlimited number of new modality combinations compared to the GUI age.

Why GUIs are Multimodal

It is probably obvious by now why GUI systems are multimodal: standard GUI interfaces take *haptic* input and present *graphics* output. Moreover, both the haptic input and the graphics output involves a range of individually different modalities.

Which Modalities are There?

Given the definition of a *modality of information representation* as a way of representing information in a particular physical medium, and given the

limitations of human perceptual discrimination we have adopted as a frame for the present discussion: which and how many input and output modalities are there? From a theoretical point of view, we would like to be able to: (i) identify all *unimodal* or elementary modalities which could be used to build multimodal interfaces and enable multimodal interaction; (ii) group modalities in one or several sensible ways, hierarchically or otherwise; and (iii) provide basic information on each of them.

From a practical point of view in design, development, and evaluation, we would like to have: (iv) a practical toolbox of all possible unimodal input/output modalities to choose from; (v) guidelines or something similar for which modality to use for a given development purpose; and (vi) some form of generalisation or extension from unimodality to multimodality.

If possible, the contents of the theory and the toolbox should have various properties characteristic of scientific theory, such as being: transparently derived from unambiguous first principles, exhaustive, well-structured, and empirically validated. We do, in fact, have much of the above in *modality theory*, a first version of which was presented in [Bernsen 1994]. Modality theory will be briefly presented in this section. Other fragments contributing to the desiderata listed above, as well as fragments missing, are discussed in the final section of this chapter.

Deriving a Taxonomy of Input/Output Modalities

Table 1 shows a taxonomy of input/output modalities. The scope of the taxonomy is this: it shows, in a particular way to be clarified shortly, all possible modalities in the three media of *graphics*, *acoustics* and *haptics*, which are currently the all-dominant media used for exchanging information with interactive computer systems. Following the principle of *symmetry* introduced above, the taxonomy only shows modalities that are perceptible to humans. The taxonomy is claimed to be *complete* in the sense that all possible modalities in those media are either shown in the taxonomy or can be generated from it by further extension downwards from the generic level. What this means will become clear as we proceed.

The taxonomy is represented as a tree graph with four hierarchical levels, called *super level*, *generic level*, *atomic level*, and *sub-atomic level*, respectively. The *second-highest* (generic) level is derived from basic principles or hypotheses. For the detailed derivation, see [Bernsen 2002]. In what follows, we sketch the general ideas behind the derivation in the form of a meaning representation tree (Figure 1) and then describe the taxonomy itself.

Table 1. A taxonomy of input and output modalities (next page).

Super level	Generic Level	Atomic level	Sub-atomic level
Linguistic modalities	1. Sta. an. graphic elements		
	2. Sta-dyn an. acoustic elements		
	3. Sta-dyn an. haptic elements	4a. Sta.-dyn. gest. discourse	
		4b. Sta.-dyn. gest. lab.	5a1. Typed text
	4. Dyn. an. graphic	4c. Sta.-dyn. gest. notation	5a2. Hand-writ text
		5a. Written text	5b1. Typed lab.
	5. Sta. non-an. graphic	5b. Written lab.	5b2. Hand-writ lab.
		5c. Written notation	5c1. Typed not.
		6a. Spoken discourse	5c2. Hand-writ not.
	6. Sta.-dyn. non-an. acoustic	6b. Spoken lab.	
		6c. Spoken notation	Legend **an** = analogue **dyn** = dynamic **gest** = gesture **lab** = labels /keywords **non-an** = non-analogue **not** = notation **sta** = static **writ** = written
		7a. Haptic text	
	7. Sta.-dyn. non-an. haptic	7b. Haptic lab.	
		7c. Haptic notation	
		8a. Dyn. written text	
	8. Dyn. non-an. graphic	8b. Dyn. written lab.	
		8c. Dyn. written notation	
		8d. Sta.-dyn. spoken discourse	
		8e. Sta.-dyn. spoken lab.	
		8f. Sta.-dyn. spoken not.	
Analogue modalities	9. Static graphic	9a. Images	
		9b. Maps	
		9c. Compos. diagrams	
		9d. Graphs	
		9e. Conceptual diagrams	
	10. Sta.-dyn. acoustic	10a. Images	
		10b. Maps	
		10c. Compos. diagrams	
		10d. Graphs	
		10e. Conceptual diagrams	
	11. Sta.-dyn. haptic	11a. Images	
		11b. Maps	
		11c. Compos. diagrams	
		11d. Graphs	
		11e. Conceptual diagrams	
	12. Dynamic graphic	12a. Images	12a1. Facial expression
		12b. Maps	12a2. Gesture
Arbitrary modalities	13. Static graphic	12c. Compositional diagrams	12a3. Body action
	14. Sta.-dyn. acoustic	12d. Graphs	
	15. Sta.-dyn. haptic	12e. Conceptual diagrams	
	16. Dynamic graphic		
Explicit structure modalities	17. Static graphic		
	18. Sta.-dyn. acoustic		
	19. Sta.-dyn. haptic		
	20. Dynamic graphic		

Basic Concepts

The taxonomy assumes that, within its scope, the meaning of physically instantiated information to be exchanged among humans or between humans and systems, can be categorised as belonging to one of the categories shown in Figure 1.

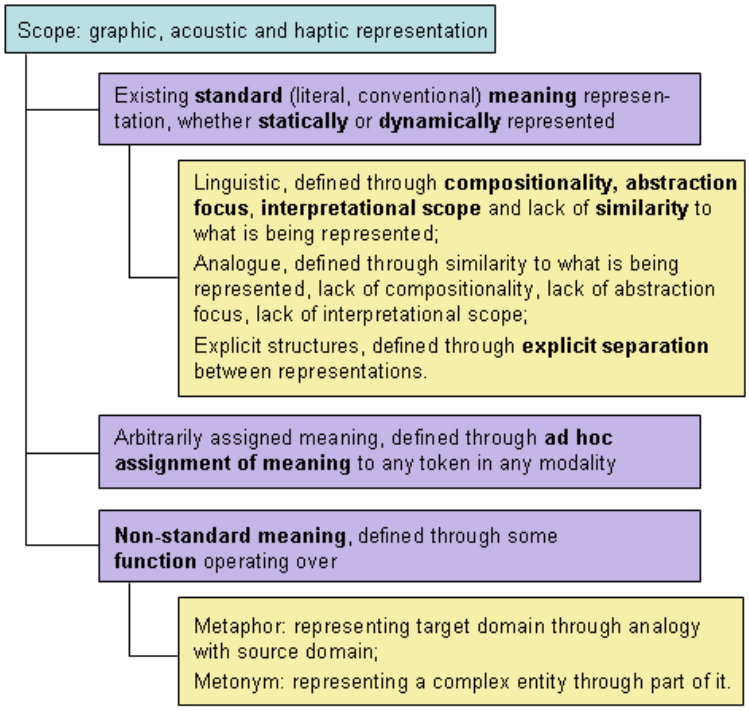

Figure 1. Varieties of meaning representation.

Figure 1 says that, at this stage, modality theory addresses meaning represented in graphics, acoustics and haptics, and that meaning representation is either *standard*, in which case it is either linguistic, analogue or explicit structures, or *arbitrary*; or meaning is *non-standard*, in which case it can be viewed as a result of applying some function, such as the functions used to create metaphorical or metonymic meaning. We now briefly explain and illustrate each concept in the meaning representation tree.

Standard meaning is (i) shared meaning in some (sub-)culture. Shared meaning is basic to communicative interaction because it allows us to represent information in some modality in an already familiar way so that people in our (sub-)culture will understand. Secondly, (ii) standard mean-

ing is opposed to (shared but) non-standard meaning as explained below. Words in the vocabulary in some language, for instance, have standard meanings which are explained in dictionaries and thesauri. An image of a computer would not have *that* meaning to Neanderthal man.

Static/dynamic: static representation is not defined in physical terms but, rather, as that which the recipient can perceptually inspect for as long as it takes, such as a GUI output screen, a blinking icon, or an acoustic alarm which must be switched off before it stops. Dynamic representations do not allow this freedom of perceptual inspection, such as a ringing telephone which may stop ringing at any moment.

Compositionality is a standard concept in linguistic analysis according to which linguistic meaning can be viewed, by way of approximation, at least, as built rule-by-rule from syntax to semantics [Jurafsky and Martin 2000]. For instance, the sentence "Mary loves John" is built in this way and systematically changes meaning if the word order is reversed, as in "John loves Mary".

Abstraction focus is the ability of language to focus meaning representation at any level of abstraction. If we write, e.g.: "A woman walks down the stairs", this is perfectly meaningful even though we are not told things like who she is, how she looks or walks, what the stairs and their surroundings look like, or whether or not the stairs go straight or turn right or left. Language can do that.

Interpretational scope: to continue the previous example, what we tend to do when reading the sentence is to construct our own (analogue, see below) representation. My representation may be very different from yours and none of the two are substantiated by what the declarative sentence about the woman walking down the stairs actually says. This is interpretational scope: we are both "right" but only as far as the standard meaning of the sentence goes. For more about interpretational scope and abstraction focus, see [Bernsen 1995]. By contrast to these properties of linguistic representation,

Analogue representation is defined through similarity between a representation and what is being represented. A drawing of a cow more or less resembles a cow – if not, we have the right to ask if what is being represented in the drawing is really a cow. However, the *word* "cow" (German: Kuh, French: vache, Danish: ko) does not resemble a cow at all. Both the drawing and the word in some language are representations rather than the real thing, of course, the difference being that the drawing is an *analogue* representation whereas the second, linguistic representation is *non-analogue*.

Modality theory also deals with more tenuous analogue relationships than those of photo-realistic images to what they represent, such as be-

tween a line diagram and what *it* represents, or between the click-clicks of a Geiger counter, or the acoustic signatures of the sonar, and what *they* represent.

Explicit separation: this notion may not look much because it deals with what we often do when creating, e.g., tables or matrices. All we do is separate columns and rows using more or less straight lines. This is often very useful for separation and grouping purposes, however, and GUIs, for instance, are full of explicit structures – in windows using multi-layered explicit structures, in pull-down menus, etc. However, explicit structures are also useful in other modalities, such as when we use a beep to mark when the user can speak in non-barge-in spoken dialogue systems, or, in particular, in haptics for the blind because the blind do naturally group objects at-a-glance as the seeing do. Grey colour is used to support perceptual grouping in Table 1.

Ad hoc assignment of meaning: spies, for instance, always did that to avoid that anyone else could understand their communication; kids do it when playing the game of having to say "yes" when you mean "no" and vice versa until you flunk it; and we all do this when, e.g., using boldface, italics, font size, and other means to assign particular (non-shared) meanings to text items. If an ad hoc assigned meaning catches on, which is what happens when new phenomena get an accepted name in a language, it becomes standard meaning.

Non-standard meaning function: although we tend not to realise, human communication is shot through with non-standard meaning [Lakoff 1987]. The reason we tend not to realise is that, e.g., metaphors which began their career as new creative inventions tend to become *dead metaphors* like when we speak of the "shoulder" of a mountain or the "head" of a noun phrase. Once dead, it takes a special focus to retrieve their origin and the words just behave as conventional standard meanings, like most Chinese characters which began their career as analogue signs. Typologies of non-standard meaning typically view each type of non-standard meaning as being created from standard meaning through application of some particular function, such as *metaphor* – e.g., "He blew the top" – using *analogy* with water cooling in a car; or *metonomy* – e.g., "The White House issued a statement saying ..." – using the familiar physical entity of The White House as *part-representing-the-whole* for the executive branch of the US government).

Modality Taxonomy

The modality taxonomy is derived from the definitions and distinctions introduced in the previous paragraph. More specifically, the relationship between Figure 1 and the taxonomy in Table 1 is as follows:

First, Figure 1 introduces a set of orthogonal distinctions which are aimed to capture the core of what it is to represent information in the physical media scoped by the theory, i.e., graphics, acoustics and haptics.

Secondly, based on those distinctions, simple combinatorics mostly account for the derivation of the taxonomy's *generic level*. All other taxonomy levels are generated from the generic level. The qualification 'mostly' refers to the fact that, since the goal of derivation is to arrive at a practical toolbox of unimodal modalities which is reasonably intuitive to use by interactive systems developers, some *fusions* of otherwise separate derived categories have taken place [Bernsen 2002]. However, these fusions are all reversible, and simply so, should future multimodal interactive systems development proceed in novel ways. A typical indicator in the taxonomy that such a *pragmatic fusion* has taken place is the definition of a modality as static/dynamic (**sta.-dyn.** in the taxonomy table), meaning that there is currently no useful point in maintaining separate modalities for static and dynamic representations of the kind specified.

Thirdly, the taxonomy does not distinguish between *standard meaning* and *non-standard meaning derivations* from standard meaning representations. In fact, it cannot because these are physically indistinguishable. The desktop metaphor representation, for instance, is a 2D static graphics plane with analogue icons, labels/keywords, explicit structures used for composing windows and other structures, etc. The fact that this representation is intended by its designers to serve as a metaphor is due to its designed similarity with an ordinary desktop. We might describe non-standard meaning, being derived from standard meaning, as a sort of third dimension relative to the 2D taxonomy. Had the taxonomy been 3D, you would have the desktop metaphor stretching out in the third dimension from the modality *analogue static graphic image* (9a).

Taxonomy Levels

Before we walk through the taxonomy levels, it is important to set one's mind to *unconventional* in order not to miss its richness. Consider that, among many other things, physically instantiated representations may be either *1D, 2D or 3D*, that analogue images, diagrams, graphs, etc., can be *acoustic* or *haptic* and not just *graphic*, and that *time*, as well as the presence or absence of *user control* of what is being represented, is essential to

the distinction between static and dynamic representation. An animated interface character (or agent), for instance, is a (2D or 3D) analogue dynamic graphic (12) image (12a) whose modalities are facial expression (12a1), gesture (12a2) and body action (12a3). And what might an acoustic compositional diagram be used for (10c)?

Super level: since the taxonomy is generated from standard meaning at the *generic level*, the top *super level* modalities only represent one among several possible *classifications* of the derived generic-level modalities. The actual classification in the figure is in terms of *linguistic*, *analogue*, *arbitrary* and *explicit structure* modalities, respectively. However, it is perfectly possible and straightforward to re-classify the generic level modalities in terms of, e.g., the underlying media, getting a super level consisting of *graphics*, *acoustics* and *haptics*, respectively, or in terms of the static/dynamic distinction. The contents of the taxonomy will remain unchanged but the structure of the tree will be modified.

In the taxonomy shown above, super-level *linguistic modalities* represent information in some natural or formal language. *Analogue modalities* represent information by providing the representation with some amount of similarity with what it represents. *Arbitrary modalities* are representations which get their meaning assigned ad hoc when they are being introduced. *Explicit structure modalities* structure representations in space or time, as when information is structured and grouped by boxes-within-boxes in a GUI window.

Generic level: relative to the super level, the generic level expands the super level modalities by means of distinctions between static and dynamic modalities, and between the three physical media. Looking at the generic-level modalities, it is hard to avoid noticing that, in particular, many, if not all of the linguistic and analogue modalities are rather unfamiliar in the sense that we are not used to thinking in terms of them. This is true: they are theoretically derived abstractions at a level of abstraction which most of us visit quite infrequently. It's like trying to think and reason in terms of *furniture* instead of in the familiar terms tables, chairs and beds. In general, the atomic-level modalities are very different in this respect.

The reader may also note that the generic-level *arbitrary* modalities and *explicit structure* modalities are not expanded at the atomic level. The reason is that, at this point, at least, no additional distinctions seem to be needed by developers. If further distinction becomes needed, they can simply be added by expanding those modalities, i.e., one, several, or all of them, at the atomic level. Similarly, the first three numbered modalities at generic level, i.e., the graphic, acoustic, and haptic *analogue linguistic elements*, remain unexpanded. This is another example of pragmatic fusion. The modalities themselves cover possible languages in all three media

which use analogue elements, like hieroglyphs or onomatopoetica, for linguistic representation. It would seem slightly contrived today to attempt to revive the analogue origins of the elements of many existing languages, which is why all discussion of languages in modality theory is done within the categories of non-analogue linguistic representation.

Sub-generic levels: there are potentially an unlimited number of sub-generic levels of which two are shown in the taxonomy in Table 1. The important point is that, in order to generate modalities at sub-generic levels, the definitions and distinctions in Figure 1 are no longer sufficient. This means that all generation beyond generic level must proceed by establishing and validating new distinctions.

Atomic level: relative to the generic level, the atomic level expands parts of the taxonomy tree based on new sets of distinctions that can be easily read off from the taxonomy tree. The distinctions are defined in [Bernsen 2002].

While the linguistic and analogue *generic-level* modalities are generally less familiar to our natural or prototypical conceptualisations [Rosch 1978] of information representation, their descendant modalities at atomic level are generally highly familiar. This is where we find, among many other modalities, GUI menu labels/keywords (5b), spoken discourse (6a), Braille text (7a), sign language (4a-c), and various static and dynamic analogue graphic representations, all thoroughly familiar. However, several of their corresponding acoustic and haptic sisters, though potentially very useful, are less familiar and may be viewed as products of the generative poser of the taxonomy.

Sub-atomic level: in the taxonomy in Table 1, only a few segments of the taxonomy tree have been expanded at sub-atomic level. The top right-hand corner shows expansion of s*tatic written text*, *labels/keywords*, and *notation* (5a-c) into their typed and hand-written varieties, respectively. This is a rather trivial expansion which is there primarily to show the principle of downwards tree expansion onto the sub-atomic level. We are all familiar with the difference it makes to both users and the system if they have to read hand-written text as opposed to typed text.

In the lower right-hand corner, the *dynamic graphic image* modalities are expanded into the non-speech, natural interactive communication modalities of visible *facial expression*, (non-sign-language) *gesture*, and *body action* (12a1-3). This distinction is argued in [Bernsen and Dybkjær, in press(b)]). If the animated interface agent also speaks, its multimodal information representation is increased even more by *acoustic speech* modalities (6a-c) and v*isual speech* modalities, i.e., mouth and lip movements during speech (8d-f).

It seems clear that the sub-atomic-level expansions of the modality taxonomy tree shown above are incomplete in various ways, i.e., multimodal interactive system developers need additional modality expansions based on the atomic level. The important point is that everyone is free to expand the tree whenever there is a need. The author would very much appreciate information about any expansions made and as well as how they have been validated.

Below sub-atomic 1evel: here are a couple of examples of highly desirable taxonomy tree expansions at sub-sub-atomic level. Although we still don't have a single complete, generally accepted or standard taxonomy of all different types of (non-sign language) *gesture* (12a2), there is considerable agreement in the literature about the existence of, at least, the following hand-arm gesture modalities, cf. [McNeill 1992]:

12a2a deictic gesture
12a2b iconic gesture
12a2c metaphoric gesture
12a2d emblems
12a2e beats (or batons)
12a2f other

We use the standard expedient of corpus annotators of adding an 'other' category for gestures which don't fit the five agreed-upon categories, thus explicitly marking the gesture modality scheme as being under development. Similarly, if there were a stable taxonomy for *facial expression of emotion*, it would constitute a set of sub-sub-atomic modalities expanding *facial expression* (12a3).

Some General Properties of the Taxonomy

In this and the following section, we list a number of properties of the taxonomy of input/output modalities and of the modalities themselves, which follow from the discussion above.

Unlimited downwards expansion. The taxonomy tree can be expanded downwards without limitation as needed, by analysing, defining, and validating new distinctions that can serve as basis for downwards expansion, as these become relevant in interactive systems design, development and evaluation. Conversely, this is also why some parts of the tree have not (yet) been expanded beyond the generic level, i.e., the arbitrary and explicit structure modalities, or the gesture modality.

Property inheritance. The taxonomy shows that modalities can be analysed and described at different levels of abstraction. We do this all the

time when working with modalities but may not be used to thinking about what we do in these terms. It follows that the taxonomy enables property inheritance. For instance, we can analyse the *linguistic* modality quite generally at super level, discovering the general properties of linguistic modalities we need for some purpose. Having done that, these properties get inherited by all *linguistic* modalities at lower levels. Thus, e.g., *spoken discourse* (atomic level) inherits all properties of the super-level *linguistic* modality as well as the more detailed properties of *acoustic linguistic* representation (generic level). Having analysed the parent properties, all we need to do to analyse the *spoken discourse* modality is to analyse, at atomic level, the new emergent properties of spoken discourse at this level. The origin of these emergent properties is clear from Table 1: it's the distinctions which enabled expansion of generic-level node 6 (*linguistic static/dynamic non-analogue acoustics*) into atomic-level node 6a (*spoken discourse*), i.e., between *discourse, text, labels/keywords* and *notation*.

Completeness at generic level. The taxonomy is claimed to be complete *at generic level* for the three media it addresses. No disproof of this claim has been found so far. However, the taxonomy is not complete at lower levels and might never be in any provable sense.

Some General Properties of Modalities

Modalities and levels of abstraction: it seems clear that a unimodal modality, i.e., a type of information representation in some physical medium, is always being thought of at some specific level of abstraction. Since the taxonomy makes this very clear, it might act as a safeguard against common errors of over-generalisation and under-specification in modality analysis.

Enumerability: modalities can only be finitely enumerated at generic level, and to do that, one has to go back to their derivation prior to the pragmatic fusions done in the taxonomy shown in Table 1. This is of little interest to do, of course, but, otherwise, it would always seem possible, in principle, to create and validate new distinctions and hence generate new modalities at sub-generic levels. In practice, though, there is a far more important question of enumeration, and this one will always have a rough-and-ready answer determined by the state of the art, i.e.: *how many different unimodal modalities do we need to take into account in interactive systems design, development, and evaluation for the time being?*

Validation: modalities are generated through distinctions, and these distinctions are more or less scientifically validated at a given time. The primary method of validation is to apply a set of generated categories to phenomena in data corpora and carefully analyse the extent to which the

categories are able to account for all observed phenomena both exhaustively and unambiguously.

Information Channels

The notion of an 'information channel' marks the most fine-grained level of modality theory and the level at which the theory, when suitably developed beyond its current state, links up with signal processing in potentially interesting ways.

Since a modality is a way of representing information in a physical medium, we can ask about the physical properties of that medium which *make it possible* to generate different modalities in it. These properties are called *information channels*. In the graphics medium, for instance, basic information channels include shape, size, position, spatial order, colour, texture, and time. From these basic properties, it is possible to construct higher-level information channels, such as a particular font type which is ultimately being used to represent the *typed text* modality.

Thus, information channels are the media-specific building blocks which define a modality in a particular medium. For instance, we could easily define a *static graphic black-and-white image* modality at atomic level (could be the new modality 9a1 in Table 1) and define its difference from modality *static graphic colour image* (9a2) by the absence of the information channel *colour*. In another example, the FACS, the Facial Action Coding System [http://face-and-emotion.com/dataface/facs/descrip - tion.jsp], starts from the fact that facial expression is being generated by some 50+ facial muscles used in isolation or in combination to form facial expressions of our mental and physical states, and specifies Action Units (AUs) for representing the muscular activity that produces momentary changes in facial appearance. The possible contraction patterns of the muscles are the information channels with which FACS operates.

Interaction Devices

Knowledge about modalities and multimodality is about how abstract information is, or can be, physically represented in different media and their information channels, and in different forms, called modalities. This knowledge has nothing to do with knowledge about *interaction devices*, and for good reason, because interaction devices come and go but modalities remain unchanged.

However, and this is the point to be made here, this simply means that designers and developers must go elsewhere to solve their problems about

which physical devices to use for enabling interaction using particular modalities. And these problems remain interrelated with the issue of modality choice. It is often counter-productive to decide to use a particular modality for interaction, such as different 3D gesture modalities for input, if the enabling camera and image processing technologies currently cannot deliver reliable recognition of those gestures.

Practical Uses of the Theory

At this point, we have addressed three of the six desiderata listed at the start of the present *Modalities* section, i.e.: (i) identify all *unimodal* or elementary modalities which could be used to build multimodal interfaces and enable multimodal interaction; (ii) group modalities in one or several sensible ways, hierarchically or otherwise; and (iv) a practical toolbox of all possible unimodal input/output modalities to choose from. This, arguably, is quite useful in practice because it enables us to know which unimodal modalities there are in the three media scoped by the theory, how they are hierarchically interrelated, and how to decompose any multimodal representation into its constituent unimodal modalities. Issue (iii) on basic information on each modality goes beyond the scope of the present chapter but will soon be available at www.nislab.dk in the form of clickable taxonomy trees providing access to basic information on each modality.

Equally clearly, however, a practical toolbox of modalities to choose from would be far more useful if it came with information about *which modality to choose for which purpose* in interactive systems development. This takes us to desideratum (v): guidelines or something similar for which modality to use for a given development purpose, or what might be called the issue of *modality functionality*.

Modality Functionality

The issue about the *functionality* of a particular modality lies at the heart of the practical question about which modality or which set of modalities to use for a given development purpose. As it turns out, modality functionality, and hence modality choice, raises a whole family of issues, i.e.:

- is modality M(a) useful or not useful for development purpose P?
- is modality M(a) more or less useful for purpose P than an alternative modality M(b)?
- is modality M(a) *in combination with* modalities M(c, c+1, ...c+n) the best multimodal choice given purpose P?

Moreover, it must be taken into account if modality M(a) is considered to be used for *input* or *output* because this may strongly affect the answers to the above questions.

Far worse, even, as regards the complexity of the modality choice problem which we can now see emerging, is the fact that, in interactive systems design, development, and evaluation, development purpose P is *inherently highly complex*. In [Bernsen and Dybkjær, in press (a)], we argue that P unfolds into *sets* of component parameters each of which has a *set* of possible values, of the following generic parameters:

- application type
- user
- user group (user population profile)
- user task or other activity
- application domain
- use environment
- interaction type
- interaction devices

Now, if we multiply such multiple sets of development purpose-specific values by the modality function questions above *and* by the sheer number of unimodal modalities in our toolbox, we get a quasi-intractable theoretical problem. Furthermore, intractability is not just due to the numbers involved but also to the fact that many of those sets of component parameters and their sets of values are likely to remain ill-defined forever.

Still, we need *practical answers* to the question about which modalities to use for a given development purpose. When, after many failed attempts to make modality theory directly applicable, we finally discovered the intractability problem just described, we identified (functional) *modality properties* as the primary practical contribution which modality theory can make.

Modality Properties

Modality properties are functional properties of modalities which characterise modalities in terms that are directly relevant to the *choice* of input/output modalities in interactive systems design, development, and evaluation.

To study the potential usefulness of modality properties, we made a study [Bernsen 1997] of *all speech functionality claims* made in the 21 paper contributions on speech systems and multimodal systems involving

spoken interaction in [Baber and Noyes 1993]. In this and a follow-up study which looked at a cross-section of the literature on speech and multimodality 1993-1998 [Bernsen and Dybkjær 1999a, 1999b], we analysed a total of 273 claims made by researchers and developers on what particular modalities were good or bad for. An example of such a claim could be: "Spoken commands are usable in fighter cockpits because the pilot has hands and eyes occupied and speech can be used even in heads-up, hands-occupied situations". It turned out that more than 95% of those claims could be evaluated and either justified, supported, found to have problems, or rejected by reference to a relatively small number of 25 modality properties. These are exemplified in Table 2.

Regarding the modality properties listed in Table 2, it is important to bear two points in mind in what follows: (i) the listed modality properties are selected examples from the longer list of properties made to evaluate those 273 claims made in the literature; and, more importantly, (ii) that longer list was made *solely* in order to evaluate those 273 claims. So the list does not have any form of theoretical or practical closure but simply includes the modality properties which happened to be relevant for evaluating a particular set of claims about speech functionality. In other words, modality theory has more to say about speech functionality, and far more to say about modality functionality in general, than Table 2.

Table 2. Modality properties.

No.	Modality	Modality Property
MP1	**Linguistic input/output**	Linguistic input/output modalities have interpretational scope. They are therefore unsuited for specifying detailed information on spatial manipulation.
MP3	**Arbitrary input/output**	Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned.
MP4	**Acoustic input/output**	Acoustic input/output modalities are omnidirectional.
MP5	**Acoustic input/output**	Acoustic input/output modalities do not require limb (including haptic) or visual activity.
MP6	**Acoustic output**	Acoustic output modalities can be used to achieve saliency in low-acoustic environments.
MP7	**Static graphics**	Static graphic modalities allow the simultaneous representation of large amounts of information for free visual inspection.
MP8	**Dynamic**	Dynamic output modalities, being temporal (serial and transient), do not offer the cognitive advantages

	output	(with respect to attention and memory) of freedom of perceptual inspection.
MP11	**Speech input/output**	Speech input/output modalities in native or known languages have very high saliency.
MP15	**Discourse output**	Discourse output modalities have strong rhetorical potential.
MP16	**Discourse input/output**	Discourse input/output modalities are situation-dependent.
MP17	**Spontaneous spoken labels/ keywords and discourse input/output**	Spontaneous spoken labels/keywords and discourse input/output modalities are natural for humans in the sense that they are learnt from early on (by most people). Note that spontaneous keywords must be distinguished from designer-designed keywords which are not necessarily natural to the actual users.
MP18	**Notation input /output**	Notation input/output modalities impose a learning overhead which increases with the number of items to be learned.

Now back to those studies. Moreover, even though (i) speech in multimodal combinations was the subject of only few claims in the first study which evaluated 120 claims, and (ii) there were many more claims about speech in multimodal combinations in the batch of 157 claims analysed in the second study, the number of modality properties used in claims evaluation only went up from 18 in the first study to 25 in the second study. The reason why these numbers are potentially significant is that *modality properties may remain tractable in number while still offering significant support* to interactive systems designers, developers and evaluators.

This likely tractability may be contrasted with several other familiar approaches. On is the *strict experimentalist approach*, in which a careful, expensive, and time-consuming study may conclude, e.g., that X % of Y children, aged between 6 and 8, the group being normal in its distribution of English speech and manual dexterity skills, in a laboratory setting, using a map application, were found to stumble considerably (and here we get tables, numbers and percentages) in their speech articulation when they used speech to indicate map locations, whereas they found it significantly easier to use pen-pointing to input the same information. - Incidentally, this short description includes values of all the generic parameters described in the previous section. - But what does this result tell practitioners who are specifying a different application, with different users and user profiles, for a different environment, for a different task, and using partially different input devices? The answer is that *we don't know*. Modality

property MP1 in Table 2, on the other hand, directly says that speech is unsuited for specifying detailed spatial information. Furthermore, MP1 does so without even mentioning speech. Instead, if relies on property inheritance from linguistic representation of information in general.

Another contrasting approach is *guidelines of the if-then type*. An if-then guideline might say, e.g., that if the environment is an office environment, if the person is alone in the office, and if the person is not a fast typist, then speech dictation might be considered as an alternative to typing. This guideline is nice and concrete, it might be true, and it might be helpful for someone who develops office applications. The problem is tractability, because how many guidelines would we need of this kind? Clearly, we would need a catastrophic number of guidelines.

On the other hand, modality properties come at a price to be paid in natural intelligence. It is that they focus on the modality itself and its properties, rather than mention values of the parameters listed in the previous section. That's why modality properties are comparatively economical in number: they leave it to the natural intelligence of developers to apply them to the parameter values which characterise the development project at hand. Arguably, however, there seems to be a rather trivial but important point here, i.e., if we want to know what a modality is or is not suited for, we need to understand, first of all, the *information representation properties* of that modality. So it would seem to be a likely prediction that the more useful a guideline is for guiding modality choice in practice, the more it resembles a statement of a modality property.

Multimodal Information Representation

Modality theory is fundamentally about the *unimodal* modalities that, as building blocks, go into the construction of multimodal information representation. Before we look into the process of construction, it is useful to briefly discuss the advantages offered by the growing number of input/output modalities that are becoming available.

Advantages of Multimodality

Since *no two modalities are equivalent*, all modalities differ amongst each other, as we have seen, in terms of their individual combination of expressive strengths and weaknesses *and* their relationship with the human perceptual, cognitive, emotional, etc. system. The central implications are that the more modalities we can choose from, (i) the *wider the range of infor-*

mation it becomes possible to express as input or output; and (ii) the higher our chances become of identifying a modality combination which has a suitable, if not optimal, relationship with the human system for at given application purpose.

If we *combine* two or several modalities, we ideally get the sum of their expressive strengths and are able to overcome the expressive weaknesses of each of them taken individually. However, it still remains necessary to make sure as well that the combination is possible for, and acceptable to, the human users.

Also, the more modalities we have at our disposal as developers, the more we can develop applications for *all users*, including people with perceptual, cognitive and other disabilities, people with different degrees of computer literacy, the 1 billion people who are illiterate, as well as users with sometimes widely different preferences for which modalities to use. This is often done by *replacing* information expressed in one modality by, practically speaking, the same information expressed in a different modality, like when the blind get their daily newspaper contents read aloud through text-to-speech synthesis.

Given these limitless opportunities, it is no wonder that multimodality is greeted with excitement by all.

Constructing Multimodality from Unimodal Modalities

Since we have a theory of unimodal modalities, and, it would appear, *only because* we have something like that, it makes sense to view multimodal representation as something which can be *constructed* from unimodal representations, analogous to many other constructive approaches in science – from elements to chemistry, from words to sentences, from software techniques to integrated systems.

This view is sometimes countered by some who do not use modality theory, by an argument which goes something like this: the whole point about multimodality is to create something *entirely new*. When modalities are combined, we get *new emergent properties* of representations which cannot be accounted for by the individual modalities on their own. Now unless one is inclined towards mysticism, this argument begs the question about whether and to which extent multimodal combinations can be analysed, or even predicted, as resulting from an ultimately transparent process of combining the properties of unimodal modalities and taking the relationship with the human system into account. As repeatedly stressed above, this process is provably a very complex one in general, but so is the field of synthetic chemistry.

In the remainder of this chapter, we introduce some distinctions in order to approach the issue of multimodal construction and remove some of the mystery which still seems to surround it.

Linear Modality Addition and Replacement

Let us define a concept of modalities which can be *combined linearly* so that the combination *inherits* the expressive strengths of each modality and does not cause any significant negative *side-effects* for the human system. It is very much an open research question which modalities can be combined in this fashion and under which conditions. However, to the extent that modalities *can* be combined linearly, it is straightforward to use the modality properties of the constituent unimodal modalities to describe the properties of the resulting multimodal representation. The modalities simply add up their expressive strengths, and that's that. Let us look at some examples.

Modality Complementarity

In a first, non-interactive example, we might take a *static graphic* piece of *text* describing, say, a lawnmower, and add a *static graphic image* of the lawnmower to it, letting the text say what the lawnmower can do and how to use and maintain it, and the picture show how it looks. For good measure, we might throw in a *static graphic compositional diagram* of the lawnmower, showing its components and how they fit together, and cross-reference the diagram with the text.

In another, interactive, example, we might put up a large screen showing a *static graphic* Sudoku gameboard and have users play the game using *spoken numbers and other spoken input keyword commands* in combination with 3D camera-captured and image-processed *pointing gesture input*. A number is inserted into, or deleted from, the gameboard by pointing to the relevant gameboard square and uttering a command, such as "Number seven" or "Delete this". A recent usability test of the system showed rather unambiguously that this multimodal input/output combination works well both in terms of input/output modality expressiveness and in terms of fitting the human system [Bernsen and Dybkjær 2007]. Modality theory would predict that the spoken input keywords (other than the numbers 1 through 9 which are familiar to all Sudoku players), being designer-designed, might cause memory problems for the users, cf. Table 2, MP 17. However, since there were only a couple of them in the application, the problems caused would be predicted to be minimal, which, in fact, they

were. In addition, it would be easy to put up the "legal" keywords as an external memory on the screen next to the Sudoku gameboard to solve the problem entirely.

These two examples are among the classical examples of *good multimodal compounds*: they work well because they use the *complementary* expressive strengths of different modalities to represent information which could not easily be represented in either modality on its own. In the first example, the complementarity i *a-temporal* because of the freedom of visual inspection afforded by the static graphics. In the second example, the complementarity is *temporal* because the speech is being used dynamically and therefore the pointing gestures have to occur at the appropriate time intervals lest the meaning of the message would be a different one – if, indeed, any contextually meaningful message would result at all. In the first example, since all component modalities are static, the multimodal representation causes no working memory problems nor any perceptual or cognitive conflicts due to pressure to process too much information at the same time – indeed, the representation as a whole acts as an external memory. The second example makes use of a modality combination which is as natural as natural language and is, in fact, learned in the process of learning a natural language. These examples do seem to well illustrate the notion of linear addition of modalities, i.e., of *gaining novel expressiveness without significant side-effects*, cognitive or otherwise, through modality combination.

Good multimodal compounds need not be as simple and as classical as the examples just discussed. We recently tested the usability of a treasure hunting game system prototype in which a blind user and a deaf-mute user collaborate in finding some drawings essential to the survival of an ancient Greek town [Moustakas et al. 2006]. For the blind user alone, the input/output modalities are: *spoken keywords output* to help the blind navigate the 3D townscape and its surroundings to find what s/he needs; *non-speech sound* musical instrument output acting as arbitrary codes for the colours of objects important in the game; *haptic 3D force-feedback output* providing the blind user with data for navigating the environment, locating important objects, and building a mental map of the environment; *haptic 3D* navigation robot arm *input* through which the blind orientates in the environment; and *haptic click notation input* through which the blind acts upon objects important to the game. The usability test of this system showed that this multimodal compound worked excellently except for the problem of remembering the arbitrary designer-designed associations between colours and musical instruments, something which, again, is predicted by modality theory [Bernsen and Dybkjær 2007]. For the purpose of the evaluated game, the colour problem would disappear if the objects

simply described their colour through speech rather than using arbitrary non-speech sounds.

Modality Redundancy

Linear addition also often works well in another abstract kind of modality combination which is often termed *redundancy*. Sometimes it is useful to represent more or less the same information in two different modalities, for instance because the information is particularly important or because of particular values of the *user* or *environment* parameters (cf. above). So, for instance, we add an acoustic alarm (Table 1, 14) to a visual alarm (Table 1, 13/16) for increased security in a process plant; or we add visual speech (Table 1, 8d-f) to speech output because the user is hard-of-hearing or because the environment is or can be noisy. The visual speech is actually *not* information-equivalent to the speech but comes close enough for the redundant multimodal representation to provide significant help to users when the speech is difficult to hear. Again, no significant side-effects would be predicted in these cases.

Modality Replacement

In linear *modality replacement*, one modality (or a combination of modalities) is replaced by another for the purpose of achieving practically the same representation of information as before. Again, many examples can be mentioned where this works sufficiently well, such as replacing *spoken discourse* for the hearing by *sign language discourse* for the deaf and hard-of-hearing; or replacing *static graphic written text* for the seeing by *static haptic Braille text* for the blind or hard-of-seeing. Yet again, no significant side-effects would be predicted.

It should be noted that several have distinguished other abstract types of modality combination in addition to complementarity, redundancy, and replacement, but space does not allow discussion of these, see, e.g., [Martin 1995] and [Nigay and Coutaz 1993].

The conclusion at this point is that in *a large fraction of cases* in which several modalities are combined into multimodal representations, the resulting multimodal representation (i) is largely a straightforward *addition* of modalities, or a straightforward replacement by functionally equivalent modalities, with no significant side-effects upon the human system; and (ii) that the knowledge of unimodal modalities can be used to good effect in predicting the functionality of the constructed multimodal compound. In other words, combining modalities can be a straightforward and predict-

able process of construction rather than the creation of a magical concoction with mysterious properties and totally unpredictable effects.

Non-linear Effects, Users, Design Detail, Purpose

However, we now need to return to the huge theoretical complexity of the general modality choice problem, this time in a multimodal context.

In fact, modality choice complexity is such that we do not recommend that *any* novel multimodal combination be launched without thorough usability testing. Modality theory can be very helpful in the analysis and specification phase, suggesting modality combinations to be used or left out, predicting their expressiveness and potential problems in relationship to the human system. But modality theory, however much further developed, cannot guarantee unimodal or multimodal application success. There are many, partly overlapping, reasons for exercising caution.

The first reason to be mentioned is *non-linear effects*. For instance, nothing might appear more straightforward than to add a voice interface to an email system so that the user can access emails without having a static graphic text interface at hand. This kind of modality replacement is sometimes called *interface migration*. However, the user soon discovers things like that the overview of the emails received is gone and not replaced by any different mechanism, and that the date-time information read aloud by the system is (i) unnatural and (ii) takes an exceedingly long time to listen to. In other words, while the modality replacement no doubt preserves *information equivalence*, something has gone wrong with the relationship with the human system. In this case, successful modality replacement is not straightforward at all because the entire structure of the email information representation has to be revised to arrive at a satisfactory solution. Modality theory can tell the developer that, despite their information equivalence-in-practice in the present case, the (non-situated) *text* modality is fundamentally different from the (situated) *discourse* modality, the former having a tendency to being far more explicit and elaborate, as illustrated by the lengthy absolute date-time information provided in the email list; and (ii) that speech, being largely a *dynamic* modality, does not enable anything like the information overview provided by *static* graphics. So the theory can advise that developers should be on the alert for non-linear effects and test for them using early mock-ups, but the actual effects just described can hardly be predicted due to their detailed nature.

A second reason is the *users*. From a theoretical point of view, *both* in terms of practical information equivalence *and* in terms of the absence of side-effects on the human system, it may be a perfect example of modality

replacement to replace *static graphic text* by *static haptic* Braille *text*. But then it turns out that, for instance, only 5% of the blind Danish users know Braille whereas some 80+% of Austrian blind users know Braille. Given this marked difference in generic parameter *user population profile*: component parameter *user background skills*, the modality replacement above might be a good idea in Austria whereas, in Denmark, one would definitely recommend replacing static graphic text by *text-to-speech* instead. Modality theory, of course, has no notion of Braille skill differences between Austrian and Danish populations of blind users. Or, to mention just one more example, the theory has little to say about the notoriously hard-to-predict *user preferences* which might render an otherwise theoretically well-justified modality addition or replacement useless.

A third reason is *design detail*. It may be true in the abstract that, for a large class of applications, the addition of a dynamic graphic animated human representation adds a sense of social interaction to interactive systems. But this advantage might easily be annulled by an animation who is perceived as being unpleasant of character, daft, weird, overly verbose, occupying far too much screen real-estate for what it contributes to the interaction, or equipped with a funny voice

A fourth reason is the *purpose* of adding or replacing modalities. If the purpose is a more or less crucial one, such as providing blind users with text-to-speech or Braille access to the linguistic modality, this is likely to overshadow any non-linear effects, specific user preferences, or design oddities. But if the purpose is less essential or inessential – such as adding entertaining animations to web pages, small-talk to spoken dialogue applications, or arbitrary musical instrument coding of colours which could just as well be described through another output modality which is actually used in the application, such as spoken keywords – then users are likely to be far less tolerant to the many other factors which are at play in creating user satisfaction.

References

Baber, C., and Noyes, J. (Eds.). *Interactive Speech Technology*. London: Taylor & Francis, 1993.

Bernsen, N. O.: Foundations of Multimodal Representations. A Taxonomy of Representational Modalities. *Interacting with Computers* 6.4, 1994, 347-71.

Bernsen, N. O. Why are Analogue Graphics and Natural Language Both Needed in HCI? In Paterno, F. (Ed.), *Design, Specification and Verification of Interactive Systems. Proceedings of the Eurographics Workshop,* Carrara, Italy, 1994, 165-179. *Focus on Computer Graphics.* Springer Verlag, 1995: 235-51.

Bernsen, N. O. Towards a Tool for Predicting Speech Functionality. *Speech Communication* 23, 1997, 181-210.

Bernsen, N. O.: Multimodality in Language and Speech Systems - From Theory to Design Support Tool. In Granström, B., House, D., and Karlsson, I. (Eds.): *Multimodality in Language and Speech Systems.* Dordrecht: Kluwer Academic Publishers 2002, 93-148.

Bernsen, N. O., and Dybkjær, L.: Working Paper on Speech Functionality. *Esprit Long-Term Research Project DISC Year 2 Deliverable* D2.10. University of Southern Denmark, 1999a. See www.disc2.dk

Bernsen, N. O., and Dybkjær, L.: A Theory of Speech in Multimodal Systems. In Dalsgaard, P., Lee, C.-H., Heisterkamp, P., and Cole, R. (Eds.). *Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, Irsee, Germany. Bonn: European Speech Communication Association, 1999b 105-108.

Bernsen, N. O., and Dybkjær, L.: Report on Iterative Testing of Multimodal Usability and Evaluation Guide. *SIMILAR Deliverable D98*, October 2007.

Bernsen, N. O., and Dybkjær, L. (in press (a)): *Multimodal Usability*.

Bernsen, N. O., and Dybkjær, L. (in press (b)): Annotation Schemes for Verbal and Non-verbal Communication: Some General Issues.

Bolt, R. A: Put-that-there: Voice and Gesture at the Graphics Interface. *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, Seattle, 1980, 262-270.

Hovy, E. and Arens, Y.: When is a Picture Worth a Thousand Words ? Allocation of Modalities in Multimedia Communication. Paper presented at the *AAAI Symposium on Human-Computer Interfaces*, Stanford, 1990.

Jurafsky, D. and Martin, J. H.: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

Lakoff, G.: *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: University of Chicago Press 1987.

Martin, J.-C.: *Cooperations between Modalities and Binding Through Synchrony in Multimodal Interfaces.* PhD Thesis (in French). ENST, Orsay, France 1995.

McNeill, D: *Hand and Mind.* University of Chicago Press, 1992.

Moustakas, K., Nikolakis, G., Tzovaras, D., Deville, B., Marras, I. and Pavlek, J.: Multimodal Tools and Interfaces for the Intercommunication between Visually Impaired and "Deaf and Mute" People. *eNTERFACE'06*, July 17th – August 11th, Dubrovnik, Croatia, *Final Project Report*, 2006.

Nigay, L. and Coutaz, J.: A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. *International Conference on Human-Computer Interaction.* ACM Press, 1993, 172-178.

Oviatt, S. and Cohen, P.: Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM*, 43/3, 2000, 45-53.

Rosch, E. Principles of Categorization. In Rosch, E. and Lloyd, B. B. (Eds.). *Cognition and Categorization.* Hillsdale, NJ: Erlbaum, 1978.