INTEGRTATING KNOWLEDGE ABOUT MODALITIES TO A MULTIMEDIA KNOWLEDGE REPRESENTATION FRAMEWORK

Y.Bachvarova † , N.Elouazizi*

†Human Media Interaction Group, University of Twente, The Netherlands
y.s.bachvarova@ewi.utwente.nl
*LUCL, Leiden University, The Netherlands
n.elouazizi@let.leidenuniv.nl

Keywords: Modality, Media, Multimodal, Knowledge representation

Abstract

In order to be satisfactorily adequate in generating relevant multimodal information, we argue that any multimedia and multimodal ontology has to incorporate three basic criteria. These are: (i) a conceptually and semantically clear distinction between the operational concept of *Modality* and *Media* (medium), (ii) describe a set of recursive formal rules that can allocate and vehicle the appropriate modality information through the most relevant media, taking into consideration human cognitive constraints of perceiving and interpreting relevant information and (iii) develop formal rules to ensure that the output knowledge about the different modalities that constitute a final multimodal presentation can be recombined, reinterpreted and regenerated. The relevant interaction of these criteria to ensure the generation of optimal and relevant multimodal information in the multimedia and multimodal systems requires the existence of a modality ontology which can formalise this interaction. This paper, which is mainly concerned with the presentation of a modality ontology is a step in that direction.

1 Introduction

The generation of the multimedia and the multimodal output has been addressed by different research communities within the fields of artificial intelligence, knowledge representation and natural language processing in general and multimedia in particular. Falling within the domain of multimedia, the efforts of research have been mainly concentrated on achieving a consensus to establish standards to describe the content of media items. The knowledge about modalities as conceptualized through these knowledge representation and metadata models though highly expressive yet it fails to allow automatic relevant selection of the most optimal modality (combinations).

We assume that the choice of an optimal combination of modalities is a key factor to ensure that a final generated multimodal presentation will properly convey the desired information. The *allocatio*n of the most expressive modalities and their subsequent most appropriate *combinations* is a knowledge intensive process which requires explicit representation of a wide range of modality aspects pertinent to the computational processes of allocation and combination. In this respect, we identify two levels of knowledge representation for the automatic generation of multimodal systems and three criteria in designing the formal ontology which will support such multimodal automatic generation process. The two types of knowledge which require proper conceptual modelling and a unified integrated formal representation are the FORM (intrinsic features/semantics) of each modality and its FUNCTION (extrinsic features/semantics). These two types of knowledge respectively correlate with levels of representation, viz. the PROFILE and CONTENT level of a modality. In the PROFILE level, the modality is described in terms of its intrinsic features; that is, the ones that describe its capacities for presenting information that can be perceptually and cognitively processed. The CONTENT level describes the information that the particular modality represents. We argue that the formal ontology which will support these two levels of identifying the modalities and the knowledge they encode have to meet three criteria. These are: (i) a conceptually and semantically clear distinction between the operational concept of *Modality* and *Media* (medium), (ii) describe a set of recursive formal rules that can allocate and vehicle the

appropriate modality information through the most relevant media, taking into consideration human cognitive constraints of perceiving and interpreting relevant information and (iii) develop formal rules to insure that the output knowledge about the different modalities that constitute a final multimedia presentation can be recombined, reinterpreted and regenerated.

This paper is organised into five sections. In section two we briefly define the notion of modality and the automatic processes of its selection in multimodal systems. In section three, we provide an eye bird view of MPEG-7 [6] metadata model and the Modality Theory [1,2] in order to exemplify the two main aspects of knowledge about modalities that has been modelled, namely content and intrinsic modality features. Having laid the ground for the introduction of our modality ontology, section four describes the different levels of the modality ontology we postulate. Section five provides a summary of the issues discussed and points to the future directions of research.

2. Defining modality and the process of its selection

First of all, let us make our stand about how we define the notion *modality* clear. In our attempt at controlling the semantic properties associated with "modality", we draw on some fundamental insights of two existing approaches. These are the approach of Niels Ole Bernsen [1] and the one of Maybury [7]. We assume together with Niels Ole Bernsen [1] that it is necessary to make a distinction between the term *modality* and the term *medium*. The term *modality* is defined as the 'mode or way of exchanging information between humans or between humans and machines in some medium. The term *medium* is defined as the physical realization of some presentation of information at the interface between human and system.

Extending on this basic assumption, we associate the term *modality* with the human channels of perception as defined in the domain of cognitive psychology (Cf. Fodor [5]). That is, we associate the term *modality* with each of the human channels of perception that are deployed in storing, decoding, delivering and processing information. This includes the human visual system, the human linguistic system and the human auditory system, etc.. However, though we define the term modality, mainly on the basis of human cognitive system, we do not exclude that it entails the interaction with the machine (artificial multimodal systems in our case) and as such it acquires characteristics that are usually associated only with artificial systems. Thus, our definition of modality draws on the form and the function of the cognitive processes associated with each channel of perception and which can ensure a relevant and optimal interaction between the human and the computer.

Now that we defined the concept of modality that we adopt, the next step is to provide a functional view about how modality operates in the context of an artificial multimodal system. To do that, we need to make our stand explicit about which kind of architecture we are adopting for the notion of modality we postulated above. Consider the graphical representation in Figure 1 below.

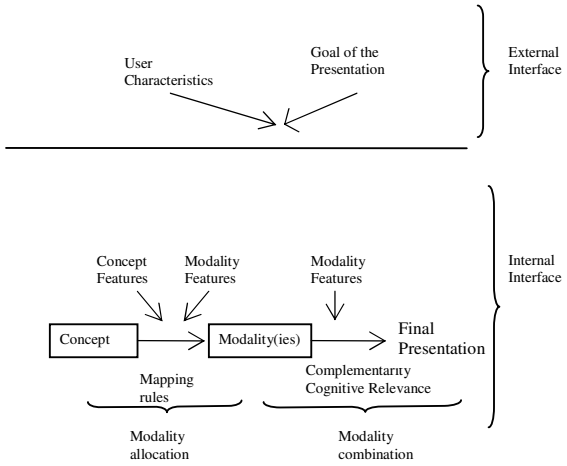

FIGURE 1: Architecture of modality selection.

Figure 1 above illustrates the kind of computational architecture which constitutes the operational domain for the computational processes and interfaces associated with modality. We identify in this architecture two interface levels and two computational derivational phases. These interface levels and computational phases are the bedrock for the generation of any multimodal output. Put more explicitly, in the process of selecting the most optimal modality we recognize at least two levels of interfaces – internal and external one [4]. The internal interface is the interface that exists between the different modalities while the external interface is the interface between the modalities and the context of the multimodal presentation (user profile, user task, goal of the presentation, discourse state). The internal and external interfaces involve processes that are parallel and interrelated rather than strictly hierarchical and sequential.

There are two main computational derivational phases for the process of selecting the optimal combination of modalities. These are the *Modality Allocation* and *Modality Combination*. The allocation phase is the one responsible for assigning the most expressive modalities that can best represent a particular concept. Modality combination is the computational module that outputs the most optimal combination from the set of modalities selected as a result of the allocation phase. Computationally speaking, modality allocation is a mapping between features of the information that has to be conveyed (in our model the concepts that has to be represented) and the intrinsic features of modalities. This implies the existence of structured representation of the knowledge describing the possibilities of each modality to represent information

content. Modality combination, on the other hand, requires conceptualization of the knowledge about the way each modality is perceived and processed by the human cognitive system.

To allow for a coherent contextualization of our proposal with the specifics of the notion of modality we adopted and the kind of computational architecture we assume for its execution, it is necessary to show how our proposal squares with the ones of our predecessors. This is what we will address in the subsequent section.

3 A short overview of the knowledge representation frameworks[1]

In this brief overview, we focus on two main knowledge representation frameworks described in [6] and [1] respectively. For each of these frameworks we describe the type of knowledge it captures and how that knowledge is being modelled. Based on that description we derive the main properties that are inherent and characterise the two main aspects of modality, namely its capacity to represent information in a particular way and the way it is perceived and processed by human cognitive system. The two modality-related-knowledge representation frameworks that we will briefly discuss here are the MPEG-7 [6] and the Modality theory [1].

Let us start with MPEG-7. MPEG-7 is a standard for describing different aspects of multimedia information at different levels of abstraction. It can describe visual features (e.g., colour), audio features (e.g., timbre), structure (e.g., moving regions and video segments), semantics (e.g., objects and events), management (e.g., creator and format), collection organization (e.g., collections and models), summaries (e.g., hierarchies of key frames) and, even, user preferences (e.g., for search) of multimedia. The key components of MPEG-7's semantic descriptions are *semantic entities* such as *objects* and *events*, *attributes* of these entities such as labels and properties, and, finally, *relations* of these entities such as an object being the patient of an event.
The MPEG 7 is an interesting metamodel as its eclectic approach enables it to describe the semantics of any modality on the basis of the golden triangle, borrowed formal semantics (viz. entities, attributes and relations). However, note that the fact of adopting the fundamental formal relation of entities and properties as provided by the framework of natural formal semantics is itself a problem. The triangle of entities, properties and relations can formalize any semantic relation that is entertained within the concepts of natural language; however, it can not set the constraints for generating the optimal outputs in an artificial domain (such as multimodal systems). The grain of the difference is that the operationalism of entities-properties-relations in natural cognition is optimal and yet poorly understood in terms of the cognitive constraints which regulate such optimality. Therefore the mapping of the semantics of the formal relation described above into an artificial environment as can be easily overloaded and hindered by its own complexity to generate optimal outputs in a predictable constrained way. Hence follows the difficulty of its use and replication across various types of multimodal systems.

Modality Theory was developed by Niels Ole Bernsen [1,2] as an attempt to answer the general problem of mapping task domain information into interactive multimodal interfaces. One of the contributions of Modality Theory is the generative taxonomy of output modalities it proposes. The taxonomy distinguishes different modalities based on a set of intrinsic features that determine how powerful is the specific modality in representing the different types of information. While this approach might prove useful for modality allocations, it is not clear how it can be replicated for the process of modality combinations. Two processes as we claim above are necessary for any type of knowledge representations within multimodal systems.

All in all, despite their claims of modelling the heterogeneous aspects of information in multimodal systems, most, if not all, the existing modality-knowledge representation frameworks share one single attribute. They lack a consensus view about how to formally trace the semantic properties that distinguish between the operational concepts of "*Medium*", "*Modality*", "Multimedia" and "*Multimodal*". Thus, given the formal, conceptual and operational primordial role of these concepts as the core atomic units in building and generating multimedia/multimodal automatic presentation, it is a methodological and implementational imperative to formally handle such slippery concepts with care. The challenge therefore, is in whether it is possible to conceptualize and design an alternative ontology for the automatic assignment of modalities and which encompasses the many advantages which exist independently in the models described above and yet be simple enough in its conceptual structure to allow a maximum of across-the-board implementation within multimodal systems. In the following section, we propose a modality ontology which is sensitive towards the set of methodological and implementational concerns discussed above.

4 Adding and integrating knowledge about modalities

In our model we identify two main levels through which the identity of modality is described – PROFILE and CONTENT level. In the PROFILE level the modality needs to be described in terms of its intrinsic features that are the ones that describe its capacities for presenting information that can be perceptually and cognitively processed. The CONTENT level describes the information that a particular modality can represent by combination or interaction with other modalities. Thus, any multimedia knowledge representation framework, we assume, has to account for these two levels of information.

[1] Note that we are by no means making justice to all the available models and literature on the topic and this goes beyond the scope of this short paper.

For a concrete illustration, consider figure 2. Note that, for expository purposes, we are describing modality ontology by the following three figures (that is figure 2, 3 and 4)[2]. Let us start with figure 2.

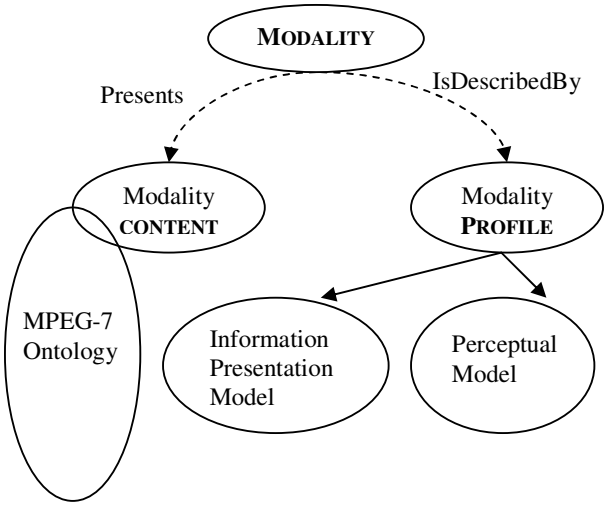

FIGURE 2: Upper level of the Modality Ontology

The graphic representation in figure 2 describes the upper level of the modality ontology we are proposing.[3] The primitive category at the core of the presented model is *modality*. The categories of Modality CONTENT and Modality PROFILE come at the next level of the ontology. The function of the Modality CONTENT is to model the information content each modality represents. Likewise, the Modality PROFILE describes the modality in terms of its cognitive representation as well as its semantic relational capacity. The intrinsic modality features described in the PROFILE fall into two main categories: features describing the suitability of each modality to represent information (INFORMATION PRESENTATION MODEL) and perceptual features (PERCEPTUAL MODEL) which describe the way each modality is perceived and processed by the human cognitive perceptual channels.

Now zooming on the INFORMATION PRESENTATION MODEL (Cf. figure 3) the latter encompasses the categories that describe the most general and robust distinctions among the capabilities of different modalities to represent information. The informational entities that the INFORMATION PRESENTATION MODEL encompasses are both linguistic and analogue (visual). The entities that belong to the Linguistic category, such as speech and text, have two most notable characteristics – they can abstract and focus [1]. That is, their referential linguistic capacity is detached from the here and now as it can refer to things that are abstract and across time and space. Hence, the linguistic representations focus at some level of abstraction on the subject matter to be communicated without the need to provide its specifics. This stands in contrast to the entities which belong to the analogue category and which depend on how the subject matter they represent looks or sounds (for example when video is used)[1]. Further note that the modalities belonging to the linguistic and analogue categories can function independently in terms of establishing their own domain specific semantic and referential relations and yet they can equally be complementary to each other.

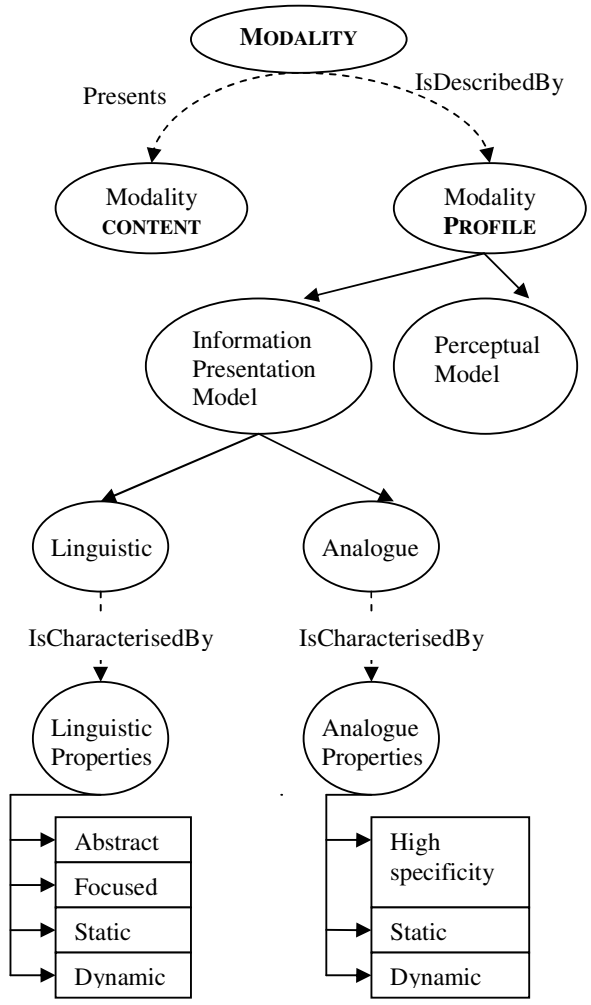

FIGURE 3: The Information Presentation Model of the Modality Ontology

Another interesting aspect of the linguistic and the analogue categories is that they allow us a flexible yet constraint governed degree of distinguishing between the static and the dynamic aspects of modalities. For example, while static representations may be decoded by the user in any order

[2] The dashed lines in the figures denote property relationship. The solid lines indicate sub-class relationship.

[3] Note that the modality ontology we propose here can be easily extended with MPEG-7 ontology in order to represent the content of the modalities which MPEG-7 has the capacity to describe.

desired and as long as desired, dynamic representations are transient and do not afford freedom of perceptual inspection [3].

Now let us consider the aspect of the perceptual features that the modality ontology represents. Consider figure 4 for an illustration.

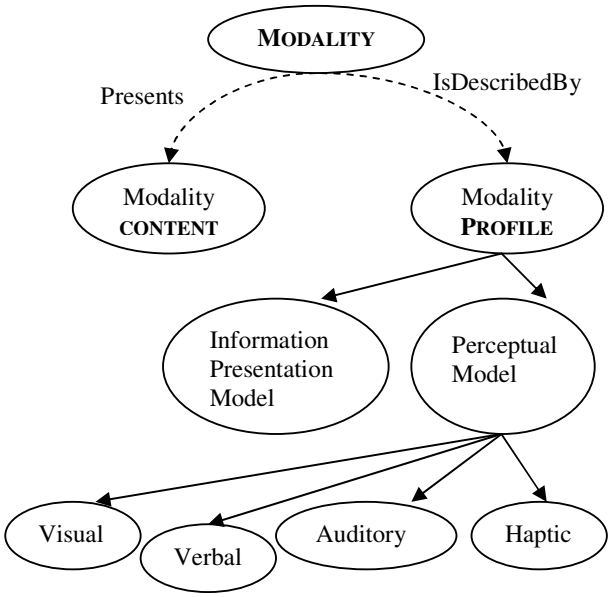

FIGURE 4: The Perceptual Model of the Modality Ontology

The perceptual Model in this ontology is intended to model the knowledge about modalities and describes the way they are perceived and processed by the human cognitive system. In modeling perception we adopt the view that language and non-language information is processed differently and through different perceptual channels. In this respect, we can talk about, at least, four perceptual channels, viz. visual, auditory, verbal and haptic modes of perception (that is of perceiving, encoding, decoding, transferring, retrieving and storing the informational content).

5 Concluding remarks and future directions

In this paper we argued for the conceptual and theoretical necessity of developing modality ontology to support the automatic assignment of modalities in the multimodal systems. After we briefly described some existing knowledge representation frameworks with a special focus on their positive aspects as well as their drawbacks, we proposed an alternative modality ontology. We introduced and described the main categories of the proposed ontology.
However, the arguments that support the postulation of this modality ontology are purely conceptual and theoretical in nature. Though the plausibility and validity of these conceptual arguments are demonstrated by identifying the existing conceptual gaps in the available taxonomies and metadata models for the design of ontologies to support the processes of automatic assignment of modalities within multimodal systems, the proposed modality ontology still requires further research to enlarge and validate. We are currently developing these two directions.

Acknowledgements

We would like to thank Betsy van Dijk for her valuable comments and suggestions on the improvement of the paper.

The work of one of the authors is part of the ICIS program (http://www.decis.nl/html/icis.html). ICIS is sponsored by the Dutch government under contract BSIK 03024.

References

[1] N.O. Bernsen. "Defining a Taxonomy of Output Modalities from an HCI Perspective". *Computer Standards and Interfaces*, 18:537–553, 1997.

[2] N.O. Bernsen, "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.

[3] W. Buxton. "Lexical and pragmatic considerations of input structures." *Computer Graphics* 17,1 (1983): 31-37.

[4] N. Elouazizi, Y. Bachvarova. "On Cognitive Relevance in Automatic Multimodal Systems". In *Proceedings of the Sixth IEEE International Symposium on Multimedia Software Engineering (ISMSE '04)* (Miami, Florida, USA, December 13-15, 2004). IEEE Computer Society, Los Alamitos, California, 2004, 418-426.

[5] J. Fodor. *The modularity of mind. An essay on faculty psychology*. Cambridge: MIT Press, 1983.

[6] MPEG7 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-7

[7] M. T. Maybury, editor. *Intelligent Multimedia Interfaces*. AAAI,Press,July,1993