

INTRODUCTION

1. THE ROLE OF MULTIMODALITY IN LANGUAGE AND SPEECH SYSTEMS

What is meant by multimodality and why is it important in language and speech systems? There are a variety of answers to the first part of this question. Definitions of multimodality can be very general, stemming from theoretical models of human information exchange. Other definitions can be based on a particular application framework. In essence, however, multimodality is the use of two or more of the five senses for the exchange of information. From this definition, we can directly infer the importance of multimodality in language and speech systems and provide a number of answers to the second part of the question. First of all, in natural, face-to-face communication, the three sensory modalities of sight, hearing and touch are all used and integrated to a considerable degree. It may be quite important also in human-system interaction to be able to make use of several modalities and modality integration. The use of several modalities may increase the naturalness of human-system communication and in turn lead to more flexible and more efficient systems which are more appealing and easier to use for the general population.

Ease of communication is not the only benefit of using several modalities in language and speech systems. From a system-design point of view, different modes of presentation can complement each other and provide different types of information. They can also reinforce each other and provide more complete information for the user during a difficult task, letting the user make use of different sensory modalities. Here the definition of modality becomes more complex. In his chapter on Modality Theory, for example, Niels Ole Bernsen makes the distinction between media and modality. In the context of this theory, a medium is the physical representation of information. Thus different media are closely related to sensory modalities such as the relationships between graphics and sight, acoustics and hearing, and haptics and touch. Modalities, on the other hand, are defined as different ways or modes of exchanging information. According to this definition, we can talk about different information channels, which can be established between the system and the user. The use of multiple channels is not only of great interest to the general design of systems, but it is also important to many people with sensory impairments as is explained by Alistair Edwards in his chapter on multimodal interaction and persons with disabilities.

Technological advances which have made possible rapid development of new interactive interfaces have also led to a considerable increase in interest in multimedia applications and multimodal information systems in general. The current status of such systems is, however, often more experimental than commercial, and

bears witness to the fact that there are still many problems to solve before such complex systems can be truly useful. The question of modality interaction, for example, is crucial to the development of multimodal systems. Which modality is best for a given task and in what situation? How do the different modalities reinforce or complement one another? In what situations or context is it a detriment to use one or the other modality? How do speech and language interact with other modes of presentation, and when is it advisable to use speech? What is the role of animated agents in such systems and how is speech perceived in a multimodal setting? These are just some of the questions that the different authors of this volume discuss and illuminate. The area of multimodality research is broad and comprises many disciplines. This book makes no pretence of covering the area completely, but rather represents selected perspectives on multimodality relevant to language and speech systems and includes several examples of current experimental and research oriented systems.

From the early oral storytelling tradition, where body and facial gestures have always played a major role, through classic theater, opera, dance, film, video and multimedia, multimodal communication has been constantly present in human communication and human culture. As Jens Allwood in his chapter on bodily communication points out, interest in the study of gestures and communication can be traced back to antiquity. Modern studies of multimodal communication, however, have their roots in the work of psychologists and linguistics during the past 50 years made possible by the advent of film and video analysis techniques. This book, therefore, takes as its point of departure, studies of human-to-human communication involving bodily communication (Allwood) and speech and gesture (McNeill et al.). The final chapter of the first section involves audio-visual speech perception (Massaro) where much of what we know is a result of experimentation using visual speech synthesis. Visual speech can also be used by the hearing-impaired to replace or reinforce the impaired modality of hearing. This is a good example of modality transform, which is elaborated on in the section on multimodality in alternative communication mediated by machine (Edwards). That chapter serves as a natural link to the final section on multimodality in communication between humans and systems. The section begins with a taxonomy and systematic analysis of input and output modalities, their representations and functions, and discusses practical problems of modality selection (Bernsen). This modality overview is followed by the final three chapters in which various aspects of experimental language and speech systems are described and tested. The first of these chapters (Brøndsted et al.) presents the architecture of a prototype system 'CHAMELEON' where input includes spoken language and pointing gestures and output comprises speech and a system pointer. The following chapter describes a computational model of natural turn-taking in goal-oriented face-to-face dialogue and presents results based on the implementation and testing of the model (Thórisson). The book concludes with a chapter on visual and auditory conversational signals for artificial talking heads and presents examples from different experimental spoken dialogue systems in several domains (Granström et al.).

2. MULTIMODALITY IN HUMAN-TO-HUMAN COMMUNICATION

2.1. Bodily Communication

A prerequisite to language and speech system design is the understanding of multimodal communication between humans. In his chapter on bodily communication, Jens Allwood presents us with a summary of the history of such research, and then places gestural and bodily communication in a human communication framework where the key concepts of intention and awareness are related to the sharing of different types of information. The chapter continues with examples of how bodily movement can be used as a primary means of expression and a review of some of the functions and content of such movements and gestures in human communication. Finally the chapter concludes with a section on the relationship between speech and gestures exemplifying three types of relationships: addition of information, change of information and reinforcement/support of information; and discusses the interaction between multimodal contributions from different communication partners. This final section is of considerable importance for system designers since one must understand, for example, when and what type of multimodal presentation is beneficial to the transfer of information. The section on the relationship between speech and gestures provides a natural transition to the following chapter on precisely that subject.

2.2. Speech and Gesture

In the chapter by David McNeill, Francis Quek, Karl-Erik McCullough, Susan Duncan, Robert Bryll, Xin-Feng Ma and Rashid Ansari on dynamic imagery in speech and gesture, the guiding argument is that gestures are part of our thinking process and that gestures and speech arise together from the same 'idea units.' They propose that the two modalities are not redundant but rather that each modality expresses its own part of the same semantic idea unit. They also introduce the concept of the 'catchment', which is a term they use for discourse units inferred on the basis of recurring gestural features. Video processing techniques are used to obtain gesture motion traces, which are compared to the fundamental frequency and power amplitude of the audio speech signal. Using the catchment concept, the authors are able to provide experimental evidence showing correlation between discourse segments signalled by gesture and those signalled by speech. The gestural features signalling the discourse segments in the material are explained and it is shown that these features also correspond to an independently derived text-based analysis of the discourse segmentation. The kind of analysis presented in this chapter and the correlation found between speech, gesture and discourse segmentation are important for both multimodal input and output considerations in speech and language systems.

2.3 Audio-visual speech perception

In his chapter on multimodal speech perception, Dominic Massaro presents a number of theoretical arguments and a large amount of empirical evidence supporting an approach to human speech perception as a multimodal phenomenon rather than as primarily an auditory one. One of the main arguments deals with the complementary nature of auditory and visual speech. When information from one modality is weak, it is complemented by stronger cues from the other modality. Massaro places spoken language understanding within an information-processing framework involving the sequential processing of different sources of information. A fuzzy logical model of perception (FLMP) has been developed in which the integration of the auditory and visual information sources takes place in an optimally efficient manner. This model is used to explain a variety of results from several audio-visual perception tests. In the tests, audio-visual synthesis in the form of a talking head is used to create test stimuli in which both audio and visual speech cues are systematically varied. The results of these tests give us insight into how humans combine and integrate audio and visual cues in speech perception. This knowledge is important in designing systems using talking heads as interactive agents. Massaro illustrates this in the context of visible speech in applications for the hearing-impaired. This type of application is the topic of the following section of the book.

3. MULTIMODALITY IN ALTERNATIVE COMMUNICATION

A special situation in multimodal human-human or human-system communication occurs when one or more human modalities are impaired or lost. The impairments can be of different types: sensory impairments such as deafness or blindness, physical impairment such as lack of motor control of the speech or manual articulatory system, or cognitive impairments such as dyslexia. Different ways of solving communication problems that occur in these cases are discussed by Alistair Edwards. Multiple modalities are very important in this context as an impairment in or a loss of one modality may be overcome by the exploitation of another. An example of this is the use of speech output from a computer to assist visually impaired users in reading a computer screen. A partial loss of a modality can also be remedied by enhancing input to that channel.

As is discussed in the chapter, this area of research puts heavy demands on the knowledge and understanding of the inherent differences between the senses and also on the different ways one particular modality can be used. Examples are given of existing and possible communication devices that may be used both in human-to-human and in human-system communication. The aims of these devices are to enhance the quality of life using the concept 'design for all.' The devices are not only useful for people with impairments but may also be useful in particular situations for everybody, for example in noisy environments. The challenges presented by the need for alternative communication are seen to stimulate innovations in system design which will be beneficial to all users.

4. MULTIMODALITY IN HUMAN-SYSTEM COMMUNICATION

4.1. Modality Theory Framework

With the increasing use of different modalities and combinations of modalities in human-system interaction, the need for a taxonomy and a systematic analysis of the modalities has evolved. Niels Ole Bernsen has addressed this challenging task in his chapter and presents a definition and discussion of a Modality Theory. The theory addresses the general problem: given any particular set of information which needs to be exchanged between the user and the system during task performance in a given context, identify the input/output modalities which constitute an optimal solution to the representation and exchange of that information. The taxonomy and theory accordingly provide a framework for describing different modalities and for deciding which modalities to choose or to avoid for a given application. Modality theory is specifically applied to speech functionality where the decision to use speech output and/or speech input in interaction design is discussed. The use of the theoretical framework in an interactive design support tool that assists in choosing between modalities in a specified situation is also demonstrated.

4.2. A Hands-Free System Integrating Language and Vision

Multimodal systems often include screen presentations of data and employ keyboards and screen pointing devices (e.g. mouse) as modes of communication for the user, which means that the user has to stay in front of a screen. This is not always necessary in multimodal systems. An example of an application where the user is not tied to a screen is described in the chapter by Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, and Kristian Olesen. Both input and output to this experimental system are multimodal using speech recognition, gesture recognition, natural language processing, speech synthesis, and a laser pointing system. The user may walk around in a room and ask for information from the system by speaking and/or by pointing at a floor-plan of a building placed on a table. The system response will be by speech and by pointing and tracing with a laser beam either at a specific position or indicating a route between locations on the floor-plan. This chapter gives us an example of how the integration of language and vision processing in a dialogue system can be achieved and shows us the potential of a hands-free system for information exchange.

4.3. A Computational Turn-Taking Model

Kristinn Thórisson's main interest in this chapter lies in the problem of turn-taking in a multimodal dialogue system, i.e. how should the system signal that it expects reactions from the user and how can the system perceive that the user is expecting the system to react. A good command of this problem will greatly improve the seamless operation of multimodal dialogue systems. Thórisson discusses data from studies of how humans behave in dialogues. He has used this data to develop a

Multimodality in Language and Speech Systems

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

2002, X, 243 p., Hardcover

ISBN: 978-1-4020-0635-7