Multimodality in Language and Speech Systems

Björn Granström, David House, and Inger Karlsson (editors)
(Royal Institute of Technology, Stockholm)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 19), 2002,
ix+241 pp; hardbound, ISBN
0-4020-0635-7, $82.00, £56.00, €89.00

Reviewed by
Michael Johnston
AT&T Labs—Research

Multimodality in Language and Speech Systems is a collection of papers that stem from a summer school on the topic held at the KTH Royal Institute of Technology in Stockholm, Sweden, in July 1999, under the auspices of the European Language and Speech Network (ELSNET). The volume's chapters address a range of related topics, including taxonomies and descriptive frameworks for analysis and examination of multimodal communication (Allwood, Bernsen), experimental analysis of the relationship between speech and hand gesture (McNeill et al.), audio/visual speech perception (Massaro), multimodality in assistive technology (Edwards), descriptions of implemented systems and architectures that support face-to-face multimodal interaction (Thórisson, Granström et al.), and an intelligent workspace (Brøndsted et al.).

As you might expect, given their origin as summer school presentations, the contributions here primarily do not present new work but rather summarize the authors' research programs or overviews of subareas of the field. As such, the volume is a good introduction to an increasingly important area of speech and language research and provides a solid entry point for more detailed reading. It should be of interest both for use in teaching and for researchers and scholars seeking an introduction to this area.

It should be noted that the contributions in this volume focus primarily on face-to-face multimodal interaction and do not provide an overview of other areas of multimodal interaction such as pen or voice interfaces to mobile devices. Also, the volume does not provide a detailed overview of computational models of multimodal language understanding and multimodal output generation. André (2003) provides an overview of these areas and could be used in teaching along with this volume, readings from Maybury and Wahlster (1998) and Cassell et al. (2000) to provide a more complete overview of the issues, theory, and practice of multimodal systems.

The chapter by Allwood, "Bodily Communication—Dimensions of Expression and Content," illustrates how body movements are essential in interactive face-to-face communication and argues for going beyond analysis of signaled, discrete, written symbols to develop a fuller picture of human communication. The article provides an excellent overview of research on bodily communication over the last century and presents a descriptive framework for analysis of multimodal communication. This framework combines Peirce's division of indexical, iconic, and symbolic information with dimensions of intentionality and awareness (indicate/display/signal). Allwood's contribution clearly illustrates the complexity of the "simultaneous multidimensional coupling" between multiple media of expression and multiple levels of content in face-to-face communication. This point is highly relevant for computational work, since it

explains why embodied conversational systems are so challenging to build: Failure to capture this complexity will lead to unnatural and stilted behavior on the part of artificial-agent communicators.

Like Allwood's, the chapter by Bernsen, "Multimodality in Language and Speech Systems—From Theory to Design Support Tool," provides a framework that can be used in the analysis of multimodal communication and the design of multimodal interactive systems. Whereas Allwood addresses the complexity of face-to-face communication, Bernsen addresses the broader range of interaction between humans, other humans, and machines, including graphical presentations and haptics. The goal of Bernsen's research program is to determine the basic properties of input and output modalities and from these to derive a comprehensive, relevant, and intuitive taxonomy of modalities and modality combinations (modality theory) and to use this theory to aid interaction designers in selecting which representational modalities to use for a given task, context, and user. This chapter provides a highly detailed elucidation and exemplification of a theory and taxonomy of output modalities and briefly describes how this has been used in the development of a hypertext encyclopedic reference tool to aid interaction designers. A number of asymmetries between output modalities and input modalities are addressed but, unlike for output, a comprehensive theory and taxonomy of multimodal input is not yet available. The chapter also summarizes research (Bernsen 1997; Bernsen and Dybkjær 1999) that shows how modality theory accounts for the great majority of claims made in the literature regarding speech functionality. One interesting aspect of Bernsen's modality theory is that, given the top-down development of the taxonomy from theoretical principles, it enables not just analysis of commonplace modalities, but also exploration of new kinds of modalities and modality combinations.

The chapter by McNeill et al., "Dynamic Imagery in Speech and Gesture," argues that human hand gestures are part of our thinking process and that speech and gesture are 'co-expressive': deriving from the same semantic source but able to express different aspects of it. This position is supported by results using the experimental paradigm developed by McNeill, Quek, and colleagues, which combines video-based motion tracking techniques with psycholinguistic analysis of discourse. The chapter presents the experimental method and analysis in detail but provides less detail on the underlying psycholinguistic theory. For this, the reader might want to consult other works (McNeill 1992, 2000). The experimental analysis demonstrates how hand use correlates tightly with the semantic content of discourse. In particular the kind of synchrony (antisymmetry or mirror symmetry) is shown to provide cues for discourse segmentation. Principles are also developed for analysis of the gesture signal, including a 'dominant motion rule' used to determine whether small hand movements are significant.

The chapter by Massaro, "Multimodal Speech Perception: A Paradigm for Speech Science," presents a very clear overview of work on audio/visual speech perception by Massaro and colleagues. The central tenet of the approach is that when evidence from multiple modes, such as audible and visible speech, are combined, the influence of one modality is greater to the extent that the other is ambiguous or neutral. This is captured by a formal model, the fuzzy logic model of perception (FLMP). The core of the chapter is the presentation of the results of a series of experiments that validate the FLMP as an accurate description of multimodal perception. The experiments address the combination of audible speech with lip movement, integration of written text and speech, word recognition, combination of paralinguistic and linguistic cues, and the combination of auditory and facial cues in the perception of emotion. The McGurk effect is also addressed. This chapter provides an excellent introduction to

the program of research pursued by Massaro and colleagues over the last 20 years and provides an entry point for more detailed reading in various books and articles such as Massaro (1998).

Edwards's chapter, "Multimodal Interaction and People with Disabilities," provides a clear (and inspiring) overview of the ways multimodal interface technology has been or could be applied to assisting users with sensory disabilities. The chapter starts with a clear presentation of the properties of different sensory channels and their relationship to modalities of communication and goes on to present a series of examples of interfaces that map one mode into another or use a combination of modes in order to assist people with disabilities.

The chapter by Thórisson, "Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action," addresses the complex problem of modeling turn-taking behavior in multimodal dialog. Starting from literature on human-human interaction, a series of hypotheses are developed regarding the properties of turn-taking behavior. The turn-taking mechanism is characterized as anticipatory, multi-level, highly parallel, and opportunistic. It involves logical combination of multiple sensory features and cues and receives higher (temporal) priority than content analysis and interpretation. Thórisson goes on to show how these hypotheses can be captured in a computational model in which interaction processing is split into three cooperating layers (reactive, process control, content) with differing temporal priorities, and he describes the implementation of the model in the Gandalf prototype. This is an interactive guide to the solar system that supports face-to-face multimodal communication with a synthetic character. A great deal of detail on the implementation is provided, though it is quite densely packed, so the reader may also want to consult Thórisson (1996; 1999) for a fuller understanding of the approach.

The chapter by Granström et al., "Speech and Gestures for Talking Faces in Conversational Dialogue Systems," provides a concise overview of work on audio-visual speech synthesis at KTH. Like Cohen and Massaro (1993), Granström et al.'s approach employs direct parameterization of a graphical model of the face (Parke 1982). In addition to presenting their approach to facial animation and audio-visual synthesis, the authors summarize two perceptual experiments. The first experiment (Teleface) examines the role of visual synthesis in speech intelligibility and its use as an aid to hearing-impaired individuals. For hearing-impaired subjects, adding a synthetic face in addition to the audio channel was found to be almost as much help as adding the natural face. The second experiment explores the relationship between eyebrow movement and intonational phrasing and prominence. Eyebrow movement was found to serve as an independent cue to prominence. The chapter concludes with a description of five different experimental dialogue systems that employ the KTH audio-visual synthesizer (Waxholm, Olga, August, AdApt, and a language tutor) and demonstrates the applicability of the technology to a broad range of application domains.

The chapter by Brødnsted et al., "Developing Intelligent Multimedia Applications," describes a platform for building applications that combine speech and vision developed at the University of Aalborg in Denmark. A sample application for providing campus information is presented. The system supports speech input and output, visual input (camera), and visual output (laser pointer). The authors provide an overview of the underlying system architecture, with a brief description of each component and an interesting example of one type of multimodal application. However, this is primarily a system overview and offers little detail on the approach to multimodal language processing and dialog management adopted.

References

André, E. 2003. Natural language in multimedia/multimodal systems. In Ruslan Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press.

Bernsen, Niels Ole. 1997. Towards a tool for predicting speech functionality. *Speech Communication*, 23:181–210.

Bernsen, Niels Ole and Laila Dybkjær. 1999. Working paper on speech functionality. Technical report, Esprit Long-Term Research Project DISC Year 2 Deliverable D2.10, University of Southern Denmark.

Cassell, Justine, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. 2000. *Embodied Conversational Agents*. MIT Press, Cambridge.

Cohen, Michael M. and Dominic W. Massaro. 1993. Modeling co-articulation in synthetic visual speech. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo.

Massaro, Dominic W. 1998. *Perceiving Talking Faces: From Speech Perception to Behavioral Principle*. MIT Press, Cambridge.

Maybury, Mark and Wolfgang Wahlster. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann, Los Altos, California.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

McNeill, David. 2000. Growth points, catchments, and contexts. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 7(1):22–36.

Parke, Frederic I. 1982. Parameterized models for facial animation. *IEEE Computer Graphics*, 2(9):61–68.

Thórisson, Kristinn. 1996. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Thórisson, Kristinn. 1999. A mind model for multimodal communicative creatures and humanoids. *International Journal of Applied Artificial Intelligence*, 13(4–5):449–486.

Michael Johnston is principal technical staff at AT&T Labs—Research. His research over the last seven years has focused on the theoretical basis and implementation of multimodal systems, and he leads the MATCH multimodal interface project. Johnston's address is Room E101, AT&T Labs—Research, 180 Park Avenue, Florham Park, NJ 07030, USA; e-mail: johnston@research.att.com.