

# The Handbook of Multimodal-Multisensor Interfaces

**Volume 3**

*Language Processing,  
Software, Commercialization,  
and Emerging  
Directions*

**Sharon Oviatt  
Björn Schuller  
Philip Cohen  
Daniel Sonntag  
Gerasimos Potamianos  
Antonio Krüger**



# **The Handbook of Multimodal-Multisensor Interfaces, Volume 3**



# ACM Books

## Editor in Chief

M. Tamer Özsu, *University of Waterloo*

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

### **The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions**

Editors: Sharon Oviatt, *Monash University*

Björn Schuller, *Imperial College London and University of Augsburg*

Philip R. Cohen, *Monash University*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*

Gerasimos Potamianos, *University of Thessaly*

Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*

2019

### **Data Cleaning**

Ihab F. Ilyas, *University of Waterloo*

Xu Chu, *Georgia Institute of Technology*

2019

### **Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework**

Robert J. Moore, *IBM Research-Almaden*

Raphael Arar, *IBM Research-Almaden*

2019

### **Heterogeneous Computing: Hardware and Software Perspectives**

Mohamed Zahran, *New York University*

2019

### **Hardness of Approximation Between P and NP**

Aviad Rubinstein, *Stanford University*

2019

### **Making Databases Work: The Pragmatic Wisdom of Michael Stonebraker**

Editor: Michael L. Brodie, *Massachusetts Institute of Technology*

2018

**The Handbook of Multimodal-Multisensor Interfaces, Volume 2:  
Signal Processing, Architectures, and Detection of Emotion and Cognition**

Editors: Sharon Oviatt, *Monash University*

Björn Schuller, *University of Augsburg and Imperial College London*

Philip R. Cohen, *Monash University*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*

Gerasimos Potamianos, *University of Thessaly*

Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*

2018

**Declarative Logic Programming: Theory, Systems, and Applications**

Editors: Michael Kifer, *Stony Brook University*

Yanhong Annie Liu, *Stony Brook University*

2018

**The Sparse Fourier Transform: Theory and Practice**

Haitham Hassanieh, *University of Illinois at Urbana-Champaign*

2018

**The Continuing Arms Race: Code-Reuse Attacks and Defenses**

Editors: Per Larsen, *Immunant, Inc.*

Ahmad-Reza Sadeghi, *Technische Universität Darmstadt*

2018

**Frontiers of Multimedia Research**

Editor: Shih-Fu Chang, *Columbia University*

2018

**Shared-Memory Parallelism Can Be Simple, Fast, and Scalable**

Julian Shun, *University of California, Berkeley*

2017

**Computational Prediction of Protein Complexes from Protein Interaction Networks**

Srikanth Srikari, *The University of Queensland Institute for Molecular Bioscience*

Chern Han Yong, *Duke-National University of Singapore Medical School*

Limsoon Wong, *National University of Singapore*

2017

**The Handbook of Multimodal-Multisensor Interfaces, Volume 1:  
Foundations, User Modeling, and Common Modality Combinations**

Editors: Sharon Oviatt, *Incaa Designs*

Björn Schuller, *University of Passau and Imperial College London*

Philip R. Cohen, *Voicebox Technologies*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*

Gerasimos Potamianos, *University of Thessaly*

Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*  
2017

**Communities of Computing: Computer Science and Society in the ACM**  
Thomas J. Misa, Editor, *University of Minnesota*  
2017

**Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining**  
ChengXiang Zhai, *University of Illinois at Urbana-Champaign*  
Sean Massung, *University of Illinois at Urbana-Champaign*  
2016

**An Architecture for Fast and General Data Processing on Large Clusters**  
Matei Zaharia, *Stanford University*  
2016

**Reactive Internet Programming: State Chart XML in Action**  
Franck Barbier, *University of Pau, France*  
2016

**Verified Functional Programming in Agda**  
Aaron Stump, *The University of Iowa*  
2016

**The VR Book: Human-Centered Design for Virtual Reality**  
Jason Jerald, *NextGen Interactions*  
2016

**Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age**  
Robin Hammerman, *Stevens Institute of Technology*  
Andrew L. Russell, *Stevens Institute of Technology*  
2016

**Edmund Berkeley and the Social Responsibility of Computer Professionals**  
Bernadette Longo, *New Jersey Institute of Technology*  
2015

**Candidate Multilinear Maps**  
Sanjam Garg, *University of California, Berkeley*  
2015

**Smarter Than Their Machines: Oral Histories of Pioneers in Interactive Computing**  
John Cullinane, *Northeastern University; Mossavar-Rahmani Center for Business and Government, John F. Kennedy School of Government, Harvard University*  
2015

**A Framework for Scientific Discovery through Video Games**

Seth Cooper, *University of Washington*  
2014

**Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers**

Bryan Jeffrey Parno, *Microsoft Research*  
2014

**Embracing Interference in Wireless Systems**

Shyamnath Gollakota, *University of Washington*  
2014

# **The Handbook of Multimodal-Multisensor Interfaces, Volume 3**

***Language Processing, Software,  
Commercialization, and Emerging Directions***

**Sharon Oviatt**

*Monash University*

**Björn Schuller**

*Imperial College London and University of Augsburg*

**Philip R. Cohen**

*Monash University*

**Daniel Sonntag**

*German Research Center for Artificial Intelligence (DFKI)*

**Gerasimos Potamianos**

*University of Thessaly*

**Antonio Krüger**

*Saarland University and German Research Center for Artificial Intelligence (DFKI)*

*ACM Books #23*



Copyright © 2019 by the Association for Computing Machinery  
and Morgan & Claypool Publishers

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan & Claypool is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

*The Handbook of Multimodal-Multisensor Interfaces, Volume 3*

Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, Antonio Krüger, editors

[books.acm.org](http://books.acm.org)

[www.morganclaypoolpublishers.com](http://www.morganclaypoolpublishers.com)

ISBN: 978-1-97000-175-4 hardcover

ISBN: 978-1-97000-172-3 paperback

ISBN: 978-1-97000-173-0 eBook

ISBN: 978-1-97000-174-7 ePub

Series ISSN: 2374-6769 print 2374-6777 electronic

DOIs:

<a href="#">10.1145/3233795</a> Book	<a href="#">10.1145/3233795.3233805</a> Chapter 8
<a href="#">10.1145/3233795.3233796</a> Preface	<a href="#">10.1145/3233795.3233806</a> Chapter 9
<a href="#">10.1145/3233795.3233797</a> Introduction	<a href="#">10.1145/3233795.3233807</a> Chapter 10
<a href="#">10.1145/3233795.3233798</a> Chapter 1	<a href="#">10.1145/3233795.3233808</a> Chapter 11
<a href="#">10.1145/3233795.3233799</a> Chapter 2	<a href="#">10.1145/3233795.3233809</a> Chapter 12
<a href="#">10.1145/3233795.3233800</a> Chapter 3	<a href="#">10.1145/3233795.3233810</a> Chapter 13
<a href="#">10.1145/3233795.3233801</a> Chapter 4	<a href="#">10.1145/3233795.3233811</a> Chapter 14
<a href="#">10.1145/3233795.3233802</a> Chapter 5	<a href="#">10.1145/3233795.3233812</a> Chapter 15
<a href="#">10.1145/3233795.3233803</a> Chapter 6	<a href="#">10.1145/3233795.3233813</a> Chapter 16
<a href="#">10.1145/3233795.3233804</a> Chapter 7	<a href="#">10.1145/3233795.3233814</a> Index/Bios/Glossary

A publication in the ACM Books series, #23

Editor in Chief: M. Tamer Özsü, *University of Waterloo*

Area Editor: Michel Beaudouin-Lafon, *Université Paris-Sud*

This book was typeset in Arnhem Pro 10/14 and Flama using ZzTeX.

First Edition

10 9 8 7 6 5 4 3 2 1

*This book is dedicated to our families, whose patience and support sustained the year-long effort required to organize, write, and manage different stages of this multi-volume project.*



# Contents

Preface **xvii**

Figure Credits **xxi**

Introduction: Toward the Design, Construction, and Deployment of Multimodal-Multisensor Interfaces **1**

    Multimodal Language and Dialogue Processing **1**

    Multimodal Behavior **2**

    Emerging Trends and Applications **3**

    Insights into the Chapters Ahead **4**

    References **18**

## **PART I MULTIMODAL LANGUAGE AND DIALOGUE PROCESSING 21**

### **Chapter 1 Multimodal Integration for Interactive Conversational Systems 23**

*Michael Johnston*

1.1 Introduction **23**

1.2 Motivations for Multimodal Input **25**

1.3 Early Approaches to Multimodal Fusion **32**

1.4 Unification-based Multimodal Fusion and Related Approaches **33**

1.5 Multimodal Grammars and Finite-state Approaches **41**

1.6 Incremental Multimodal Integration Using Event Logic and Visual Statecharts **47**

1.7 Multimodal Reference Resolution and Multimodal Dialog **50**

1.8 Applications of Machine Learning to Multimodal Integration **53**

1.9 Conclusion **59**

Focus questions **66**

References **67**

## **Chapter 2 Multimodal Conversational Interaction with Robots 77**

*Gabriel Skantze, Joakim Gustafson, Jonas Beskow*

- 2.1 Introduction 77**
- 2.2 The Importance of the Face in Interaction 80**
- 2.3 Giving the Robot a Face 82**
- 2.4 Modeling Human-Robot Interaction 83**
- 2.5 Turn-Taking 89**
- 2.6 Grounding and Feedback 93**
- 2.7 Joint Attention 95**
- 2.8 Conclusions 97**
- References 98**

## **Chapter 3 Situated Interaction 105**

*Dan Bohus, Eric Horvitz*

- 3.1 Introduction 105**
- 3.2 Situated Spoken Language Interaction 110**
- 3.3 Engagement 116**
- 3.4 Conclusion 135**
  - Acknowledgments 137
  - Focus Questions 137
  - References 138

## **Chapter 4 Software Platforms and Toolkits for Building Multimodal Systems and Applications 145**

*Michael Feld, Robert Neßelrath, Tim Schwartz*

- 4.1 Introduction 145**
- 4.2 Definitions 145**
- 4.3 Architecture of Dialogue Systems 146**
- 4.4 Dialogue Management Architectures 154**
- 4.5 Fusion and Communicative Functions 157**
- 4.6 Multimodal and Cross-Modal Reference Resolution 162**
- 4.7 Review of Existing Dialogue Platforms 164**
- 4.8 SiAM-dp—the Situation-Adaptive Multimodal Dialogue Platform 170**
- 4.9 Current Trends in Dialogue Architectures 180**
  - Focus questions 182
  - References 183

**Chapter 5 Challenge Discussion: Advancing Multimodal Dialogue **191****

*James Allen, Elisabeth André, Philip R. Cohen,  
Dilek Hakkani-Tür, Ronald Kaplan, Oliver Lemon,  
David Traum*

- 5.1** Introduction **191**
- 5.2** Discussion Questions **193**
- 5.3** Conversation **196**
- References **216**

**Chapter 6 Nonverbal Behavior in Multimodal Performances **219****

*Angelo Cafaro, Catherine Pelachaud, Stacy C. Marsella*

- 6.1** Introduction **219**
- 6.2** Embodiment: The Mind, Bodies, and Nonverbal Behavior **220**
- 6.3** Toward Building Multimodal Behaviors Control Models **226**
- 6.4** Approaches **228**
- 6.5** An Annotated Behavior Generation Example in VIB **241**
- 6.6** Conclusion and Future Trends **245**
- Focus Questions **249**
- References **252**

**PART II MULTIMODAL BEHAVIOR **263******Chapter 7 Ergonomics for the Design of Multimodal Interfaces **265****

*Alexis Heloir, Fabrizio Nunnari, Myroslav Bachynskyi*

- 7.1** Introduction **265**
- 7.2** The Generic Design Process **266**
- 7.3** Physical Ergonomics: Data Collection **269**
- 7.4** Physical Ergonomics: Experimental Models **278**
- 7.5** Motion Capture-based Biomechanical Simulation **282**
- 7.6** Summary and Future Research Directions **288**
- Focus questions **291**
- References **291**

**Chapter 8 Early Integration for Movement Modeling in Latent Spaces **305****

*Rachel Hornung, Nutan Chen, Patrick van der Smagt*

- 8.1** Introduction **305**
- 8.2** State of the Art for Motion Modeling **313**

8.3	Early Multimodal Integration for Motion Modeling	322
8.4	Use Case Implementation	334
8.5	Conclusion	339
	Focus questions	340
	References	340

## **Chapter 9 Standardized Representations and Markup Languages for Multimodal Interaction 347**

*Raj Tumuluri, Deborah Dahl, Fabio Paternò, Massimo Zancanaro*

9.1	Introduction	347
9.2	How the Standards Fit Together	353
9.3	The Importance of Declarative Languages for Describing Multimodal Interaction	357
9.4	Model-based Specifications for Multimodal Interaction	358
9.5	Modality Fusion and Media Synchronization	369
9.6	Multimodal Fission and Media Synchronization	372
9.7	Lessons Learned From the Implementation of Multimodal Standards	377
9.8	The Future and Open Challenges	381
9.9	Conclusions	385
	Focus Questions	385
	References	386

## **Chapter 10 Multimodal Databases 393**

*Michel Valstar*

10.1	Introduction	393
10.2	Need for Data	394
10.3	Existing Databases	398
10.4	Creating Your Own Database	412
	References	419

## **PART III EMERGING TRENDS AND APPLICATIONS 423**

### **Chapter 11 Medical and Health Systems 425**

*Daniel Sonntag*

11.1	Introduction	425
11.2	Clinical Systems	429
11.3	Non-Clinical Systems	433

11.4	Case Studies	438
11.5	Future Directions	458
11.6	Conclusion	462
	Focus Questions	464
	References	465

## Chapter 12 Automotive Multimodal Human-Machine Interface 477

*Dirk Schnelle-Walka, Stefan Radomski*

12.1	Introduction	477
12.2	HMI Evolution	477
12.3	Challenges and Opportunities	481
12.4	Multimodal In-Car Interaction	486
12.5	Summary and Outlook	512
	Focus Questions	514
	References	515

## Chapter 13 Embedded Multimodal Interfaces in Robotics: Applications, Future Trends, and Societal Implications 523

*Elsa A. Kirchner, Stephen H. Fairclough, Frank Kirchner*

13.1	Introduction	523
13.2	Inherently Safe Robots—a Prerequisite for Human-Robot Cooperation	527
13.3	Definition and Relevance of Embedded Multimodal Interfaces	534
13.4	Embedded Multimodal Interfaces in Robotic Applications	542
13.5	Future Trends: Self-Adapting Embedded Multimodal Interfaces and Societal Implications	554
13.6	Supplementary Digital Materials: Exoskeleton’s Mode Change Supported by Embedded Brain Reading—Approach and Evaluation	565
	Focus questions	569
	References	570

## Chapter 14 Multimodal Dialogue Processing for Machine Translation 577

*Alexander Waibel*

14.1	Introduction	577
14.2	Technology	581
14.3	Evolution of System Prototypes and Deployments	588
14.4	Multimodal Translingual Communication	604

14.5	Conclusion	613
	Focus Questions	614
	References	615

## **Chapter 15 Commercialization of Multimodal Systems 621**

*Philip R. Cohen, Raj Tumuluri*

15.1	Introduction	621
15.2	Aeronautics	621
15.3	Robotics	625
15.4	Biometrics	636
15.5	Assistants and Avatars	638
15.6	Product Search	643
15.7	Virtual and Augmented Reality	644
15.8	Field Force Automation	646
15.9	Insurance	647
15.10	Personal Care Products	648
15.11	Emotion Recognition	650
15.12	Summary	651
	Focus questions	651
	References	652

## **Chapter 16 Privacy Concerns of Multimodal Sensor Systems 659**

*Gerald Friedland, Michael Carl Tschantz*

16.1	Introduction	659
16.2	Calls for and Types of Privacy	662
16.3	Prior Work on Privacy Threats and Responses	667
16.4	Privacy Risks and Possible Attacks	674
16.5	Case Studies of Privacy Violations Using Multimedia Analyses	677
16.6	Future Directions For Research	685
	Focus Questions	692
	Acknowledgments	696
	References	696

**Index 705**

**Biographies 747**

**Volume 3 Glossary 761**

## Preface

The content of this handbook is most appropriate for graduate students and of primary interest to students studying computer science and information technology, human-computer interfaces, mobile and ubiquitous interfaces, affective and behavioral computing, machine learning, and related multidisciplinary majors. When teaching graduate classes with this book, whether in a quarter- or semester-long course, we recommend initially requiring that students spend 2 weeks reading the introductory textbook, *The Paradigm Shift to Multimodality in Contemporary Interfaces* (Morgan & Claypool Publishers, *Synthesis Lectures on Human-Centered Informatics*, 2015). With this orientation, a graduate class providing an overview of multimodal-multisensor interfaces then could select chapters from the current handbook, distributed across topics in the different sections. As an example, in a 10-week course the remaining 8 weeks might be allocated to reading select chapters on (1) theory, user modeling, and common modality combinations (2 weeks); (2) prototyping and software tools, signal processing, and architectures (2 weeks); (3) language and dialogue processing (1 week); (4) detection of emotional and cognitive state (2 weeks); and (5) commercialization, future trends, and societal issues (1 week). In a more extended 16-week course, we recommend spending an additional week reading and discussing chapters on each of these five topic areas, as well as an additional week on the introductory textbook, *The Paradigm Shift to Multimodality in Contemporary Interfaces*. As an alternative, in a semester-long course in which students will be conducting a project in one target area (e.g., designing multimodal dialogue systems for in-vehicle use), some or all of the additional time in the semester course could be spent (1) reading a more in-depth collection of handbook chapters on language and dialogue processing (e.g., 2 weeks) and (2) conducting the hands-on project (e.g., 4 weeks).

For more tailored versions of a course on multimodal-multisensor interfaces, another approach is to have students read the handbook chapters in relevant sections and then follow up with more targeted and in-depth technical papers. For

example, a course intended for a cognitive science audience might start by reading *The Paradigm Shift to Multimodality in Contemporary Interfaces*, followed by assigning chapters from the handbook sections on (1) theory, user modeling, and common modality combinations; (2) multimodal processing of social and emotional information; and (3) multimodal processing of cognition and mental health status. Afterward, the course could teach students different computational and statistical analysis techniques related to these chapters, ideally through demonstration. Students might then be asked to conduct a hands-on project in which they apply one or more analysis methods to multimodal data to build user models or predict mental states. As a second example, a course intended for a computer science audience might also start by reading *The Paradigm Shift to Multimodality in Contemporary Interfaces*, followed by assigning chapters on (1) prototyping and software tools; (2) multimodal signal processing and architectures; and (3) language and dialogue processing. Afterward, students might engage in a hands-on project in which they design, build, and evaluate the performance of a multimodal system. In all of these teaching scenarios, we anticipate that professors will find this handbook to be a particularly comprehensive and valuable current resource for teaching about multimodal-multisensor interfaces.

## Acknowledgments

In the present age, reviewers are one of the most precious commodities on Earth. First and foremost, we'd like to thank our dedicated expert reviewers, who provided insightful comments on the chapters and their revisions, sometimes on short notice. This select group included Antonis Argyros (University of Crete, Greece), Vassilis Athitsos (University of Texas at Arlington, USA), Nicholas Cummins (University of Augsburg, Germany), Randall Davis (MIT, USA), Jun Deng (audEERING, Germany), Jing Han (University of Passau, Germany), Anthony Jameson (DFKI, Germany), Michael Johnston (Interactions Corp., USA), Thomas Kehrenberg (University of Sussex, UK), Gil Keren (ZD.B, Germany), Elsa Andrea Kirchner (DFKI, Germany), Stefan Kopp (Bielefeld University, Germany), Marieke Longchamp (Laboratoire de Neurosciences Cognitives, France), David McGee (Nuance, USA), Vedhas Pandit (University of Passau, Germany), Diane Pawluk (Virginia Commonwealth University, USA), Jouni Pohjalainen (Jabra), Hesam Sagha (audEERING, Germany), Maximilian Schmitt (University of Augsburg, Germany), Dirk Schnelle-Walka (Harman International, Germany), Mark Seligman (Spoken Translation, Inc., USA), Gabriel Skantze (KTH Royal Institute of Technology, Sweden), Jim Spohrer (IBM, USA), Zixing Zhang (Imperial College London, UK), and the handbook's main ed-

itors. We'd also like to thank the handbook's eminent advisory board, 13 people who provided valuable guidance throughout the project, including suggestions for chapter topics, assistance with expert reviewing, participation on the panel of experts in our challenge topic discussions, and valuable advice. Advisory board members included Samy Bengio (Google, USA), James Crowley (INRIA, France), Marc Ernst (Bielefeld University, Germany), Anthony Jameson (DFKI, Germany), Stefan Kopp (Bielefeld University, Germany), András Lőrincz (ELTE, Hungary), Kenji Mase (Nagoya University, Japan), Fabio Pianesi (FBK, Italy), Steve Renals (University of Edinburgh, UK), Arun Ross (Michigan State University, USA), David Traum (USC, USA), Wolfgang Wahlster (DFKI, Germany), and Alex Waibel (CMU, USA). We all know that publishing has been a rapidly changing field, and in many cases authors and editors no longer receive the generous support they once did. We'd like to warmly thank Diane Cerra, our Morgan & Claypool Executive Editor, for her amazing skillfulness, flexibility, and delightful good nature throughout all stages of this project. It's hard to imagine having a more experienced publications advisor and friend, and for a large project like this one her experience was invaluable. Thanks also to Mike Morgan, President of Morgan & Claypool, for his support on all aspects of this project. Finally, thanks to Tamer Özsü and Michel Beaudouin-Lafon of ACM Books for their advice and support. Many colleagues around the world graciously provided assistance in large and small ways—content insights, copies of graphics, critical references, and other valuable information used to document and illustrate this book. Thanks to all who offered their assistance, which greatly enriched this multi-volume handbook. For financial and professional support, we'd like to thank DFKI in Germany, Monash University (Australia), and Incaa Designs, an independent 501(c)(3) nonprofit organization in the United States. In addition, Björn Schuller would like to acknowledge support from the European Horizon 2020 Research & Innovation Action SEWA (agreement no. 645094).



## Figure Credits

**Figure 1.9** From G. Mehlmann. and E. André. 2012. Modeling multimodal integration with event logic charts. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. pp. 125–132. Santa Monica, CA. Used with permission.

**Figure 1.11** Based on S. Oviatt and P. Cohen. 2000. Perceptual User Interfaces: Multimodal Interfaces that process what comes naturally. *Communications of the ACM* 43.3, pp. 45–53.

**Figure 1.12** From M. T. Vo. 1998. *A framework and toolkit for the construction of multimodal learning interfaces*. Ph.D. Thesis, Carnegie Mellon University, CMU-CS-98-129. Used with permission.

**Figure 3.2** Based on D. Bohus and E. Horvitz. 2009d. Open-world dialog: challenges, directions and prototype. In *Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Pasadena, CA.

**Figure 3.3** Based on Bohus and Horvitz, 2009b.

**Figure 3.4** Based on D. Bohus and E. Horvitz. 2009b. Dialog in the open world: platform and applications. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI 2009, pp. 31–38, Boston, MA.

**Figure 3.5** Based on D. Bohus and E. Horvitz. 2009a. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGdial'2009, pp. 244–252. London, UK.

**Figure 4.1** From M. T. Maybury and W. Wahlster. 1998. Intelligent user interfaces: An introduction. In M. T. Maybury and W. Wahlster, editors, *Readings in Intelligent User Interfaces*, pp. 1–13. Morgan-Kaufmann, San Francisco, CA. Used with permission.

**Figure 4.5** Based on S. L. Oviatt and P. R. Cohen. 2015b. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*, Ch. 7. Morgan & Claypool Publishers, San Rafael, CA.

**Figure 6.5** Permission to use this photo was granted by the Institute for Creative Technologies (<http://ict.usc.edu>).

**Figure 6.8** Permission to use this photo courtesy of Dr. Hannes H. Vilhalmsson (<http://www.ru.is/faculty/hannes/>).

**Figure 7.1** From ISO/IEC. 1998. Ergonomic requirements for office work with visual display terminals (VDTs)—part 11: Guidance on usability. The International Organization for Standardization. New York, NY.

**Figure 7.4** Based on J. Leijnse, N. H. Campbell-Kyureghyan, D. Spektor, and P. M. Quesada. 2008. Assessment of individual finger muscle activity in the extensor digitorum communis by surface emg. *Journal of Neurophysiology*, 100(6): 3225–3235.

**Figure 9.6** Courtesy of Openstream.com. Used with permission.

**Figure 11.2** From D. Sonntag and M. Möller. 2010. A multimodal dialogue mashup for medical image semantics. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pp. 381–384. ACM, New York. Used with permission.

**Figure 11.5** From RadSpeech. 2011. RadSpeech Project. <https://www.dfki.de/RadSpeech>. Accessed 10/31/2018. Used with permission.

**Figure 12.1** From <https://ustwo.com/blof/hmi-where-we-are-now/>. Accessed 06/14/2016. Used with permission.

**Figure 12.2** From <https://ustwo.com/blof/hmi-where-we-are-now/>. Accessed 06/14/2016. Used with permission.

**Figure 12.3** Based on R. Neßelrath and M. Feld. 2013. Towards a cognitive load ready multimodal dialogue system for in-vehicle human-machine interaction. In *Adjunct Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Eindhoven pp. 49–52.

**Figure 12.4** Based on D. Kern, and A. Schmidt. 2009. Design space for driver-based automotive user interfaces. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 3–10. Essen, Germany.

**Figure 12.7** From H. Richter and A. Wiethoff. 2011. Augmenting future in-vehicle interactions with remote tactile feedback. In *Adjunct Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '11, pp. 162–163. Used with permission.

**Figure 12.9** From [http://www.continental-corporation.com/www/pressportal\\_com\\_en/themes/press\\_releases/3\\_automotive\\_group/interior/press\\_releases/pr\\_2016\\_05\\_10\\_wheel\\_gestures\\_en.htm](http://www.continental-corporation.com/www/pressportal_com_en/themes/press_releases/3_automotive_group/interior/press_releases/pr_2016_05_10_wheel_gestures_en.htm). Accessed 6/14/2016. Used with permission

**Figure 12.11** From <http://www.bmwblog.com/2012/07/10/introducing-the-bmw-i-drive-touch-controller/>. Accessed 6/14/2016. Used with permission

**Figure 12.12** Based on R. A. Young. 2014. Self-regulation reduces crash risk from the attentional effects of cognitive load from auditory-vocal tasks. *SAE International Journal of Transportation Safety*, 250(October).

**Figure 12.13** Based on R. A. Young. 2014. Self-regulation reduces crash risk from the attentional effects of cognitive load from auditory-vocal tasks. *SAE International Journal of Transportation Safety*, 250(October).

**Figure 12.14** From H. Hofmann. 2014. *Intuitive speech interface technology for information exchange tasks*. Ph.D. Thesis. Universität Ulm, Germany. Used with permission.

**Figure 12.15** From H. Hofmann. 2014. *Intuitive speech interface technology for information exchange tasks*. Ph.D. Thesis. Universität Ulm, Germany. Used with permission.

**Figure 12.15** Based on D. L. Strayer, J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman. 2015. *Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems*. AAA Foundation for Traffic Safety.

**Figure 12.18** Based on <http://continental-head-updisplay.com/de/>.

**Figure 12.19** From D. L. Strayer, J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman. 2015. *Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems*. AAA Foundation for Traffic Safety. Used with permission.

**Figure 12.20** From Q. Rao, T. Tropper, C. Grunler, M. Hammori, and S. Chakraborty. 2014. Implementation of in-vehicle augmented reality. In *Mixed and Augmented Reality (ISMAR)*, 2014 IEEE International Symposium on Mixed and Augmented reality—Science & Technology, pp. 3–8. München, Germany. <http://www.3drealms.de/ismar-2014-ar-in-vehicle-implementation/>. Used with permission

**Figure 13.2** From *COMPI: Robot Arm with Compliant Control*. YouTube, uploaded by German Research Center for Artificial Intelligence, 8/5/2019. <https://youtu.be/mDODMNMC5zc>.

**Figure 13.5** From E. A. Kirchner, J. de Gea Fernández, P. Kampmann, M. Schröer, J. H. Metzen, and F. Kirchner. 2015. *Intuitive Interaction with Robots—Technical Approaches and Challenges*, pp. 224–248. Springer Verlag GmbH Heidelberg. Used with permission.

**Figure 13.7** From *iMRK: Intelligent Human-Robot Collaboration*. YouTube, uploaded by German Research Center for Artificial Intelligence, 6/10/2016. <https://www.youtube.com/watch?v=VoU3NbTyFtU>.

**Figure 13.8** From *VI-Bot: Final Active Exoskeleton*, YouTube, uploaded by German Research Center for Artificial Intelligence, 10/5/2011. [https://www.youtube.com/watch?v=3RhcgvRz\\_O8](https://www.youtube.com/watch?v=3RhcgvRz_O8).

**Figure 13.9** From M. Folgheraiter, M. Jordan, S. Straube, A. Seeland, S. K. Kim, and E. A. Kirchner. 2012. Measuring the improvement of the interaction comfort of a wearable exoskeleton. *International Journal of Social Robotics*, 4(3): 285–302. Used with permission.

**Figure 13.10** From M. Folgheraiter, M. Jordan, S. Straube, A. Seeland, S. K. Kim, and E. A. Kirchner. 2012. Measuring the improvement of the interaction comfort of a wearable exoskeleton. *International Journal of Social Robotics*, 4(3): 285–302. Used with permission.

**Figure 13.11** From *aBri: Recognition of Warnings During Teleoperation*. YouTube, uploaded by German Research Center for Artificial Intelligence, 6/25/2013. <https://youtu.be/8WEVZz6bpJU>.

**Figure 13.12** From E. A. Kirchner and R. Drechsler. 2013. A formal model for embedded brain reading. *Industrial Robot: An International Journal*, 40(6): 530–540. Used with permission.

**Figure 13.13** From *IMMI—Intelligent Man-Machine Interface*. YouTube, uploaded by German Research Center for Artificial Intelligence, 2/5/2016. [https://www.youtube.com/watch?v=zeFp\\_JBSBxA](https://www.youtube.com/watch?v=zeFp_JBSBxA).

**Figure 13.15** From E. A. Kirchner, S. K. Kim, M. Tabie, H. Wöhrle, M. Maurus, and F. Kirchner. 2016a. An intelligent man-machine interface—multi-robot control adapted for task engagement based on single-trial detectability of p300. *Frontiers in Human Neuroscience*, 10:291. ISSN 1662-5161. Used with permission.

**Figure 13.16** From *Exoskeleton Control via Biosignals*. YouTube, uploaded by German Research Center for Artificial Intelligence, 2/5/2016. <https://www.youtube.com/watch?v=BRpbZFOXdRk>.

**Figure 13.17** Based on E. A. Kirchner, M. Tabie, and A. Seeland. 2014. Multimodal movement prediction—towards an individual assistance of patients. *PLoS ONE*, 9(1): e85060, 01.

**Figure 13.18** From *aBri: Movement Prediction for Exoskeleton Control*. YouTube, uploaded by German Research Center for Artificial Intelligence, 6/25/2016. <https://youtu.be/fi4MKPTFg0>.

**Figure 13.19** Based on M. Folgheraiter, M. Jordan, S. Straube, A. Seeland, S. K. Kim, and E. A. Kirchner. 2012. Measuring the improvement of the interaction comfort of a wearable exoskeleton. *International Journal of Social Robotics*, 4(3): 285–302.

**Figure 15.1** USMC employee. Public Domain.

**Figure 15.4** Photo © 2015 Intuitive Surgical, Inc. Used with permission.

**Figure 15.5** Courtesy of Mobius Bionics LLC. Used with permission.

**Figure 15.6** Top row left to right: Pepper courtesy of SoftBank, all rights reserved; Pillo courtesy of Pillo, Inc. Middle row, left to right: iPal courtesy of AvatarMind Robot Technology; Mabu courtesy of Catalia Health Inc. Bottom row: Robokind Zeno from D. Cameron, A. Millings, S. Fernando, E. Collins, R. Moore, A. Sharkey, V. Evers, and T. Prescott. April 2015. The effects of robot facial emotional expressions and gender on child-robot interaction in a field study. In *4th International Symposium on New Frontiers in Human-Robot Interaction*. Used with permission.

**Figure 15.7** From A. Ross and N. Poh. 2009. Multibiometric systems: overview, case studies, and open issues. In M. Tistarelli, S. Z. Li, R. Chellappa, editors, *Handbook of Remote Biometrics for Surveillance and Security*, pp. 273–292. Springer. Used with permission.

**Figure 15.8** From Daylen [CC BY 4.0].

**Figure 15.10** Courtesy of Soulmachines.com. Used with permission.

**Figure 15.11** Courtesy of © Aelo Inc. Used with permission.

**Figure 15.13** Courtesy of Vusiz. Used with permission.

**Figure 15.14** Courtesy of Openstream.com. Used with permission.

**Figure 15.15** Courtesy of L’Oreal. Used with permission.

**Figure 15.16** Courtesy of Wearable X. Used with permission.

**Figure 16.6** Courtesy of Raffi Krikorian. Used with permission.

# **Introduction: Toward the Design, Construction, and Deployment of Multimodal-Multisensor Interfaces**

This third volume of the Handbook takes the contents of the first two volumes—namely, the motivations, foundational concepts, basic modality combinations, component analyses, and recognition and fusion techniques—to the next level. Volume 3 discusses how to design and build functioning multimodal-multisensor systems that can sustain real-world use. Now in worldwide use, natural input modalities such as speech have eclipsed the keyboard and mouse for mobile devices, and for many applications for which the user must engage in a search or question-answering. But currently, fielded voice-only systems have barely scratched the surface of what a high-performance multimodal system can provide. Studies of unimodal vs. multimodal interfaces have confirmed that multimodal ones are more accurate and preferred by users [Oviatt et al. 1997]. Comparisons of a multimodal speech and pen-based system with a graphical user interface (GUI) deployed by the US Government for the same task revealed a four-fold speed-up and significant user preference for the multimodal interface [Cohen et al. 2015]. Furthermore, as Volumes 1 and 2 have shown, multimodal-multisensor interaction can support entirely new classes of applications, such as emotion and trait (e.g., depression) detection, for which no GUI has ever existed. Building on Volumes 1 and 2, Volume 3 attempts to enable researchers into multimodal technologies and systems to climb still higher to attain more powerful multimodal human-AI interaction.

## **Multimodal Language and Dialogue Processing**

Part I of this volume investigates multimodal language and dialogue processing. In order to complete the multimodal interface revolution, researchers need to understand how to build systems quickly and effectively to engage users in multimodal

## **2** Introduction

dialogues, using whatever combination of modalities is most appropriate to the problem at hand. No single modality or set of modalities is privileged, and architectures need to be built that can easily accommodate their unique properties. First, we need to realize that utterance semantics and pragmatics are not merely linguistic concepts, but rather take on a more comprehensive meaning in a multimodal context. Human multimodal performances do not channel all the user's meaning into a single modality, but rather meaningful elements of the communication may be allocated to different modalities. A system's multimodal understanding then needs to reconstruct the intended meaning in order to decide how to respond. The need for a meaning representation has recently become a matter of some dispute in the natural language literature, as researchers in the deep neural network (DNN) paradigm emphasize end-to-end training of dialogue systems and eschew the idea of explicit meaning representations in favor of latent ones, which (somehow) cause dialogues to be properly conducted. Although the jury is still out, full end-to-end trained DNN dialogue systems, as opposed to DNN-based approaches to component technologies and to modality combinations (see Volumes 1 and 2), have avoided many of the hard problems that are so evident in true dialogues, especially multimodality. Part I opens with a chapter discussing how meaning fragments derived from individual modalities can combine at a semantic level to produce an integrated meaning representation. Part I then investigates in Chapters 2 and 3 how multimodal dialogues are situated in an open world, and how an embodied agent, such as a robot or avatar, would need to engage with one or more interlocutors to understand when it is being addressed, how to select the next agent to speak, etc. Part I then considers in Chapter 4 software and toolkits that enable the rapid construction of multimodal dialogue systems, a topic to which we return later in the book. Finally, the future of multimodal dialogue research is considered by an outstanding panel of experts in dialogue understanding in this volume's Challenge Discussion found in Chapter 5.

### **Multimodal Behavior**

Multimodal-multisensor interfaces do not merely support multimodal dialogue systems, but also natural interaction using the human body (e.g., gesture). Part II begins with Chapter 6, which discusses how a multimodal interface can generate coordinated verbal and non-verbal performances as part of a dialogue. The chapter emphasizes an international standard, which has supported many systems across the multimodal research community. In order for an avatar or robot to generate

believable action, systems need to understand and model how the human body moves. Chapter 7 discusses the assessment and modeling of human body motion from an ergonomics perspective, and Chapter 8 discusses a range of machine learning techniques for recognizing body motions. The generation of body motion discussed in Chapter 6 can be informed by the precise human body motion analysis discussed in Chapter 7. Moreover, the ergonomic goals discussed in Chapter 7 can benefit from the machine learning techniques for movement modeling discussed in Chapter 8, and the data captured by techniques discussed in Chapter 13 on embedded multimodal-multisensor interfaces for exoskeletons.

## **Emerging Trends and Applications**

Part III investigates additional encompassing topics, including a variety of support tools for building applications and doing multimodal research, as well as emerging and commercialized multimodal-multisensor applications and their privacy implications. In order to build systems that interoperate multiple devices, recognition technologies, and databases, the community needs standardized tools. Chapter 9 the first chapter of Part III, discusses the W3C multimodal standard, which numerous companies have adopted. Previous chapters lament the unavailability of multimodal corpora that enable machine learning. In order to create useful corpora, data gathering experiments need to be conducted that identify optimal ways to collect high-quality multimodal databases. Chapter 10 in Part III provides advice on collecting data, surveys existing databases, and discusses how to create new ones.

Multimodal-multisensor applications have increased in number and scope for the past decade, accelerating recently as voice and image recognition technologies have dramatically improved. Part III continues with four emerging application areas of multimodal-multisensor interface research, including applications to medical and health systems (Chapter 11), automotive human-machine interfaces (Chapter 12), embedded interfaces in robotics (Chapter 13), and machine translation systems (Chapter 14). In fact, companies have been extremely active in offering products based on multimodal-multisensor interface technologies. The penultimate chapter in Part III (Chapter 15) discusses many of these applications.

Finally, Volume 3 closes with a thought-provoking chapter on the privacy implications of multimodal-multisensor interfaces, especially at the Internet scale. All multimodal scientists should take seriously the impact of leaked private information based on natural modality capture. The topic is more complex than meets the

eye due to the compounding of multiple data types and inferences extracted from them, which expose more personal details about users than ever before.

In what follows, we discuss the specifics of each of this volume's chapters.

## Insights into the Chapters Ahead

### Part I. Multimodal Language and Dialogue Processing

In Chapter 1, Johnston examines the topic of multimodal *integration*, which takes place at the semantic level and derives a complex meaning representation from the understanding of multiple modalities. Examples of modality combinations that lead to multimodal integration include spoken language and gesture, speech and sketch, etc. [Cohen and Oviatt 2017]. Integration requires a common target meaning representation framework into which the individual meaning fragments derived from each modality are merged. A major difficulty occurs because people frequently do not simultaneously provide the modalities to be integrated. Rather, they may sequence the inputs such that the integration process must wait for the trailing modality before fusing their meanings [Oviatt et al. 1997]. A multimodal integration subsystem thus needs to be able to process simultaneous, sequential, or missing modalities, and to provide an integrated meaning representation for unimodal or multimodal inputs. Early approaches were generalized to employ the process of *unification*, an operation derived from theorem-proving and logic programming [Johnston et al. 1997]. Unification is well-suited to the task of integrating the semantic information from each modality in that it combines complementary information but rules out inconsistent information. In this chapter, Johnston presents details of multimodal integration via unification and shows how it leads to mutual disambiguation [Oviatt 1999], resulting in a more stable multimodal user interface than unimodal processing. The chapter then presents and contrasts alternative approaches to multimodal integration based on two types of finite-state models, the first is an approach in which words and gestures are handled as lattice elements, and a second approach in which they are handled as events in a state chart. The former approach offers speed as one processing advantage, whereas the latter approach is focused on incrementality of processing, i.e., the ability for a system to provide output while the input is occurring rather than wait until multimodal input is completed.

Chapter 1 continues with a discussion of applications of machine learning to multimodal integration. Early machine learning-based approaches to multimodal integration used various kinds of classifiers and/or mutual information processing

to align modalities and select the unimodal interpretations to combine. They then relied on heuristic or unification processes to perform the actual integration of information. Johnston suggests that future efforts at multimodal integration may apply the deep learning paradigm [Baltrušaitis et al. 2018], an approach that has many advantages and disadvantages, some of which are discussed in Chapter 5’s expert Challenge Discussion. A major limiting factor is there is at the present time no labeled corpus sufficient to train a supervised machine learning system to perform multimodal integration.

The next two chapters discuss multimodal *situated* dialogue. In Chapter 2, Skantze, Gustafson, and Beskow discuss issues that arise when interacting with robots multimodally. Human-robot interaction is by definition situated in a physical environment, which means that speaker(s)’ and robot(s)’ attunement to the physical situation plays a major role in motivating and conducting the conversation. Whereas other chapters in this Handbook delve into technical details of unimodal and multimodal recognition and fusion, Chapter 2 focuses on the building of a robot that can communicate multimodally. The chapter emphasizes the importance of the face in conducting a dialogue, especially the effects of gaze and head pose on turn-taking, grounding of referring expressions, and establishing joint attention. Unlike 2D avatars that appear to be looking in the same direction, no matter what direction they are viewed from (the so-called “Mona Lisa effect”), robots’ 3D faces can clearly indicate that the robot is looking toward someone and away from someone else. The chapter discusses how this ability enables the participants to select others for speaking turns [Bohus and Horvitz 2010]. The authors argue that mutual gaze between interlocutors is essential in establishing and maintaining joint attention, which is needed to coordinate interaction. The robot’s face can in principle enable emotion displays through the motion of eyebrows, eyes, mouth, head pose, etc., but most robots are incapable of providing such subtle facial cues. An exception is the Furhat robot which incorporates rear projection of a dynamic face onto a translucent head [Al Moubayed et al. 2013], thereby avoiding the difficult issues of building mechanical actuators to move simulated facial muscles. The chapter concludes its discussion of face-to-face communication with an analysis of turn-yielding and turn-holding cues that include, apart from gaze, cues from prosody, syntax, breath, and gesture on turn-taking.

Chapter 3 by Bohus and Horvitz provides more detail on multimodal, multi-party, situated interaction. The authors specifically investigate dialogue engagement, which involves initiating, maintaining, and terminating an interaction [Sidner et al. 2004]. The chapter discusses how to represent and reason about engagement state, actions, and intentions. Engagement state is modeled as a binary

## **6** Introduction

variable indicating whether an agent (including the system) is engaged in an interaction with another agent. The system can signal and track a person's intention to engage/disengage with the system via multimodal perception (user gaze, verbal signals, etc.). For engagement actions, the system can establish, maintain, or terminate engagement, which can be realized by a number of low-level actions (greetings, head or hand gestures, etc.) when the parties are appropriately situated in the physical space, such as within a given distance. The chapter provides a valuable discussion comparing probabilistic heuristic rules for tracking the above variables (e.g., for probability of disengagement) vs. machine-learned approaches. The heuristic rules can be inspected and debugged, but cannot easily be extended or scaled to many sources of data available from sensing devices. On the other hand, machine-learned classifiers for predicting engagement or disengagement may perform better than the heuristic rules, but may require extensive feature engineering and/or labeled data. The chapter illustrates how machine learning can overfit to the data, such that a system trained on engagement-related data collected in one physical environment might perform poorly with data collected from a different environment. It also discusses how a system that continually learns may alter its behavior, causing the user to behave differently, potentially resulting in errors. The chapter concludes with a discussion of how a system can decide to act, such as to engage a person in conversation. Should the system engage now or wait? When acting in the face of uncertainty, a system could wait longer in the hope of obtaining more reliable information (such as the user's intention to engage), but at a potential cost, including more sensing, lost opportunities, loss of reactivity, etc. To manage such tradeoffs, the authors discuss developing a system's ability to forecast future actions, such as disengagement and turn-taking.

The previous topics set the stage for Chapter 4 by Feld, Nefelrath, and Schwartz, which discusses the components involved in building multimodal dialogue systems, and shows how they can be incorporated into toolkits that facilitate their construction. The chapter begins with an overall architecture for multimodal dialogue systems, incorporating components for multimodal input and output, multimodal fusion and fission (splitting a semantic output into coordinated multimodal streams, such as voice, body motion, GUI etc.), and dialogue/meta-dialogue management. All components have access to and depend on context. Multimodal input streams incorporate all those discussed in this Handbook, but especially voice, gesture, gaze, etc. Multimodal integration and fusion then combines the meanings of the input streams into a coherent semantic contribution (see Chapter 1). The output of this process is then sent to a dialogue/meta-dialogue manager that determines how to respond. The dialogue manager subcomponent (see Chapter 5), determines the actions to perform, including speech, digital, and physical acts,

whereas the meta-dialogue manager coordinates turn-taking, engagement, etc. (see Chapters 2 and 3). Multimodal fission includes both the lower-level planning of a system response as well as the coordinated realization of that behavior across the available modalities (see Chapter 6). The chapter emphasizes the importance of representation in presentation planning with the dictum “no presentation without representation” [Wahlster 2006]. This approach is in contrast to so-called “end-to-end” trained dialogue systems that attempt to avoid all internal representations (e.g., [Vinyals and Le 2015]). Based on these general component descriptions, Chapter 4 then describes a number of multimodal dialogue toolkits, some of which are commercially available. The commercially offered toolkits must not only handle the multimodal processing, but also abstract the backend systems’ capabilities in order that the entire toolkit can provide reusable, domain-independent components. In doing so, the toolkits will rely on ontologies or other knowledge representations that describe those capabilities. However, it is a major undertaking to populate those knowledge representations in a rapid and generalizable way, especially to support machine learning. Finally, the chapter compares multimodal architectures and toolkits with current trends in dialogue systems, including virtual assistant architectures offered by major IT firms, chatbots, and machine-learned “slot-filling” dialogue systems, all of which are just beginning to consider the issues discussed in this handbook.

Chapter 5 is a Challenge Discussion among experts in dialogue processing on the topic of the technologies and approaches needed to advance multimodal dialogue. The experts include: James Allen (Univ. of Rochester/IHMC), Elisabeth André (Univ. of Augsburg), Phil Cohen (moderator, Monash Univ.), Dilek Hakkani-Tür (Amazon), Ron Kaplan (Stanford Univ.), Oliver Lemon (Heriot-Watt Univ.), and David Traum (Institute for Creative Technologies, Univ. of Southern California). The experts were asked to consider the following general questions:

- How can dialogue systems collaborate with their users?
- Can the current machine-learned approaches be extended to handle this kind of expected conversational behavior?
- How can dialogue systems derived from corpora of human-human or human-computer dialogues that support so-called “chatbot” dialogues be combined with dialogue techniques that support transactional dialogues?
- What are the “end-to-end” trained dialogue systems best used for?
- How can current dialogue technologies be extended to handle multimodal interaction?

## **8** Introduction

- What kinds of hybrid symbolic/statistical reasoning can support next generation multimodal dialogue?
- Why do machine-learned dialogue systems need to learn to converse again in another domain?
- How can dialogue management be made to be domain independent?
- What are promising areas for dialogue research in the near to medium term?

Most of the participants indicated that collaborative dialogue will require the conversants to recognize and respond to each others' plans and intentions. For many dialogue situations, people are assumed to be engaging in planful behavior when they engage a system, rather than merely talking to pass the time (but see the next question). Such systems will need to engage in plan recognition as a core capability, and to maintain domain-independent abstractions of the lifecycle of plans and goals, in which goals are adopted, attempted, abandoned, revised, refined, etc., which then are reflected in their utterances. Although the participants agreed on the limitations of end-to-end trained chatbot systems, they were identified as being particularly helpful in role-play dialogues, such as in support of virtual agents as job interviewers that the human engages in order to practice. Because the goal is that the user expends a significant amount of time thinking about the task (e.g., applying for a job), the system does not need to understand much about the user's responses, but only needs to maintain the illusion of dialogue. Of course, determined users could break such a system, but then again they would not have had a willing suspension of disbelief or desire to practice. It was pointed out that there are hybrid dialogue systems that incorporate data-driven responses with representations of dialogue state. Multimodal interaction was identified as being particularly challenging for a number of existing natural language processing technologies, especially those that require the system to process utterances in both forward and backward directions. Thus, such a system would need to wait until an utterance is finished before being able to process such time-critical examples as simultaneous gestures, or back-channels ("uh-huh"), which cannot be delayed more than 200 ms without disrupting the conversation, etc. Whereas the architectures discussed in Chapters 1 and 4 are directed toward applying and fusing context and information derived from various modalities, the Challenge Discussion identified that much current research into spoken dialogue architectures in the natural language processing literature has been investigating simple unimodal dialogue interactions. In order to apply machine learning to multimodal dialogue systems, there must be corpora of relevant multimodal dialogue data, which to

date are lacking. A major topic of discussion centered around transfer of learning and domain independence. Current engineered dialogue systems operate at a high level of abstraction based on plans and goals, enabling them to be relatively domain independent. To extend to a new domain requires connecting the backend data sources to the system's vocabulary and ontology. However, such systems can lack robustness to natural language variation. On the other hand, machine-learned dialogue systems can be robust to variation in their data, but have difficulty generalizing to new topics/domains for which there is little or no data. Essentially, they need to learn to converse from scratch for each domain. The panel viewed this topic as a major challenge.

## **Part II. Multimodal Behavior Modeling**

In Chapter 6, Cafaro, Pelachaud, and Marsella describe the state-of-the-art in multimodal behavior generation for embodied conversational agents. Multimodal performances are planned and executed to convey a desired message, subject to constraints on the interpersonal, physical, social, and cultural situation. Generally, multimodal performances are divided into generation of spoken and nonverbal outputs, with spoken output typically produced by a controllable text-to-speech system that can adjust volume, speaking rate, pitch accent, and intonation. Nonverbal outputs involve generation of a character's face and body motion to convey emotion, as well as different types of gestural effects (e.g., indicating referents via deictic gestures), and rhetorical functions. Chapter 6 provides an extensive discussion of how to build controllers for multimodal behavior generation. First, it describes the kinds of data that have been collected for facial and gestural performance, as well as their annotation in support of machine learning. Then it presents the SAIBA framework (Situation, Agent, Intention, Behavior, Animation) that is an international standard for multimodal behavior generation [Kopp et al. 2006]. Within this framework, there are processes for communicative intent planning, behavior planning, and behavior realization. Communicative intent planning results in expressions in the Function Markup Language (FML) that describe the communicative action(s) to be performed, locations of pitch accents, etc. FML expressions are input to behavior planning, which results in expressions in the Behavior Markup Language (BML) that describe the type and timing of head and body motions, deriving gestures from a "gesticon" (a set of primitive gestural elements). Finally, a behavior renderer produces timed outputs of the various streams of data. Researchers have built systems to coordinate production of FML and BML via rule-based, statistical, and hybrid methods. Given data in the SAIBA framework, the chapter provides descriptions of a number of complete systems that render behavior for embodied conversational

## 10 Introduction

agents, and illustrates their operation with a working example. Regarding future directions, machine learning of coordinated behaviors via neural networks is being attempted, but requires substantial annotated corpora, which as yet do not exist. The chapter concludes with numerous technical challenges for multimodal behavior, including behavioral coherency, timing, and incremental performance, as well as interactive challenges in which the system's output affects the user inputs because people will tend to mimic the virtual agent's performance [Giles and Smith 1979].

Chapter 7 by Heloir, Nunnari, and Bachynski discusses general techniques for measuring human body posture and motion in support of building next-generation interfaces. Such interfaces may be multimodal, in the use of natural modalities like gesture, but multimodal-multisensor technologies can also be used to perform the needed data collection. The field of ergonomics has long been providing posture and motion analyses for standard desktop mouse/keyboard interfaces in workplaces to estimate health risks and to assist in the design of new equipment. The goal is to establish how the body moves in relation to the human-machine system to be designed. The field also provides normative information to rehabilitation medicine that enables researchers and practitioners to assess whether the patient's body motions conform to a desired target motion [Da Gama et al. 2015]. (See also Chapter 14 for a commercial example). The chapter thoroughly describes the techniques used, which can be generally classified into physical measurement models and motion-capture based biomechanical models. Physical measurement models involve capturing data about the mechanical, electrical, and physiological processes of the human body, often by recording data from multiple sensors attached to it. Mechanical sensors include electromagnetic, inertial motion units, data gloves; electrical sensors measure neuromuscular activation, while physiological sensors capture such processes as heart motion, heart rate, respiration, and pupil size. However, these traditional physiological sensing techniques are difficult to employ when dealing with next-generation mobile interfaces. The chapter advocates for motion-capture based measurements, particularly optical techniques that require minimal setup and lower cost. Given the observed motion, a musculoskeletal model of the person can be built that predicts corresponding forces and movements. The chapter points out that such models are applicable to analysis of larger movements, but not for smaller finger gestures, and that there need to be correlational studies of the resulting models with users' perceived fatigue, stress, and tension in order to fully understand the impact of the models.

In Chapter 8, Hornung, Chen, and van der Smagt present a thorough analysis of the uses of *early fusion* for analyzing movement modeling. They identify that motion

modeling is important for rehabilitation, control of robots, and natural rendering of human and animal motion. The thesis of the chapter is that in recognizing and classifying movements, early fusion of modalities, such as video, electromyography (signals generated by muscle contraction), and sound will enable correlations to be exploited that would otherwise go unnoticed by a late-fusion approach. The chapter provides an overview of motion modeling principles, and then delves into a deep learning neural network approach using *variational autoencoders* (VAE, see Glossary of Chapter 8) that derives a lower-dimensional “latent” space within which the motions are encoded, all without expert-derived features. The authors show that this approach reduces dimensionality and separates motions in a 2D space better than classical principal component analysis, even with noisy data. Regarding multimodal analyses of movement, the authors advocate use of a mixed latent representation that combines latent representations from individual modalities, and especially a technique that omits data, and even entire modalities, during neural network training. The neural network learns the correlations of the modalities both with and without the full set of data, such that it can perform the classification during actual operations, even if data is corrupted or missing altogether. The chapter presents an extensive use case of motion modeling, providing recommendations and rationales for a specific neural network VAE architecture, including parameter settings and the use of specific types of neural units within the network. The chapter concludes by restating advantages that this early fusion approach offers for motion modeling, including reduction in dimensionality, inference of missing modality information, and ability to generate motions, such as for computer graphics. It claims the approach will be most accurate if the different modalities observe the same physical behavior. The authors suggest that it may be possible to use this approach for observation of different, although correlated, behavior, such as speech and gesture, but it remains to be studied whether the approach would be beneficial.

### **Part III. Multimodal Standards and Tools, Emerging Technologies, Commercialization and Societal Implications of Multimodal Systems**

The final part of this volume begins with a subsection discussing infrastructure tools that are very useful, if not essential, to the design of multimodal interfaces, namely industry standards and multimodal databases.

Chapter 9 by Tumuluri, Dahl, Paternò, and Zancanaro presents World Wide Web Consortium (W3C) industry standard languages for specifying multimodal interactive systems. The chapter first discusses the benefits of declarative specifications of multimodal interaction that support interoperability of multimodal

processing components and capabilities. Implementation details of those components/capabilities are not considered in the standard. Then, the chapter provides descriptions and examples of a number of languages that are variants of the XML markup language, including: InkML (digital ink), EmotionML (emotion), EMMA (multimodal fusion and fission), SCXML (Statechart XML, for interaction management), SSML (speech synthesis markup language), SMIL (Synchronized Multimedia Integration Language), VoiceXML (speech recognition), etc. Some of these languages are intended to be embedded in more encompassing descriptions, such as EmotionML's being embedded in EMMA or SMIL. EMMA specifies the kinds of information that a multimodal fusion/integration component needs and produces, such as the XML descriptions of the results of understanding individual modalities, including their timing, and the resulting multimodal fusion. EMMA does not specify how fusion is done. Beyond specifying in these languages the types of data being communicated, the W3C standard recommends a standard architecture, which includes: Modality Components (MCs), an Interaction Manager that routes messages, a Runtime Framework that support lifecycle events, and a component registration and discovery service that communicates with modality components through a Resource Manager. The MCs announce their capabilities to the Resource Manager, who ensures that the Interaction Manager can then route processing requests to them. Using the W3C specification, numerous companies have implemented multimodal systems, including AT&T, Microsoft, Openstream, and others. Such systems have included a wide array of modalities, including spoken language input and output, sketch, handwriting, computer vision, mouse/keyboard, GUIs, and others. Among the applications are personal assistants for the elderly, smart cities, and multimodal interfaces for banking, insurance, warehousing, and pharmaceutical applications. Finally, the chapter discusses lessons learned and future challenges for this methodology for building multimodal systems.

In Chapter 10, Valstar discusses multimodal databases. Although numerous chapters have noted the lack of multimodal corpora suitable for their particular interests, there have in fact been many corpora created for specific tasks such as emotion recognition. Often, the databases involve audio and video capture, but increasingly, other sensor data are being captured, including from digital pens, and physiological sensors (heart rate, galvanic skin response, EEG), etc. The chapter is written as a “how to” guide to the aspiring multimodal researcher in collecting high-quality data, as well as a resource that identifies where on the internet researchers can find existing corpora, how they were collected, and how they were annotated. The chapter discusses some rarely mentioned logistics involved in multimodal data collection, including advice on: the interplay between experimental design

and the size of the resulting databases, ethical questions that arise now under the European GDPR legislation with regard to how and where data can be stored and accessed (e.g., data that can be used to identify individuals might not be storable in a cloud system in another legal jurisdiction), the benefits of pilot testing, and how to distribute data. Finally, an invaluable section of the chapter concerns tips that should be considered when collecting a database, such as technical advice regarding methods to synchronize many sensors.

The next portion of Part III consists of chapters discussing four emerging application areas for multimodal-multisensor interaction, including applications to healthcare, automotive cockpits, embedded robotics, and machine translation.

In Chapter 11, Sonntag discusses multimodal-multisensor interfaces for clinical and non-clinical applications of medical and health systems. The chapter presents overviews of some of the major issues involved in applying these new modalities to the construction of cyber-physical healthcare systems in settings that require high-confidence software, offer embedded real-time operation, and provide support to overloaded clinical staff. Sonntag identifies active multimodal interfaces that enable the user to be more expressive in updating an electronic medical record by speaking, drawing, and handwriting their notes and annotations (in addition to entry via standard GUIs), using standard medical ontologies. In addition, multimodal-multisensor interfaces can passively track their users, e.g., tracking eye gaze or body motion via computer vision, which can enable quantitative assessment of the usage of clinical EMR systems, as well as control of clinical systems via gesture recognition. The chapter mentions a major research trend to build multimodal-multisensor interfaces for robotic surgery, especially to shorten the time it takes a surgeon to become proficient (see also Chapter 15 for a discussion of commercial efforts to incorporate such interfaces into robotic surgery). In terms of non-clinical systems, multimodal interfaces are tracking people's facial expressions, gaze direction and emotional state via combined use of computer vision and speech recognition. Other wellness applications are being built around data collected from body-worn sensors, such as smartwatches, Fitbits™, etc., supporting such applications as activity monitoring for seniors, health tracking for patients after they have been discharged from hospital, and social companions for depressive individuals. The chapter provides three case studies of current research into use of multimodal and multi-sensor interfaces. The first is a clinical application—a multimodal dialogue system for annotation and retrieval of radiological images using voice and sketch [Sonntag et al. 2010, Sonntag et al. 2012] (see also Cohen and Oviatt [2017] and Chapter 1 in this Handbook for general discussion of multimodal speech and sketch technology). The second volume case study is neurocognitive

testing, in which a patient uses a digital pen and paper rather than the classic ordinary pen and paper to take an exam. Digital pen and paper versions of classical tests such as the clock drawing test [Davis et al. 2014, Prange et al. 2018] are useful for assessing Alzheimer's disease and dementia, and other tests (Trail-making, Rey-Osterreith, etc.) are useful for assessing traumatic brain injury [Salzman et al. 2010, Tiplady et al. 2003] and Parkinson's disease [Darnall et al. 2012]. Numerous advantages of the multimodal version of these tests are discussed, both in terms of increased accuracy and objectivity, as well as better use of clinicians' time. In a third case study, the chapter proposes to expand the digital pen sensing to include electrodermal activity in order to assess stress, arousal, and affect.

Chapter 12 by Schnelle-Walka and Radomski discusses the state of multimodal human-machine interfaces (HMIs) in automobiles. It presents the progression of automotive cockpit interfaces from a plethora of knobs and buttons to today's interfaces that add voice, touchscreens, GUIs, eye gaze, heads-up displays, etc. These new interface modalities are designed both to control the operation of the car, and also to deliver "infotainment"—information (e.g., maps, data on points of interest) and entertainment (e.g., radio, music). However, it is not obvious which combination of modalities best serves the driver's needs. Of paramount importance is the requirement that the interface not raise the driver's cognitive load excessively, leading to distraction. The United States National Highway Traffic Safety Administration (NHTSA) has provided guidelines for automotive interface technology that recommends that the user interface not draw the driver's focus of attention off the driving task for more than 2 s, and not result in an interaction cycle lasting longer than 12 s. These objectives are (perhaps surprisingly) difficult to meet. The chapter discusses various research studies of automobile HMIs that investigate typical modalities and modality combinations. In particular, the chapter discusses in-car studies sponsored by the American Automobile Association and conducted at the University of Utah that have indicated that HMIs in vehicles take too much time and create too much distraction, although the situation is improving. Notably, in the most recent study of on-the-road performance of 17 vehicle types and 4 types of tasks, voice interaction resulted in lower cognitive demand, visual demand, and interaction time as compared to existing car touchscreen or button interfaces for adjusting audio entertainment and placing phone calls [Strayer et al. 2017]. However, no modality met the NHTSA guidelines for interacting with the in-car navigation systems. A second trend, as seen in the Tesla vehicle is to provide large touchscreens and GUIs to control the various car systems, but those invariably require drivers to look at the screen, thus taking their eyes off the road. Automobile manufacturers are not only striving to build multimodal HMIs, both for lowering

driver cognitive load, but they are also looking forward to passenger interactions with infotainment systems in autonomous vehicles for which *driver* cognitive load is not an issue. The chapter concludes with recommendations for studies of combined use of modalities (e.g., speech + gaze), effects on cognitive load of heads-up displays on the windshield, and the design of the automobile HMI to support the “handover” problem for autonomous driving, when the car detects that it is no longer able to drive safely.

In Chapter 13, Kirchner, Fairclough, and Kirchner discuss embedded multimodal-multisensor interfaces for robotics. A number of important distinctions are drawn in the chapter. First, they distinguish between explicit vs. implicit interaction between human and robot. For example, humans may explicitly attempt to control the robot via spoken command. The robot may also respond implicitly to a user by passively sensing user states, such as mental workload, fatigue, etc. Some of the signals given by people are overt, available to be sensed from a distance (e.g., via microphones or computer vision), whereas others are covert (e.g., their psychophysical or neurophysiological states) and generally require body-based sensing. As its primary example, the chapter considers arm exoskeletons that are driven by biosignals from the arm and the brain, although it raises similar issues for robotics in general. First, the authors discuss how multi-sensor interfaces are required for safety with and around robots, pointing out that the robots need to sense location and force in order to avoid incorrectly contacting objects and people. In the case of exoskeletons, the forces exerted by the exoskeleton on the human arm must be carefully controlled, such that the exoskeleton’s motion can be prevented if it moves too far, too fast, or must remain stationary. Critically, the authors show how brain signals that indicate implicitly the user’s intention to move the arm+exoskeleton can prepare the exoskeleton to release the motion “lockout” and become more sensitive to the user’s force; when the user then exerts force, the exoskeleton can provide a smooth response. The chapter shows that a temporally cascaded interface (see also [Oviatt and Cohen 2000], and [Zhai et al. 1999]) consisting of passive (brain) and active (human arm force) sensing is essential to the operation of the exoskeleton and leads to accurate arm placement. The embedded multimodal interface is shown to facilitate interactions in virtual worlds (or potentially in remote worlds, such as in space) when the user is also equipped with a head-mounted display and can monitor the arm’s remote location. The chapter concludes with a discussion of the trust and privacy issues entailed by embedded multimodal-multisensor interfaces that capture biosignals. Such systems are making explicit or implicit inferences about the behavior and physiology of the user, translating those inferences into system responses. Users need to be able to predict what the system is doing, and adapt to

it, and vice versa. The user needs to be able to trust that s/he and the system will converge. Regarding privacy, the authors point out examples of data being captured that could perhaps have negative ramifications for the user. To combat such possibilities, the authors argue that users should own their multimodal-multisensor data and explicitly consent to its usage (see also Chapter 16).

Chapter 14 by Waibel presents multimodal processing for machine translation. Spoken language translation is often regarded as a combination of three processes—speech recognition, machine translation, and speech synthesis in the target language. Each of these technologies has undergone a major revolution in capability. Starting in 1990 and continuing until quite recently, the same hidden Markov model statistical techniques, which were trained on bilingual corpora such as the proceedings of the Canadian parliament, became used for training both machine translation and speech recognition systems. This chapter discusses the deep neural network technology currently being used and the evolution of system prototypes, resulting in today's widely used speech translation systems. Once the personal digital assistant (PDA) device and smartphone were available, handheld devices became a preferred platform on which to deploy spoken translation, enabling numerous applications, including medical and humanitarian operations across the globe. Clients could engage the medical or humanitarian personnel, speak to the device, and see and/or hear the providers' responses translated into their own language. Besides mobile devices, workstation-based systems were used in the European parliament to assist simultaneous interpreters via multimodal interfaces. Multimodality thus enters the picture both in presenting results of speech recognition and translation, and in supporting their editing on screen-based hardware. The latter was important because it was found that correcting recognition errors by switching modalities provided a more effective system that overcame stubborn recognition errors. Multimodal interaction for translation also enabled different document types to be translated. An important example, now incorporated into Google Translate™ and Bixby Vision™, is road sign translation, in which a person takes a photo with a mobile phone of a sign in one language, which is then translated into that same sign using another language. This involves optical character recognition rather than speech recognition, but the basic system is similar. Another powerful example of multimodal translation is the real-time translation of lectures for multilingual audiences, in which a running translation of the lecture is presented alongside translated versions of the graphical lecture material, which informs the speech recognition system so that technical terms are properly translated. The chapter concludes with a discussion of other emerging multimodal translation applications, including speech translation displayed via augmented reality glasses.

In Chapter 15, Cohen and Tumuluri present a survey of the rapidly developing commercial applications of multimodal-multisensor interfaces. The chapter begins with discussion of one of the most impactful and expensive multimodal-multisensor interfaces, the helmet of the F-35 Joint Strike Fighter. The pilot is literally and virtually surrounded by a multimodal-multisensor fusion interface that integrates information from multi-spectral sensing, multiple aircraft, and ground-based sensing into a coherent 3D image. The pilot's situational awareness is also supported by head and eye-tracking, and he can control aircraft systems with voice interaction. Chapter 15 discusses multimodal-multisensor interfaces for robots, including autonomous vehicles, surgical robots, prosthetic arms, and social robots. The authors discuss the advantages of multimodal biometrics, which are used in many other applications of multimodal interaction, especially for those demanding security. Other multimodal technologies embedded in commercial applications include infotainment in automobiles (discussed in Chapter 12), multimodal emotion recognition (e.g., for in-car usage or advertising), and multimodal-multisensor augmented reality headsets that are used, for example, in warehouse automation. Finally, the chapter discusses progress in developing consumer-facing multimodal avatars and virtual assistants that have been used in many different applications, including mobile phone-based assistants, product search, health, personal care, field force automation, and insurance.

Finally, Chapter 16 by Friedland and Tschantz closes Volume 3 with a thought-provoking analysis of the privacy concerns that arise from multimodal-multisensor systems. The chapter first provides a rigorous analysis of privacy for structured data based on the concepts of *k-anonymity* and *ε-differential privacy* from computer science (see the Glossary in Chapter 16). However, unstructured multimodal-multisensor data provides substantial new challenges. The chapter discusses how a determined investigator might link supposedly anonymous records from different data sources to gather identifying information about a person. Although there may be laws and regulations requiring anonymity, statistical properties of the data may allow unwanted inferences to be made. The chapter highlights many types of inferences that determined analysts could make by correlating multiple data sources, including multimodal-multimedia ones. For example, users of social media often want to share image or video content with others. Although the subject of the content may have given permission for publication of the imagery, people in the background of the image may not have. With modern face recognition technologies, they could be automatically “tagged” as being in the image/video. Moreover, with geo-tagging metadata supplied by the camera itself, information about the creation time and location of the photo is then available, which can make “cybercasing” possible, by which thieves determine whom to target based on object

detection and location data from digital photos. The authors point out that close examination of any digital camera image provides a unique noise signature of the camera's sensor. Thus, one could tell if two images were taken by the same camera, thereby leaking even more information. Similar inferences could perhaps be made for audio recordings via background audio scene analysis and voice recognition. The chapter highlights potential risks to privacy protection from the many types of multimodal-multimedia sensors that are proliferating in our devices. The key message of the authors' discussion is that it is the "aggregate of the user's complete online footprint that needs protection." They argue for mitigation research on how to cloak multimodal data from recognition algorithms, such as to digitally blur images of people in the backgrounds of photos. The chapter concludes with a discussion of the importance of educating the next generation of users, i.e., teenagers and younger adults, about digital privacy. The authors' Teach Privacy Project aims to provide educators with classroom tools that show how one's privacy can be breached with seemingly innocuous behavior, such as sending a Tweet, and offers the students ten principles for online privacy protection. Finally, the chapter contains an extensive set of focus questions designed to extend the reader's understanding of these very important concepts.

## References

- S. Al Moubayed, G. Skantze, and J. Beskow. 2013. The Furhat back-projected humanoid head—Lip reading, gaze and multiparty interaction. *International Journal of Humanoid Robotics*, 10(1). DOI: [10.1142/S0219843613500059](https://doi.org/10.1142/S0219843613500059)Cited. 5
- T. Baltrušaitis, C. Ahuja, and L-P, Morency. 2018. Challenges and applications in multimodal machine learning. In S. L. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107993](https://doi.org/10.1145/3107990.3107993). 5
- D. Bohus and E. Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of the 12th International Conference on Multimodal Interfaces/7th International Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI 2010, Beijing, China. DOI: [10.1145/1891903.1891910](https://doi.org/10.1145/1891903.1891910). 5
- P. R. Cohen and S. Oviatt. 2017. Multimodal Speech and Pen Interfaces. In S. L. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan and Claypool Publishers, San Rafael, CA. DOI: [10.1145/3015783.3015795](https://doi.org/10.1145/3015783.3015795). 4, 13

- P. R. Cohen, E. C. Kaiser, M. C. Buchanan, S. Lind, M. J. Corrigan, and R. M. Wesson. 2015. Sketch-Thru-Plan: a multimodal interface for command and control. *Communication*, 58(4): 56–65. ACM. [1](#)
- A. Da Gama, P. Fallavollita, V. Teichrieb, and N. Navab. 2015. Motor rehabilitation using kinect: a systematic review. *Games for Health Journal: Research, Development, and Clinical Applications*, 4(2). DOI: [10.1089/g4h.2014.0047](https://doi.org/10.1089/g4h.2014.0047). [10](#)
- R. Davis, D. Libon, R. Au, D. Pitman, and D. Penney. 2014. THink: Inferring cognitive status from subtle behaviors. In *Proceedings of the Innovative Applications of Artificial Intelligence, Association for the Advancement of Artificial Intelligence*. AAAI. DOI: [10.1609/aimag.v36i3.2602](https://doi.org/10.1609/aimag.v36i3.2602). [14](#)
- N. D. Darnall, C. K. Donovan, D. C. Aktar, S. H. Tseng, P. Barthelmess, P. R. Cohen, and D. C. Lin. 2012. Application of machine learning and numerical analysis to classify tremor in patients affected with essential tremor or Parkinson's disease. *Gerontechnology*, 10(4): 208–219. DOI: [10.4017/gt.2012.10.4.002.00](https://doi.org/10.4017/gt.2012.10.4.002.00). [14](#)
- H. Giles and P. Smith. 1979. Accommodation theory: optimal levels of convergence. In H. Giles, R. N. St. Clair, editors, *Language and Social Psychology*. Baltimore: Basil Blackwell. Baltimore, MD. [10](#)
- M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. 1997. Unification-based multimodal integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 281–288. Association for Computational Linguistics. DOI: [10.3115/976909.979653](https://doi.org/10.3115/976909.979653). [4](#)
- S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: the behavior markup language. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, pp. 205–217. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/11821830\\_17](https://doi.org/10.1007/11821830_17). [9](#)
- S. L. Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 576–583. ACM Press. DOI: [10.1145/302979.303163](https://doi.org/10.1145/302979.303163). [4](#)
- S. L. Oviatt and P. R. Cohen. 2000. Multimodal interfaces the process what comes naturally. *Communications of the ACM*, 43(3): 45–53. DOI: [10.1145/330534.330538](https://doi.org/10.1145/330534.330538). [15](#)
- S. Oviatt, A. DeAngeli, and K. Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of Conference on Human Factors in Computing Systems*, pp. 415–422. ACM Press. DOI: [10.1145/258549.258821](https://doi.org/10.1145/258549.258821). [1, 4](#)
- A. Prange, M. Barz, and D. Sonntag, 2018. A categorisation and implementation of digital pen features for behaviour characterisation. *DFKI Technical Report*. [14](#)
- K. Salzman, P. R. Cohen, and P. Barthelmess. July 2010. ADAPX digital pen: TBI cognitive forms assessment. *Report to US Army Medical Research and Materiel Command*, Fort Detrick, MD. [14](#)

## 20 Introduction

- C. Sidner, C. Kidd, C. Lee, and N. Lesh. 2004. Where to look: A study of human-robot engagement. In *Proceedings of the 2004 International Conference on Intelligent User Interfaces*, pp. 78–84. ACM Press. DOI: [10.1145/964442.964458](https://doi.org/10.1145/964442.964458). 5
- D. Sonntag and M. Möller. 2010. A multimodal dialogue mashup for medical image semantics. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pp. 381–384. ACM, New York. DOI: [10.1145/1719970.1720036](https://doi.org/10.1145/1719970.1720036). 13
- D. Sonntag, C. Schulz, C. Reuschling, and L. Galarraga. 2012. Radspeech's mobile dialogue system for radiologists. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pp. 317–318. ACM, New York. DOI: [10.1145/2166966.2167031](https://doi.org/10.1145/2166966.2167031) 13
- D. L. Strayer, J. M. Cooper, R. M. Goethe, M. M. McCarty, D. Getty, and F. Biondi. September 2017. Visual and cognitive demands of using in-vehicle infotainment systems. *AAA Foundation for Traffic Safety*. 14
- B. Suhm, B. Myers, and A. Waibel. 1996. Interactive recovery from speech recognition errors in speech user interfaces. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 861–864. IEEE. DOI: [10.1109/ICSLP.1996.607738](https://doi.org/10.1109/ICSLP.1996.607738).
- B. Tiplady, R. Baird, H. Lutcke, G. Drummond, and P. Wright. 2003. Use of a digital pen to administer a psychomotor test. *Journal of Psychopharmacology*. 17(supplement 3):A71. 14
- O. Vinyals and Q. Le. 2015. A neural conversational model. In *Proceedings of the International Conference on Machine Learning*. 7
- W. Wahlster. 2006. Dialogue systems go Multimodal: The SmartKom experience. In W. Wahlster, editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pp. 3–27. Cognitive Technologies Series, Springer, Heidelberg, Germany. DOI: [10.1007/3-540-36678-4\\_1](https://doi.org/10.1007/3-540-36678-4_1). 7
- S. Zhai, C. Morimoto, and S. Ihde. 1999. Manual and gaze input cascaded(MAGIC) pointing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99)*, pp. 246–253. ACM Press, NY. DOI: [10.1145/302979.303053](https://doi.org/10.1145/302979.303053). 15



**PART**

# **MULTIMODAL LANGUAGE AND DIALOGUE PROCESSING**



# Multimodal Integration for Interactive Conversational Systems

Michael Johnston

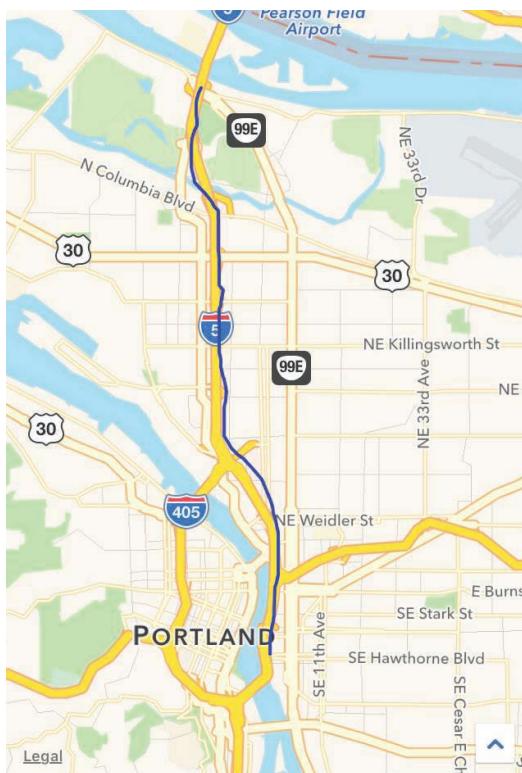
## 1.1

### Introduction

This chapter discusses the challenges inherent in combining information conveyed through multiple different modes of input and summarizes different technical approaches that have been taken to this problem of *multimodal integration* or *multimodal fusion*.<sup>1</sup> Specifically, we focus on deliberate input made by users to multimodal interactive conversational systems where their contribution is distributed across multiple modes.

A typical example of this kind of multimodal interaction would be input to an application running on a tablet, smartphone, or other device with a touch-sensitive screen where the user issues commands, queries, or statements consisting of a synchronized combination of spoken input with touching or drawing on the display. For example, in the case of a mobile navigation application, the user might speak a query such as “what is the traffic like along this route?” while simultaneously sketching a line on an interactive map display. The blue line in Figure 1.1 is an example of this kind of ink input. In this case, the semantic content of the user’s contribution to the dialog is distributed over two different modes. Prototype systems supporting this kind of interaction include QuickSet [Cohen et al. 1997], MATCH [Johnston et al. 2002a], SmartKom [Wahlster 2006], and Interact [Selfridge and Johnston 2015]. Deployed systems include Speak4It™ [Johnston and Ehlen

1. See also [Baltrušaitis et al. 2018] for a definition of multimodal fusion vs. other kinds of multimodal processing tasks.



**Figure 1.1** Ink input on map display with voice input: “what is the traffic like along this route?”

2010], Sketch-Thru-Plan [Cohen et al. 2015], which was deployed with the U.S. Marine Corp. and U.S. Army, and EVA the Enterprise Virtual Assistant [Openstream 2018]. See Cohen and Oviatt [2017] for additional details on various multimodal speech and pen systems. Other kinds of multimodal interaction involve combination of speech inputs with *gestures* made with the hands or various kinds of pointing devices towards a distant display [Bolt 1980, Corradini et al. 2002, Kaiser et al. 2003, Worsley and Johnston 2010, Cohen et al. 1999, Duncan et al. 1999, Sharma et al. 2003].

There are other types of fusion of multimodal inputs that we do not address in this chapter. Some of these involve fusion of signals at lower levels than the semantic level addressed here. For example, we do not address audio-visual speech recognition (e.g., [Potamianos et al. 2003, 2017]) where audio features are fused with features extracted from visual lip movement in order to perform robust recog-

nition of human speech. In audio-visual speech recognition, signals are integrated in order to predict the phonemes being produced by the speaker, while in the approaches we discuss here semantic content conveyed in gesture is combined with semantic content conveyed in speech in order to determine a combined meaning representation. Another related area of multimodal fusion research that we do not address concerns observation of multiple input streams either in a interactive system or in processing of recordings in order to classify various characteristics of individuals. For example, combination of visual cues with auditory features such as prosody in order to predict the emotional state of an individual or, for example, who is dominant in a meeting scenario [Aran and Gatica-Perez 2010], or the attitude of a reviewer in a recorded video clip of a product review [Chatterjee et al. 2015]. These approaches differ from multimodal input to interactive conversational systems in that they are applied either to recordings or, in the more interactive setting, they are used to passively identify characteristics of an individual rather than to fuse the components of a multimodal command or query that the user actively makes to the system as their contribution to move the dialog forward.

In Section 1.2, we outline the motivations for supporting multimodal commands to interactive systems. Section 1.3 surveys early approaches to multimodal integration. In Section 1.4 we explain and exemplify the evolution of more declarative approaches using *unification* to model multimodal integration. Section 1.5 describes an alternative approach in which multimodal integration is conceived of as part of language understanding and an implementation of this approach using finite-state methods. In Section 1.6, we discuss the use of event logic and visual *statecharts* to capture multimodal integration and *incremental interaction management*. Section 1.7 discusses approaches that model multimodal integration as part of *reference resolution* and other multimodal integration algorithms that extend beyond a single turn of input. In Section 1.8, we discuss a number of approaches that apply machine learning techniques to integration of multiple streams of user input. Section 1.9 concludes the chapter with a discussion comparing the different approaches and their relative advantages and disadvantages. A *Glossary* provides technical terms for the chapter and a list of *Focus Questions* follows Appendix 1.A.

## 1.2

### Motivations for Multimodal Input

The central motivation for supporting multimodal input to interactive systems is that it enables more natural interaction. Arguably, before the invention of the telephone, the majority of synchronous spoken human-human communication was face-to-face and inherently multimodal in nature. Interlocutors naturally combine

## Glossary

**An Attributed Relational Graph** (ARG) is a graph structure augmented with attributes presenting information regarding nodes in the graph. Attributed relational graphs have been used in computer vision and in work on multimodal reference resolution [Chai et al. 2004] where the graph represents a sequence of possible referents from either gestures or the discourse context and the attributes contain properties of the referents (e.g., color, size).

**An Augmented Transition Network** (ATN) is a *recursive transition network* augmented with a set of procedural operations and variables. ATNs have been applied to numerous tasks in natural language processing including parsing [Woods 1970].

**Chart parsing** is an algorithm for parsing strings with respect to a grammar. It uses a *dynamic programming* approach in which partial hypothesized results are stored in a tabular representation for re-use. This avoids re-computation of parses of subtrees shared among multiple parses and avoids the combinatorial explosion that can result from a brute force approach to application of rules in order to construct a parse.

**Concept accuracy** is a measure of the semantic correctness of input processed by an interactive system. This is in contrast to word accuracy and sentence accuracy that are measures of the accuracy of transcription of words by speech or handwriting recognizers. At the utterance level, the interpretation of an input is *concept accurate* if the meaning captured by the system corresponds to the user's intended meaning. Sub-turn measures of concept accuracy involving, for example, measures such as number of individual concepts successfully interpreted, are also used but need to be carefully defined as there is no generally accepted standard for assigning partial credit in determining the interpretation of an utterance.

**Co-reference Resolution or Reference Resolution** is the process of determining for a given word or phrase which other word or phrase, or entity in the application or environment that word or phrase refers to. Reference resolution may apply within an utterance, between utterances, or between a phrase and an entity presented visually in a graphical interface or present within the physical environment of the listener or in which an interactive system operates, the latter case is referred to as *multimodal reference resolution*.

**Deictic reference** is a subtype of reference where a word or phrase cannot be fully understood without additional contextual information, such as the identity of the speaker ("me") or their location ("here") or something they are pointing at ("that dog").

**Destructive unification** is an operation that combines information from two feature structures. In cases where a path in the feature structure does not unify the value from the second feature structure is taken.

**Glossary** (*continued*)

**A Directed Acyclic Graph (DAG)** is a graph consisting of a finite number of nodes and edges, where the edges are directed from one node to another, and such that there is no way to start at a particular node and loop back to the same node by following a sequence of directed edges (that is, there are no cycles in the graph). DAGs can be used to model many different types of information. They provide the formal mathematical underpinning of feature structures.

**Dynamic programming** is an approach to solving a complex computational problem by exploiting the structure of the problem in order to break it down into more simple sub-problems, storing their solutions, so they can be re-used without re-computation. In *chart parsing* for example, the parse trees for sub-constituents of a parse are stored and re-used in multiple different candidate parses of the input.

An **Edit Transducer** or **Edit Machine** is a *finite-state transducer* whose function is to transform a transducer (representing the input) into another transducer (representing the output) through a series of edit operations such as deletion, insertion, and substitution of symbols. These operations are captured using transitions in the edit transducer. For example, deletion of the symbol  $x$  can be represented using the transition  $x : \epsilon$  while insertion of  $x$  can be represented as  $\epsilon : x$ , where  $\epsilon$  is the epsilon transition. Substitution of  $x$  with  $y$  is represented using the transition  $x : y$ . The edit operation is achieved through *finite-state composition* of the input transducer  $I$  with an edit transducer  $E$ , annotated as  $I \circ E$ . Edit operations may be weighted and this captured by associating costs with arcs in the edit transducer.

**Feature structures** are a formal representation used in a variety of different grammatical formalisms including head-driven phrase structure grammars [Pollard and Sag 1994] and lexical functional grammar [Kaplan and Bresnan 1995]. A feature structure consists of a set of attribute-value pairs. The values may be atomic or feature structures themselves. Numerical indices in a feature structure indicate that the values of particular feature paths must be equal, in the sense that they must be unified (see *unification* below). Feature structures (sometimes called attribute-value matrices) have an underlying interpretation as *directed acyclic graphs* (DAGs) where the arcs are labeled with the features and the nodes contain the values. Use of numerical indices in feature structure notation corresponds to shared structure in the underlying directed acyclic graph.

A **finite-state automaton**, or **finite-state acceptor** is a finite state machine that operates over a single stream of input symbols. A finite state automaton defines a set of strings corresponding to all of the possible paths from the state space it defines. Formally, a finite-state automaton is a 5-tuple  $< S, I, B, F, T >$  where  $S$  is a finite set of states,  $I$  is a finite set of symbols,  $B$  is a finite set of beginning states, and  $F$  is a finite set of ending states.  $T$  is a finite set of transitions between states, each labeled with a symbol from  $I$  or the epsilon ( $\epsilon$ ) symbol. Finite-state automata may also be weighted with costs associated with arcs and states.

**Glossary (continued)**

**Finite-state composition** is an operation on a pair of finite-state transducers. Given as input a finite-state transducer  $T$  with the input symbol alphabet  $I$  and output symbol alphabet  $O$  and a finite-state transducer  $R$  with the input symbols alphabet  $O$  and output symbols  $O_2$ , there is a composition of the two finite-state transducers if there is a string  $y$  drawn from the alphabet  $O$  that  $T$  produces as output and  $R$  accepts as input. Finite-state composition of  $T$  and  $R$  is represented as  $T \circ R$ .

A **finite-state transducer** is a finite-state machine that operates over two streams, an input stream, and an output stream. Formally, a finite-state transducer is a 6-tuple  $\langle S, I, O, B, F, T \rangle$  where  $S$  is a finite set of states,  $I$  is a finite set of input symbols,  $O$  is a finite set of output symbols,  $B$  is a finite set of beginning states and  $F$  is a finite set of ending states.  $T$  is a finite set of transitions between states, each labeled with an input label from  $I$  or the epsilon ( $\epsilon$ ) symbol and an output label from  $O$  or the epsilon ( $\epsilon$ ) symbol. Finite-state transducers may also be weighted with costs associated with arcs and states. In weighted transducers the best path is the path with the lowest overall cost.

**Frames** are a data structure that have been used in a variety of artificial intelligence applications including natural language processing, computer vision, and knowledge representation and reasoning. A frame consists of a series of slot names and values that define a stereotypical object. In frame representations, typically the value of a frame slot may itself be another frame. The term *frame* was introduced by Minsky [1974]. Frames are similar to feature structures in that they consist of sets of attributes and values with the potential to embed further sets of attributes and values within values. Feature structures differ from frames in the capability of structure sharing and the underlying mathematical foundation of feature structures as directed acyclic graphs.

**Gestures** are inputs provided by a user through body movement, usually via hands and/or arms. In some instances, gestures are produced through contact with a display and involve touch or drawing on a screen using a finger or stylus (sometimes called touch gesture, pen gesture, or ink gesture). In other instances, gestures involve hand and arm movement in space and do not involve contact with a surface.

**Gesture recognition** is the process of automatically classifying and interpreting gestures made to an interactive system. The means of capturing and recognizing a gesture depends on the particular type of gesture. To recognize a drawing made on a display for example the trace made by a finger or stylus can be captured directly and the resulting series of *strokes* can be fed into a classification algorithm. For other types of gesture without direct physical contact with a touch-sensitive surface, video cameras or other types of sensors, some of which are worn by the user, can be used to track the location of the hand and arm.

**Glossary** (*continued*)

**Incremental processing** refers to the ability of a system to provide feedback or interpretation on the input that it is receiving before the input has completed. One example is incremental speech recognition, where a visual interface provides feedback on recognized words as the user speaks.

A **lattice** is a *finite-state automaton* that captures a range of possible different interpretations of user input. For example, a lattice from a speech recognizer will represent different possible speech recognition results for a particular input. Each path through the lattice from a designated start state to one of more end states is a different potential recognition of the input.

**Multimodal grammar.** The most common use of the term *grammar* is a formal mechanism for accounting the possible combinations of words. A multimodal grammar in contrast is a formal mechanism that accounts for the possible combinations of words and symbols from other modalities such as gesture.

A **multimodal interaction manager** is a system component in an interactive multimodal system that manages the flow of interaction (performs *multimodal interaction management*). A multimodal interaction manager may operate at the turn level, within the turn (*incremental interaction management*), or both. For example, a multimodal interaction manager may be responsible for application of temporal constraints for consideration of speech and gesture inputs as part of a single turn. An incremental interaction manager may trigger incremental feedback to the user while they are still speaking, such as highlighting an object that has been mentioned by voice or pointed at using hand gestures.

**Multimodal integration** or **multimodal fusion** is the process of combining content from different modes of input in order to determine their combined meaning.

An **N-best list** is a list of length  $N$  of different alternative hypotheses produced by a component of a system such as a speech recognizer, gesture recognizer, or natural language understanding component. Each member of an N-best list represents a different possible result. Generally, they are associated with scores or probabilities assigned by the component generating the result, and the results appear in order with the highest score or highest probability result first. For example, an N-best list from the speech recognizer would represent different possible decodings of the speech signal into strings of words.

Overlay of feature structures is a type of *unification* operation similar to *destructive unification*. In overlay and destructive unification, if there are clashing values as the two feature structures A, B are combined the values in B appear in the result.

**Glossary (continued)**

**A Recursive Transition Network** (RTN) is a graphical construct that can be used to capture the legal strings of a language. An RTN consists of a series of nodes, transitions between those nodes, a defined start state and a defined set of end states. The transitions are labeled with symbols from the language and paths through the network from the start state to an end state enumerate the legal strings of the language. Critically in an RTN, the label on a transition may also be a reference to another RTN, enabling re-use of sub-networks and recursive description of the language. RTNs are generally implemented as pushdown automata and have equivalent computational power to context-free grammars.

**Statecharts** first presented by [Harel \[1987\]](#), extend the concept of state machines and state diagrams with concepts supporting hierarchy, concurrency, and communication, enabling the description of complex systems including the operation of devices, human computer interfaces, and communication protocols.

**Strokes** are the individual components of a gesture made by a user. For example, in order to make an “X” gesture the user might make two strokes one from the top left to the bottom right, then raise the finger or stylus and make a second stroke from the top right to the bottom left.

**A subcategorization frame** is a construct, generally a list, which captures a series of phrases that a word or phrase needs to combine with. For example, the verb “give” in English subcategorizes for a noun phrase (NP) and a prepositional phrase (PP), e.g., “give [the book]NP [to john]PP.” In the grammatical theory Head-driven Phrase Structure Grammar [[Pollard and Sag 1994](#)], a subcategorization frame is a list structure represented as feature structure which captures the words or phrases that a word or phrase needs to combine with. In [Johnston \[1998\]](#), the concept of a subcategorization frame is extended to multimodal inputs and a spoken phrase such as “put that there” is said to “subcategorize” for two gestures, the first corresponding to “that” and the second corresponding to “there” in the spoken input.

**Typed feature structures** are a feature structure representation where each feature structure is assigned a type within a hierarchical structure. When feature structures are unified, the types are required to be in an ancestor relation in the hierarchy and the result is the more specific type. See [Carpenter \[1992\]](#) for more details.

**Unification** of feature structures is an operation which takes two feature structures as input and determines whether or not they are compatible with each other, that is they unify, and if they do returns a feature structure containing the combination of information from the two input feature structures. This operation corresponds to graph unification of the two underlying directed acyclic graphs that the feature structures represent. Feature structure unification is a generalization of term unification, an operation used in logic programming.

spoken contributions with pointing and other types of gestures [McNeill 1992]. In addition to enabling unimodal spoken communication over great distances, communications technologies such as email and text messaging have increased the amount of asynchronous unimodal communication. Interestingly now, however with the prevalence of smartphones, and adoption of video calling, advances in communication technology are enabling multimodal human-human communication at a distance. The ongoing convergence of the Web with telephony, driven by technologies such as voice over IP, prevalent access to high-speed mobile data networks, and sophisticated handheld computers and smartphones, now enables widespread deployment of multimodal interfaces which combine graphical user interfaces with natural modalities such as speech, pen, and touch. In addition to mobile devices, use cases such as the connected car, control of smart devices in the home (and more broadly the Internet of Things (IoT), wearable computing, and interaction with robots are all inherently multimodal contexts for interaction.

Speech is an extremely powerful mechanism for interaction. In comparison with graphical interfaces, speech provides a single point of entry for the user to provide a large, in fact infinite, range of different commands, and avoids the need for complex on-screen menus. Speech is also extremely powerful in that it allows for reference to objects not currently visible on the display and provides a natural mechanism for statement of complex constraints [Cohen 1992]. However, speech is not the best modality for everything. Particular types of content are particularly well matched to the affordances of particular input modes. In the words of Rudnicky and Hauptman [1992], “certain tasks and functions cry out for particular modalities.” Drawings or gestures are particularly well suited to specification of complex spatial information, such as routes on a map, as in (Figure 1.1). As speech capabilities are added to devices supporting a graphical user interface, such as smartphones, tablets, and interactive wall displays the user should not have to stop using non-verbal modes such as touch, drawing, and hand gesture—for optimal performance these modes should be seamlessly integrated with speech input. Another key motivation for multimodality is that through having support for more than one mode, multimodal interfaces allow users to switch from one mode to another in order to avoid errors [Oviatt and VanGent 1996], or to adapt to changes in the physical or social environment (for example, switching from voice to pen for entering sensitive information such as credit card numbers).

There is also a significant, and growing, body of experimental work providing empirical evidence that compared to their unimodal counterparts, multimodal interfaces have significant advantages both in terms of user preference and task performance [Hauptmann 1989, Nishimoto et al. 1995, Cohen et al. 1998, Oviatt

1999, Cohen et al. 2015]. Specifically from the point of view of system performance, several studies show that multimodal interfaces can provide increased robustness through mutual compensation for errors in the individual modes [Oviatt 1999, Johnston and Bangalore 2005]. In a 3D multimodal interaction setting, [Kaiser et al. 2003] report reductions in error rate of as much as 67% through mutual disambiguation.

## 1.3

### Early Approaches to Multimodal Fusion

Systems capable of accepting multimodal commands combining spoken or typed natural language and gestures of various kinds have existed since the early 1980s. Bolt [1980] describes a system, “Put-that-there” which combines hand gestures captured with a magnetic field sensor with inputs to a speech recognizer. Neal and Shapiro [1991] describe a complex multimodal system for interaction with maps that supports multimodal integration and multimedia output planning, although speech was only used for output not as an input modality. Allgayer et al. [1989] presents a multimodal interface to an expert system that assists users working on tax forms.

A common feature of these early systems is that interaction is primarily driven by the speech or language modality, with gesture as a secondary dependent modality. Combination of information from the gesture input into the combined multimodal meaning is triggered by the appearance of expressions in the speech input whose reference needs to be resolved, such as definite and deictic noun phrases (e.g., “this ship,” “the blue square”). In these approaches, multimodal integration is essentially a procedural add-on to an existing speech or text understanding system. In the case of Neal and Shapiro [1991], it is achieved through the addition of operations accessing gesture to an *Augmented Transition Network* (ATN) for parsing and interpreting word sequences, while in the case of Allgayer et al. [1989] gestures are accessed through extensions to a unification-based *chart parser*. The Shoptalk system [Cohen et al. 1989] takes a similar approach but using definite clause grammars. In the case of “Put-that-there” [Bolt 1980], procedural subroutines for interpreting the speech recognizer output access the data from the magnetic sensor to determine the location or object referred to by a deictic pronoun. Latoschik [2002] also develops an approach using an ATN for capture and integration of multimodal utterances in a virtual reality environment. For their crisis management application, Sharma et al. [2003] describe a speech-driven approach in which the appearance of particular key phrases or semantic categories in the speech input stream triggers a

temporally constrained search in a series of events emitted by a free hand gesture recognition system.

None of these approaches develop a common meaning representation across modes or a declarative mechanism for capturing multimodal integration strategies. Work by [Koons et al. \[1993\]](#) represented an important step in this direction. They developed prototype systems supporting simultaneous input from hand gestures, gaze, and speech. One of these systems, which supported interaction with a blocks world, supports iconic and pantomimic gestures in addition to deictic pointing gestures. A *frame* representation, common across modes, is used to represent the contribution of each modality. Frames are built for the natural language input and for gaze and gestures, capturing, e.g., fixations, blinks, and posture and orientation of the hand for hand gesture. As pointed out by [André \[2002\]](#), while [Koons et al. \[1993\]](#) present a uniform representation formalism, they do not have a declarative process for modality integration, rather they have algorithms that apply directly to the frames. [Waibel et al. \[1996\]](#) develop a related frame-merging approach and applied it to integration of speech and pen input. In the next section, we discuss unification-based approaches to multimodal fusion in which declarative approaches to capturing the process of multimodal integration were developed.

## 1.4

### Unification-based Multimodal Fusion and Related Approaches

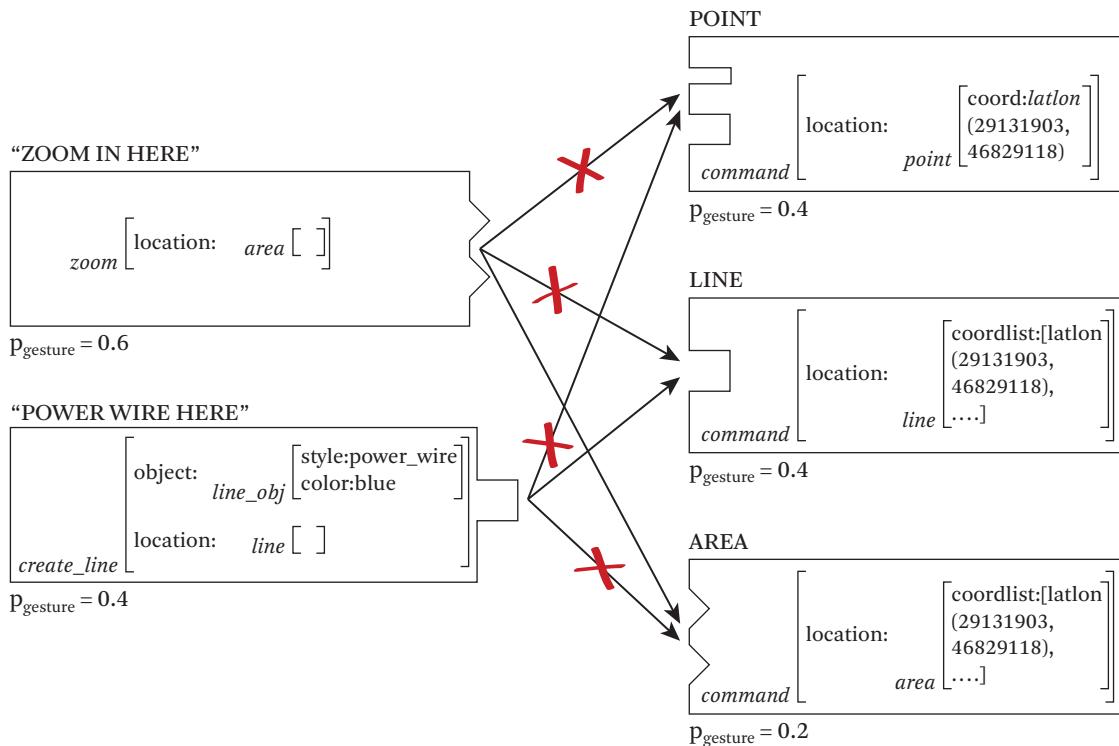
A number of different approaches to multimodal fusion draw on unification and related graph-theoretic operations to provide a more declarative model of integration of content from different modes. Unification is an operation that determines whether two pieces of partial information are consistent, and if they are, combines them into a single result. As such it is well suited to capturing constraints on combination of semantic content from different modes. Unification has been used extensively in linguistic models of syntactic and semantic phenomena [[Pollard and Sag 1994](#)] and has also been applied in work on knowledge representation [[Brachman et al. 1991](#)].

Unification is frequently described as an operation applying to *feature structure* representations [[Carpenter 1992](#)]. Feature structures are hierarchical data structures each consisting of a series of attributes and values and allow for the embedding of feature structures as values. Indices in feature structure representations indicate the sharing of values. These feature structure representations represent *directed acyclic graphs* (DAGs) and the indices correspond to structure sharing in the graph. For example, structure sharing in a graph representing the combination of a subject noun phrase and a verb phrase in a sentence might be used to capture

agreement between the *number*: feature of the subject and the *number*: feature of the verb (subject-verb agreement in English).

[Johnston et al. \[1997\]](#) present the first application of unification to multimodal input processing. In that approach, multimodal integration is modeled as unification of *typed feature structures* [[Carpenter 1992](#)] representing partial meanings assigned to speech and gesture inputs. Typed feature structures are a feature structure representation where each feature structure is assigned a type within a hierarchical structure of types. When feature structures are unified, the types are required to be in an ancestor relation in the hierarchy and the result is the more specific type. See [Carpenter \[1992\]](#) for more details on typed feature structures, unification, and the mathematical foundations of these operations. Speech or gesture input that partially specifies a command is represented as an underspecified feature structure in which certain features are not instantiated. The application domain addressed is Quickset [[Cohen et al. 1997](#)], a speech/pen interface to a map display on a tablet for interaction with distributed interactive simulations—the predecessor to Sketch-Thru-Plan (STP) [[Cohen et al. 2015](#)]. Multimodal commands are used to manipulate the map (e.g., “zoom in here”) and to position symbols representing the location of entities and features (e.g., “power line along here”) on the map. For example, a spoken command such as “power wire here” specifies an action and captures the fact that the location of the command should be a route line (as opposed to an area or circle or point). An accompanying gesture drawn on a display does not contribute an action but does contribute a location of a particular type, along with the specific content e.g., a series of geographic coordinates. Figure 1.2 provides an illustrative example of multimodal integration modeled as unification. In this framework, both speech recognition and gesture recognition components yield *N-best lists* of results. A natural language parser builds a feature structure representation for each of the potential speech recognition results and a gesture interpretation module builds a feature structure representation for a sketch made on the display. Figure 1.2 shows the feature structures associated with speech on the left and those associated with gesture on the right for the “power wire here” example. Feature structure types are shown here as subscripts on the left of each feature structure. The puzzle piece graphic around each structure visualizes the compatibility of the respective feature structures with respect to the unification operation.

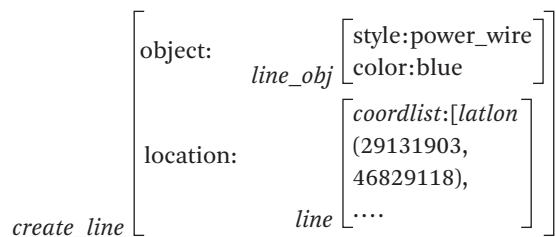
A multimodal integration algorithm attempts to unify each of the cartesian product of the speech and gesture N-best lists of feature structures. In this case, only two combinations are possible: “zoom in here” with AREA and “power wire here” with LINE. The semantic constraints on integration of speech and gesture are captured declaratively in the typed feature structure representation and a general



**Figure 1.2** Multimodal integration through unification of typed feature structures.

procedure of feature structure unification determines which combinations are possible. In cases where more than one combination is possible, each multimodal combination is assigned a probability using a weighted combination of normalized speech and gesture recognition scores ( $p_{multimodal} = \lambda p_{speech} * (1 - \lambda)p_{gesture}$ ).  $\lambda$  here is a parameter between 0 and 1 controlling the influence of speech vs. gesture on the combined score of multimodal results. In this example, the system selects the combination of “power wire here” with the line gesture and resulting feature structure representing the combined semantics for the multimodal utterance is as in Figure 1.3.

This process of multimodal integration results in mutual compensation for errors in the individual modalities. In this case, both the top scoring result for speech and the top scoring result for gesture were incorrect: “zoom in here” and *POINT*, but through the process of multimodal fusion the correct interpretation combining elements lower down in each N-best list was selected: “power line here”



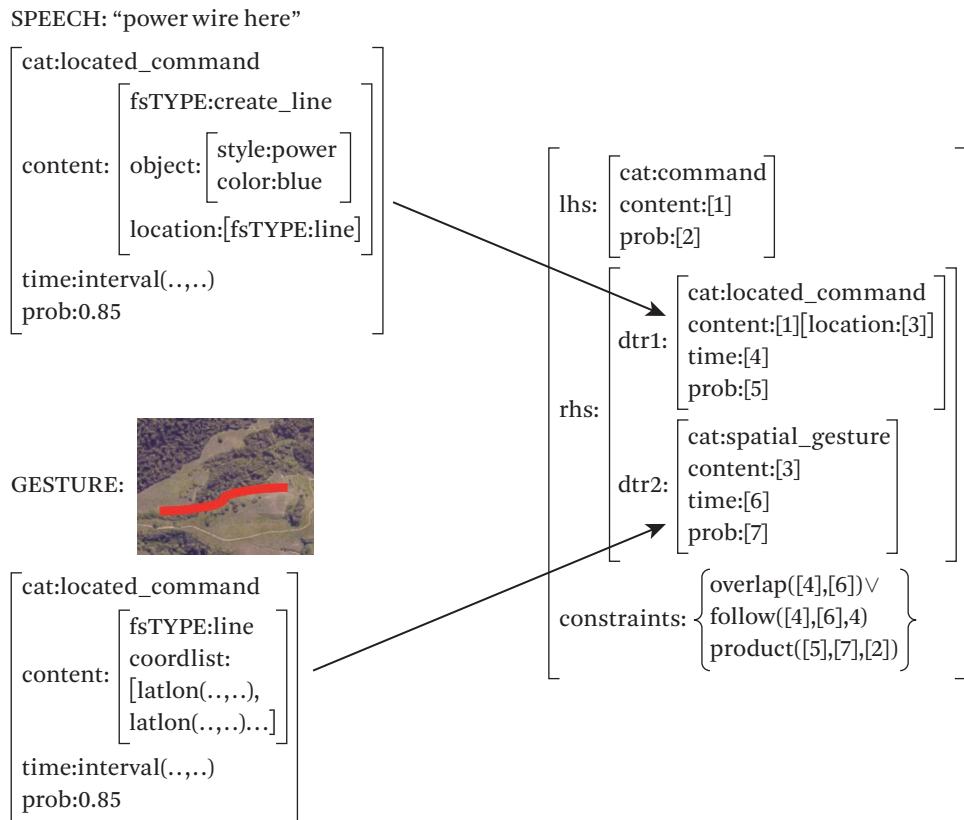
**Figure 1.3** Combined feature structure semantics for multimodal input “power wire here” with line gesture.

and *LINE*. Oviatt [1999] presents a detailed empirical evaluation of this unification-based approach to multimodal integration in the context of the QuickSet pen/voice system. In that study, one in eight of the commands that were recognized correctly by the system was successful because of multimodal disambiguation. Focusing specifically on the spoken language understanding (speech+NLU), there was a 41% relative reduction in error rate resulting from the use of a unification-based multimodal architecture.

In addition to mutual compensation, this approach also has the advantages of moving away from being speech-driven, putting both of the modalities on an equal footing. The semantic constraints on multimodal fusion, such as route following commands going with lines, are captured declaratively using the feature structure representation and unification. However, the integration algorithm itself remains procedural, with temporal and any other constraints on multimodal fusion being captured directly in the code of the multimodal integration algorithm that applies the unification operation. Another limiting factor is that the approach only supports combination of a single spoken command with a single gesture and does not scale to gesture only commands involving combination and visual parsing of multiple gestures. In Wu et al. [1999], [2002], additional mechanisms are added to the unification-based multimodal integration mechanism in order filter out combinations of features structures that are never able to unify and a trainable approach to combination of probabilities associated with each mode is developed (See Section 1.8 for more details). In Section 1.7 we discuss application of a related operation to unification, *overlay* [Alexandersson and Becker 2001] in handling discourse phenomena in multimodal systems.

Later work extended the unification-based approach to multimodal input, drawing on some ideas from information-based syntax and semantics [Pollard and Sag 1994] in order to provide a declarative representation not just of the semantic con-

straints on fusion but of the multimodal integration strategy itself including any spatial and temporal constraints [Johnston 1998]. In this approach, the multimodal integration strategy is modeled as application of a series of *multimodal grammar* rules encoded as typed feature structures. Returning to our example, for combination of a command that requires a location (like “power wire here” above) with a gesture input such as a line, a general purpose multimodal grammar rule would be used, represented as a feature structure (Figure 1.4). The *lhs:* feature indicates the left-hand side of the rule and *rhs:* the right-hand side. The two categories on the right-hand side are in daughter features *dtr1:* and *dtr2:*. The feature structure encodes a combination rule of the form  $LHS \rightarrow DTR1\ DTR2$ , where each constituent being combined is represented as a feature structure. The semantic content that



**Figure 1.4** Example of a multimodal feature structure rule (on the right) for combination of speech and gesture being applied to “power wire here” with line gesture (on the left).

was unified before in the unification-based multimodal integration approach above is now held in the feature *content*::. The feature *fsTYPE*: is used here to indicate the overall type of the feature structure in which it appears. The *cat*:: feature provides a high level categorization of the type of content and is used as a filter to limit application of the feature structure rules. This representation captures the fact that input of the category *located\_command* can be combined with one of the category *spatial\_gesture*, so long as the *location*:: feature of the *located\_command* unifies with the *content*:: of the spatial gesture and the time intervals associated with speech and gesture overlap or the speech follows within 4s of the gesture. This temporal constraint is captured using functional constraints. The constraint *overlap*([4],[6]) is true if the intervals shared through co-indices [4] and [6], the spoken command and the gesture, that is, overlap; that is if the end of one is later than the start of the other. The constraint *follow*([4],[6],4) is true if the time interval of the spoken command ([4]), follows within 4 s of the time interval of the gesture ([6]); that is, the time between the end of the gesture and beginning of the speech is no more than 4 seconds.<sup>2</sup> Structure-sharing [1] is used to pull the combined meaning into the resulting *content*:: value, and another function *product*([5],[7],[2]) is used to capture the combination of the probabilities associated with the two combining interpretations. In this approach, as in unification-based multimodal integration above, the outputs of speech understanding and gesture interpretation are captured as typed feature structures representing partial interpretations. Essentially these chunks of meaning serve as the “words” combined in a second level of multimodal parsing.

In order to support combination of the multimodal inputs a new formulation of chart parsing was needed. Generally, parsing algorithms such as chart parsing assume that the input is a linear sequence of discrete tokens. Multimodal parsing is not subject to the same constraints of combination of items in linear order, since speech and gesture may overlap in time, and gesture interpretations may overlap in space. To address this, [Johnston \[1998\]](#) developed a multi-dimensional extension of the classic CKY chart parsing algorithm [[Cocke and Schwartz 1970](#), [Kasami 1965](#), [Younger 1967](#)] which removes the assumption that the combining constituents are in linear order and adjacent (Equation 1.1).

$$mchart(X) = \cup mchart(Y) * mchart(Z) \quad (1.1)$$

where  $X = Y \cup Z$ ,  $Y \cap Z = \emptyset$ ,  $Z \neq \emptyset$ .

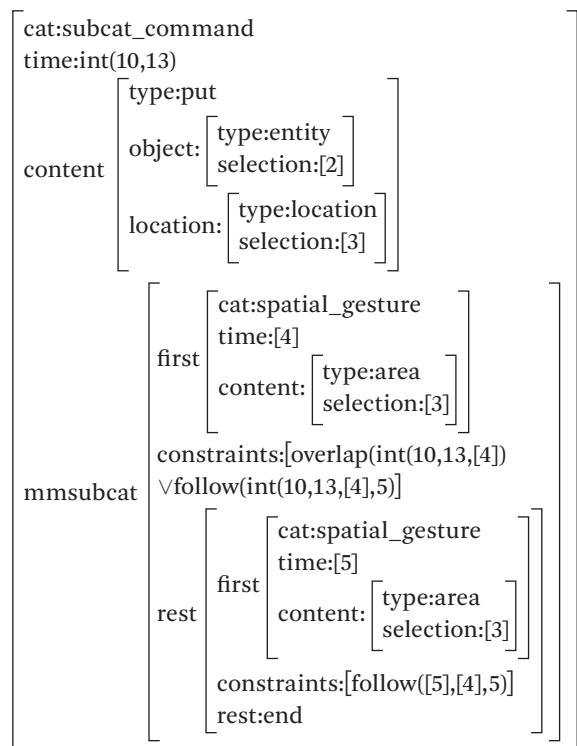
---

2. These temporal constraints are based on an empirical study of the synchronization of inputs in multimodal interaction detailed in [Oviatt et al. \[1997b\]](#).

$X$ ,  $Y$ , and  $Z$  here are edges representing elements of input and their combinations in the parse chart. The  $*$  is the operation for the combination of edges (in this case application of multimodal feature structure rules. The  $\cup$  represents the union of all of the permitted combinations of the input elements  $Y$  and  $Z$  with respect to the multimodal combination rules represented by  $*$ . The constraint that remains is that the combining constituents should not intersect, that is you cannot use an input more than once in the derivation of the multimodal parse. This constraint is important in that it limits the number of possible constituent combinations that need to be considered by the parser. Note, however, that some inputs to interactive systems may be multi-functional in nature, for example a head nod intuitively can signal both understanding and agreement. If these separate functions were to be handled within this multimodal parsing mechanism each would be added as a separate constituent in the chart. Alternatively, the combined interpretation could be added as one of the N-best interpretations of the head nod.

The unification-based approach to multimodal parsing supports combination of speech with multiple gestures and also extends naturally to supporting parsing of complex unimodal gesture commands. These are handled using feature structure schemata for combination of multiple gestures and spatial and temporal constraints on their combination can be directly encoded in the combination schemata. There is a body of related work specifically extending parsing to handling two dimensional graphical input, in the context of parsing flowcharts, equations, visual programming languages, and so on, using a number of different grammatical formalisms [Lakin 1986, Helm et al. 1991, Crimi et al. 1991]. Wittenburg et al. [1991] apply unification-based grammars to parsing two dimensional graphical languages and employ a bottom-up Earley style tabular parser [Earley 1970].

In this unification-based approach to multimodal parsing, combination of speech and gesture elements can be captured by rule schemata capturing general multimodal integration patterns, or alternatively the speech and gesture understanding mechanisms can be enriched in order to capture more information about what other content an input needs to be combined with, which then drives application of a smaller set of general rule schemata. These two approaches can be compared to rule-based approaches to syntax, vs. more lexicalized approaches such as categorial grammar [Steedman 1996] and head-driven phrase structure grammar [Pollard and Sag 1994] that capture combinatory information directly in the lexical representation. For example, Figure 1.5 shows a representation for the spoken input “put that there” which captures, in a multimodal subcategorization frame (*mmsubcat:*) that this spoken command needs to be combined with two area gestures which provide the entity to be moved and the location to move it



**Figure 1.5** Multimodal subcategorization for “put that there.”

to. Functional constraints embedded in the *subcategorization frame* capture the temporal order requirement on the two gestures and the requirement of overlap or follow between the time interval of the speech and the time interval of the gesture. The notation  $int(t_1, t_2)$  is used to represent a time interval starting at  $t_1$  and running to  $t_2$ . Structure-sharing in the graph (indices, e.g., [2] [3]) capture how content from the two gestures contributes to the overall combined meaning.

The unification-based approaches separate the concerns of spoken parsing and understanding from multimodal integration. The unification-based integration and unification-based parsing approaches model multimodal fusion after spoken language understanding (and gesture understanding), a case of late fusion involving a second level of multimodal parsing and understanding. This modularity and resulting separation of concerns has advantages for system development. For example, different types of natural language processing components can be coupled with the multimodal parser. Also, the same multimodal parser can be used with natural

language components supporting different languages. The unification-based parsing approach with multimodal subcategorization is effective in enabling the combination of speech with sequences of gestures. However, as more complex spoken inputs are considered, for example cases with numerals and conjunction, the task of building the multimodal subcategorization from a spoken utterance becomes increasingly unwieldy [Johnston 2000]. One approach to address this problem is to remove the separation and combine spoken and multimodal parsing into a single process which applies directly to words and symbols representing the other modes. This is the topic of the next section.

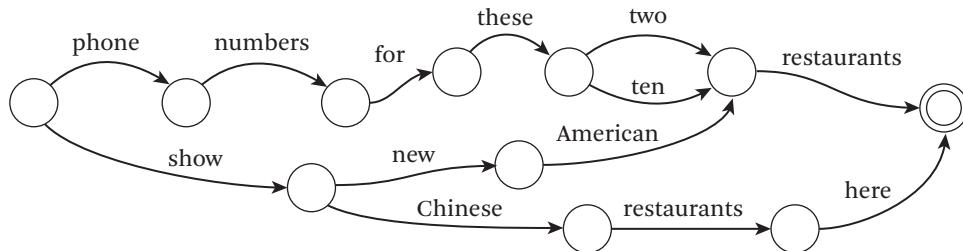
## 1.5

### Multimodal Grammars and Finite-state Approaches

The approaches we have considered so far explicitly separate the task of language understanding from that of multimodal integration. For example, in the unification-based approaches spoken language understanding and gesture understanding components apply separately to generate sets of candidate meaning representations for speech and gesture inputs, and these are then combined by a multimodal integration component. An alternative approach views understanding of the individual modes and their integration as a single holistic process of multimodal language understanding [Johnston and Bangalore 2001, Johnston and Bangalore 2005, Bangalore and Johnston 2009].

In a “one-stage” approach, the multimodal integration component applies directly to streams of input containing token symbols from each mode, and a multimodal grammar or model processes these inputs, integrates them, and creates a combined meaning representation for the multimodal input. The tokens of input for spoken or written modalities are somewhat more straightforward to establish since we have fairly clear conventions in most languages for tokenization of spoken or written input into sequences of words. For example, a spoken input to a multimodal local search application (see, for example, [Johnston et al. 2002b]) might consist of the words “show chinese restaurants here.” Since for natural inputs such as speech, a recognition component may be uncertain of the specific word sequence, various representations are used to capture multiple alternatives, including ordered lists of alternatives (N-best lists) or more compact representations such as *lattices* (Figure 1.6) which capture as a *finite-state automaton* a range of possible word strings that may have been expressed in the spoken utterance.

In order to build a model that applies directly to multimodal input streams, the first challenge is to decide how to tokenize and represent input for non-verbal modalities such as touch, pen drawings, or 3D gesture input. Unlike speech and



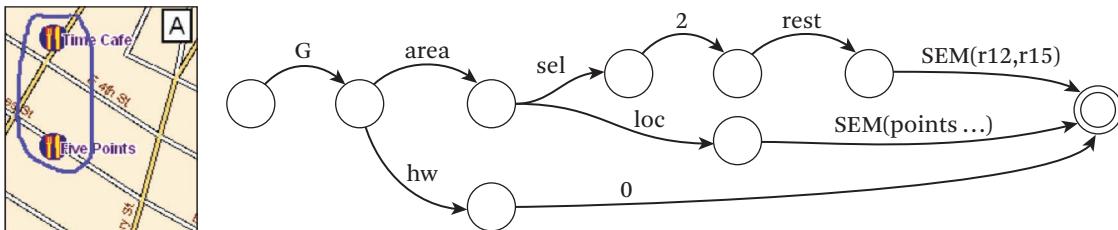
**Figure 1.6** Lattice capturing possible recognitions of speech input.

written/typed text, there are no established, generally accepted schemes for tokenization and representation of these modalities.<sup>3</sup> Generally, the developer of a multimodal system will have to come up with a representation scheme for the non-verbal mode.<sup>4</sup> In order to develop multimodal integration capabilities for the MATCH system, Johnston et al. [2002a] develop a token language for representing possible contributions of electronic ink (see Bangalore and Johnston [2009] for more details). This representation results from the application of both handwriting recognition and gesture classification to a series of *strokes* of ink made by the user. As a concrete example, consider a case in which the user draws ink on a map around icons representing the locations of businesses (Figure 1.7). Possible interpretations of this ink input are captured using a series of token symbols which abstract out properties resulting from handwriting recognition and gesture classification, and consideration of the relation between the gesture and the map itself, entities which may be being selected and so on.

For this particular case, on the surface a fairly simple gesture, the representation in Figure 1.7 captures the fact that this is an area gesture (*G area*) that may either be a selection (*sel*) or a location (*loc*). In the case of a selection, it is a selection of two (2) restaurants (*rest*). The same ink, processed by handwriting recognition (*hw*), could be input of a zero (0). The *SEM( . . . )* tokens are used to contain specific content such as a sequences of points or IDs representing specific selected entities. One can imagine the design of similar token languages for other types of input. For simple touch input to graphical user interfaces elements (such as buttons),

3. Of course there are schemes for sign languages such as American Sign Language, but there are not widely agreed upon and established languages for free-form arm gesture or pen sketches.

4. There are representation schemes for electronic ink, e.g., InkML [Watt et al. 2011] but they do not provide a standard tokenization and representation, rather it provides an XML representation for capturing the signal, e.g., a sequence of strokes of electronic ink.

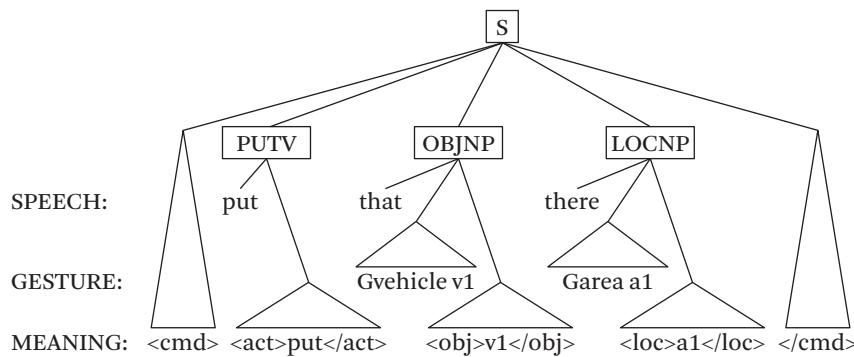


**Figure 1.7** Lattice representing possible interpretations of ink gesture input.

the token language could capture the type of entity touched and an identifier, e.g., *Gvehicle id123 Gwarehouse id456*. Note that the combination of strokes into more complex symbols (e.g., multiple lines combined into a square, head, and line of arrow, military symbology) are not directly represented in the gesture lattice at this level, rather strokes are combined by gesture recognition algorithms and the results are used to populate the gesture lattice representation.

Having established representation schemes for the different modes, the next challenge is how to encode a grammar that can apply to multiple streams of input. Grammar representations (and language models) for speech and text generally capture the possible (or likely) sequences of tokens, generally words, in the case of those modes. A context-free grammar for speech or text consists of a series of production rules, e.g., ( $S \rightarrow NP\ VP$ ,  $NP \rightarrow Det\ N$ ) enumerating possible expansions of non-terminal symbols into other non-terminal symbols and eventually into terminal symbols (words), e.g., ( $Det \rightarrow$  “the”,  $N \rightarrow$  “dog”). In order to extend grammar representations to address multimodal language, [Johnston and Bangalore \[2001\], \[2005\]](#) take the approach of enriching the representation with complex terminal symbols which contain triples of speech (words), gesture (gesture symbols, e.g., *area*, *sel*, *loc*), and meaning symbols (e.g., XML representation of the combined meaning). Since not all input will contribute to all three of these symbol streams (speech, gesture, meaning), and to facilitate alignment, these components of the terminal symbol may also contain the reserved symbol epsilon  $\epsilon$  indicating the absence of content for a particular stream.

The example in Figure 1.8 illustrates the concept of a multimodal grammar and its application to multiple input streams (SPEECH, GESTURE) to create a combined output stream (MEANING). In this language, the terminals such as  $\epsilon:SEM:SEM$  represent a copying operation where content elements appearing on the gesture input stream are copied into the meaning output stream. In the forthcoming finite-state implementation of this approach, this is captured using a sequence of



**Figure 1.8** Sample multimodal grammar for “put that there.”

**finite-state composition** operations. In this example, the symbol  $a1$  stands for the set of points indicated by the area gesture, and  $v1$  is a symbol indicating the identity of a vehicle pointed at by the user. Note that in this multimodal grammar approach the grammar does not encode detailed temporal constraints. Instead, the grammar captures the sequence of the words and the sequence of gestures and through parsing creates an alignment of speech and gesture. [Bangalore and Johnston \[2009\]](#) found this mechanism to be sufficient for support of a complex pen/voice system, though for more freeform inputs such as 3D hand/arm gesture, where segmentation of gesture events is significantly harder, order information without more fine-grained temporal constraints may likely not be sufficient.

A multimodal grammar in the form of a set of production rules could be applied directly to multiple input streams using an appropriate parsing strategy, accounting, for the complex terminals, with, for example, an extension of chart parsing to support alignment of multiple token streams. [Johnston and Bangalore \[2001\], \[2005\]](#) take a different approach choosing instead to compile the multimodal grammar into a cascade of finite-state transducers. This has the important advantage that it enables direct application of the resulting multimodal understanding component to lattice representations of speech and gesture (e.g., Figures 1.6 and 1.7). A multimodal parser running directly from rules would have to consider all possible permutations of speech paths with gesture paths.

Finite-state methods have a long history in speech and language processing including applications to speech recognition [[Riccardi et al. 1996](#), [Pereira and Riley 1997](#)], phonology [[Kartunnen 1991](#), [Kaplan and Kay 1994](#)], morphology [[Koskenniemi 1984](#)], chunking [[Abney 1991](#), [Joshi and Hopely 1997](#)], parsing [[Roche 1999](#)], and machine translation [[Bangalore and Riccardi 2002](#)] In the case

of multimodal understanding addressed by [Johnston and Bangalore \[2001\]](#), there are two input streams (speech and gesture) and a third output stream (meaning). In principle, the multimodal grammar could be compiled directly into a three tape finite-state device [[Rosenberg 1967](#)], using standard techniques for finite-state approximation of context-free grammars [[Nederhof 1997](#)]. This would be a three-tape finite state automaton:  $W:G:M$ , words/gesture/meaning) corresponding to the structure of the terminal symbols in the multimodal grammar representation. However, in practice, libraries and toolkits for finite-state processing only support finite-state machines and finite-state transducers [[Mohri et al. 1998](#), [Hetherington 2004](#), [Allauzen et al. 2007](#)]. [Johnston and Bangalore \[2001\], \[2005\]](#) implement multimodal understanding through composition of a cascade of finite-state transducers compiled from a multimodal grammar. The first of these  $R : G \rightarrow W$  is composed with the input speech and gesture lattices and performs an alignment and mapping of speech and gesture symbols into a combined alphabet of multimodal tokens. The second of these  $T : G \times W \rightarrow M$  maps from those multimodal tokens into their combined meaning representation.

One of the characteristics of gestures made on a map display is that they convey specific semantic content such as map coordinates or the identities of selected items, such as the restaurants in the example in Figure 1.7. This content needs to be incorporated into the resulting combined semantic representation for the multimodal input. With feature structure grammars, this can be achieved through structure sharing in the underlying graph or by some kind of copying operation. When modeling multimodal integration as finite-state composition it is not practical and unwieldy to encode all of the possible different specific items, such as coordinates on a map or all IDs of entities on a complex display, in the finite-state transducers used to model integration. [Bangalore and Johnston \[2009\]](#) address this in the finite-state approach by representing gesture as a transducer  $I:G'$  which has the gesture symbols with specific content on the input tape and gesture symbols with specific content replaced with a reserved symbol *SEM* on the output tape. A projection of the output side is used in the finite-state compositions, then when the meaning representation is read off from the final transducer, the  $G$  symbols associated with that meaning are used as a signature to select out the appropriate specific content from the  $I:G'$  transducer. A more detailed explanation of this abstraction mechanism and the finite-state processing cascade, along with an illustrative worked example, are available in Appendix 1.A.

[Bangalore and Johnston \[2004\], \[2009\]](#) describe an extension to the finite-state approach that increases the robustness of the system to unexpected inputs. The multimodal grammar approach described above allows for the expression of

complex multimodal commands. If, however, the model used for speech recognition is a deterministic model taken from the projection of the output  $W$  tape of the  $G:W$  machine, the resulting system can be brittle with respect to unexpected inputs. If instead a statistical language model (SLM) is used for recognition, speech recognition performance improves but multimodal understanding will fail on utterances that are recognized but fall outside of those enumerated in the multimodal grammar. To address this challenge, they explore a range of different techniques including application of a statistical classifier to multimodal input, learning a machine translation from SLM output to the inputs accepted by the multimodal grammar, and an edit-based approach in which an *edit transducer* is composed with the SLM output before application of the multimodal language processing cascade described above. The approaches were evaluated on realistic multimodal data collected from experimental subjects using the MATCH pen/voice system [Johnston et al. 2002a]. The best performing approach in terms of *concept accuracy* was the edit transducer approach. This approach achieved concept accuracy of 63.2% on SLM speech recognition output with 75% word accuracy. In comparison, direct application of the finite-state understanding without the edit transducer yielded only 38.9% concept accuracy. In contrast, using a grammar-based language model with the grammar-based finite-state understanding yielded 50.7% concept accuracy. The combination of improved language modeling (the statistical language model) with the edit-based approach provided a 12.5% absolute improvement in concept accuracy.

Since the speech and gesture inputs to the multimodal integration mechanism here are represented as lattices, the approach also allows for compensation for errors; that is, it allows for information from one modality to overcome errors in the other. For example, if the top scoring path in the speech lattice is wrong but it is semantically incompatible with the gesture, then a lower scoring but correct speech path ( $W$ ) may be selected through the finite-state composition with ( $G$ ) as they are composed with the  $R(G:W)$  transducer. The approach in fact allows for mutual compensation for errors; that is, even if both the speech and gesture top paths are wrong, both errors may be corrected through multimodal fusion. For a multimodal directory assistance task, Johnston and Bangalore [2005] found the tight coupling of speech and gesture processing through finite-state multimodal integration resulted in a 2% absolute improvement in word accuracy which corresponded to a 23% relative reduction in sentence level ASR error. In relation to the results from Oviatt [1999], which demonstrated the potential for mutual compensation in a unification-based multimodal architecture, these results provided further evidence for the benefits in terms of overcoming errors that can be achieved through multimodal integration.

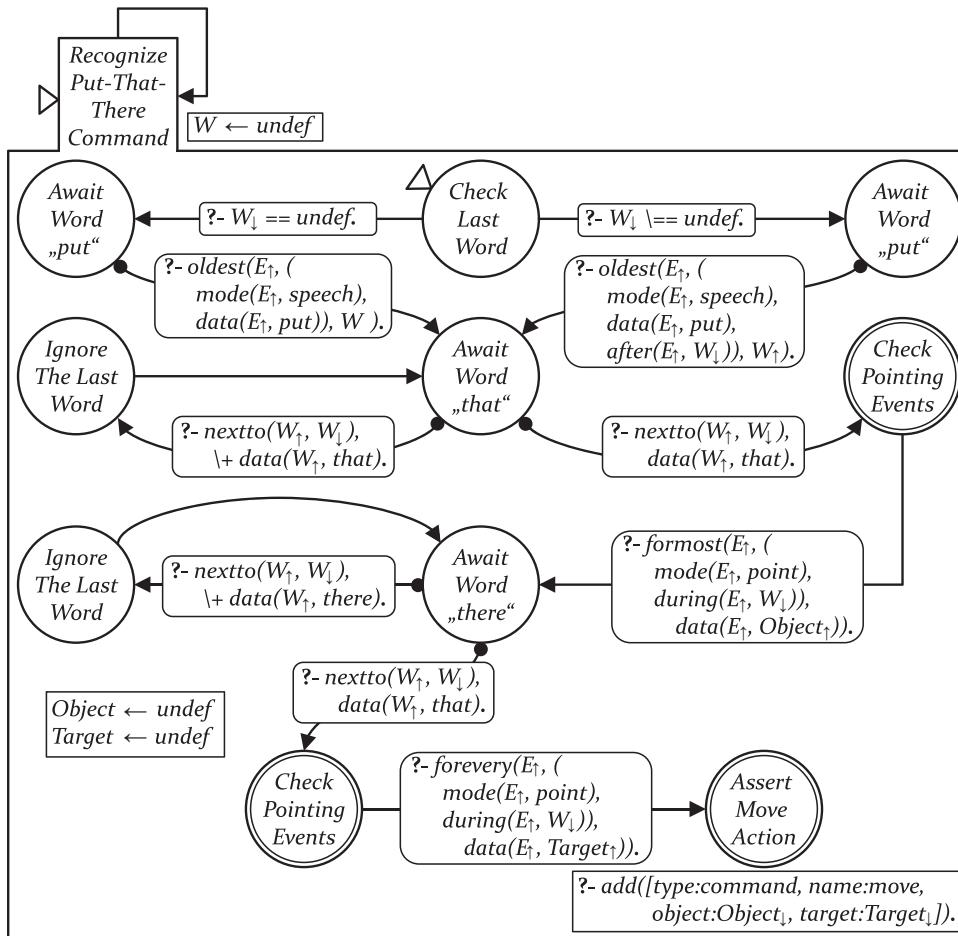
Both the unification-based approach and the finite-state approach to multimodal fusion have significant potential to improve robustness through overcoming errors. The individual studies are hard to compare though as they are for very different applications and modes. With sufficiently rich lattices, the finite-state approach should provide more potential for compensation as it scales readily to support higher levels of ambiguity in the inputs. The unification-based parsing approach applying to N-best lists from the individual modes should be able to offer the same level of error compensation, but as the N-best lists become sufficiently large the approach runs into computational complexity concerns as in the worst case the multidimensional chart parsing algorithm is exponential with respect the number of input elements [Johnston 1998].

## 1.6

### Incremental Multimodal Integration Using Event Logic and Visual Statecharts

The approaches to multimodal integration discussed so far focus on the combination of inputs from multiple modes within a single turn of input from the user in order to assign a combined meaning. They do not address the management of the internal dynamics within a turn: the incremental processing of inputs and dynamic feedback to the user (*multimodal interaction management*). To some extent this is a function of the commonly available capabilities of speech recognition engines when these approaches were developed. More recently, it has become commonplace for multimodal systems to support incremental speech recognition where the user receives immediate feedback on what the system thinks they are saying. Incrementality is also important for handling of other continuous input streams such as gesture and 3D pointing gestures where it can be used to provide immediate feedback on an entity selected or brought into focus.

Mehlmann and André [2012] develop an approach to multimodal integration that specifically seeks to address the incremental and continuous nature of multimodal input. They represent multimodal inputs as streams of multimodal events containing information about the mode, input tokens (e.g., recognized words, objects pointed to or gazed at), timing information, and confidence scores. An event logic consisting of a series of relations and operators over these events captures various kinds of spatial and temporal constraints and quantification over events. This event logic is used to express constraints on transitions in a statechart language that is used to process incoming multimodal input symbols, drive feedback to the user, and perform multimodal integration. The event logic is implemented in Prolog and uses the built-in theorem proving and term unification mechanisms of the language to prove and instantiate event logic expressions with respect to



**Figure 1.9** Visual statechart with event logic controlling transitions. (From Mehlmann and André [2012], used with permission)

the event database as it changes over time. A visual programming environment is used to construct statecharts for different kinds of multimodal inputs and desired system behavior. Figure 1.9 is an example statechart that supports fusion of the incoming speech input “put that there” with a sequence of two gestures. The transition network captures a sequence of statements in event logic that need to be true in sequence in order to interpret and integrate the particular multimodal command, in this case a command to move a block to a particular tile in an interface. The event logic expressions at the beginning of the statechart are triggered by the

appearance of the word “put” in the speech mode. Once the word “that” is found, the system looks for pointing events. Further progress through the graph is based on finding a pointing event at an object. The  $during(E,W)$  constraint is used to check temporal alignment of the spoken word and the pointing gesture. Similarly, when “there” is observed in speech it triggers a check for a pointing gesture at a target. The final event logic expression adds a command for a move specifying the object (*Object*) pointed at and the target (*Target*) as its two arguments. In this approach, the possible sequences of multimodal input and their resulting multimodal semantic interpretation are essentially captured in an augmented transition network (ATN) [Woods 1970] where the conditions on transitions in the ATN are expressions in an event logic language that must be satisfied with respect to a database of events that unfolds over time.

Like the multimodal grammar approaches described in Section 1.5, the event logic approach captures multimodal integration as a process that applies directly to tokens of input in multiple simultaneous input streams. Unlike those approaches, however, it does not model the input streams directly as sequences to be consumed, more like the unification-based parsing approach the inputs are represented as a collection of events. In the unification-based approach, the events are at the level of whole speech utterances and individual gestures, while in the event logic approach they are at the level of words and gestures. In the unification-based parsing approach, the events are combined with respect to the application of a multimodal parsing algorithm that applies feature structure rules to the inputs in order to determine feasible combinations, while here the integration patterns are captured using statecharts augmented with expressions in the event logic. In the finite-state approach, transitions are conditioned on the appearance of tokens in the input sequences while in the event logic approach the transitions are conditioned on proving of complex constraints against a Prolog database of events, where the events are occurrence of spoken words or gestures. The event logic approach also differs from unification-based multimodal integration, finite-state multimodal integration, and other previous techniques in that it attempts to provide a declarative model not just of the integration of modes, but also of the *incremental interaction management* of the resulting multimodal user interface. As in the multimodal grammar rule schemata of Johnston [1998], timeouts on consideration of inputs and actions to provide feedback to the user, rather than being coded in a surrounding multimodal interaction manager, are captured in the statechart and associated event logic expressions. See also Chao and Thomaz [2012] for a related approach using timed petri nets for modeling concurrency and synchronization of multiple modalities in a human-robot interaction use case.

While it provides a declarative representation not just of multimodal integration but of the interaction management strategy and incremental system behavior within the turn, issues remain for the event logic approach. It is not clear how it will scale from detailed specification of the handling of simple examples to handling of the broad range of speech that needs to be supported in a robust multimodal system. There may also be issues in the handling of errorful input, especially in the context of incremental speech recognition, where early hypotheses are revised as more speech is processed and recognized. Also, the words in a speech recognition hypothesis are not independent, they depend on the recognition hypotheses for the surrounding words. In a model where each word constitutes a separate event the system will need to keep track of which words are part of the same hypothesis.

## 1.7

### Multimodal Reference Resolution and Multimodal Dialog

The approaches to multimodal integration we have discussed so far are primarily focused on modeling the combination of content from different modes within a single turn of input from the user. A subset of multimodal integration phenomena, specifically resolution of *deictic references* to objects (“that,” “this one”) or locations (“here,” “there”) and implicit variants of the same can alternatively be addressed as part of a reference resolution mechanism. One of the benefits of this approach is that deictic expressions can in some cases either be resolved with respect to a gesture or with respect to the dialog context. For example, in the sample dialog in Figure 1.10 the referent for the expression “that ball” is the “blue ball,” unless there is an explicit gesture to another ball in the scene.

Huls et al. [1995] develop a multimodal prototype (EDWARD) supporting combinations of typed input and pointing gestures made with a mouse. They capture resolution of referential expressions to multimodal actions using a salience-based context model drawing on the idea of “context factors” from Alshawi [1987]. The salience value for each potential referent is determined based on combination of weights associated with a series of linguistic and perceptual features. The weights associated with particular features are configured to decay over successive turns of interaction. Huls et al. [1995] found that this approach handled all of the referring expressions in their test set (125 examples). Developing on this approach, Kehler and colleagues [Kehler et al. 1998, Kehler 2000] examine multimodal ref-

“Put the *blue ball* on the table” “Place the red cube next to *that ball*”

**Figure 1.10** Discourse resolution of deictic phrase.

erence resolution in the context of a prototype system supporting speech and pen input. Using a Wizard-of-Oz methodology, a small set of 174 referring expressions was collected from users completing travel booking scenarios. They found that all of the data could be accounted for using a simple decision list algorithm without reference to the type of referential expression or its distributional properties (e.g., syntactic position). The first step is to constrain the search based on the content of the referring expression, e.g., for “the hotel” consider only referents that are hotels. If that does not resolve the referent, then if there is a simultaneous gesture to an object that is the referent, otherwise the referent is the currently selected object. The simplicity of the solution here may well relate to the simplicity of the multimodal system they consider rather than being a general result. Even in pen-based systems, and certainly with gaze and 3D gesture input, there will often be more than one potential referent selected by a gesture. In the application domain in [Kaiser et al. \[2003\]](#), there are many potential objects a 3D gesture may refer to, and combination of speech and gesture resulted in a 67% relative reduction in error in the resolution of pointing gestures. Similarly in the 2D environment in the real estate application described in [Chai et al. \[2004\]](#) there are multiple referents for gestures made on a crowded map display.

[Chai et al. \[2004\]](#) develop an approach to multimodal reference resolution that addresses inputs with multiple gestures and imprecise gestures, such as pointing or circling on a densely populated map display. In their approach, the sequence of referring expressions in the speech input is represented as an *attributed relational graph* (ARG), capturing semantic and temporal properties of each referring expression and the sequence in which they appear. Potential referents both from gesture input and from the dialog history are also represented as an attributed relational graph. This captures the sequence of gestures and interlinks entities from the history into the graph so they can be matched against spoken expressions. *Multimodal reference resolution* is achieved using a graph matching algorithm which attempts to find the most likely alignment of the speech graph to the combined gesture and history graph. Chai et al. conducted a user study applying the approach to an experimental map-based multimodal system for real estate search. Like the finite-state approach, this graph matching approach represents sequences of spoken expressions and sequences of gestures as networks to be aligned. The graph-matching approach goes further in adding in support for resolving referring expressions with respect to the dialog history. Like the unification-based approach, we see another application of constructs from graph theory though somewhat different formally from typed feature structures and graph unification. In the unification-based approaches, DAGs are used as a semantic representation for multimodal

inputs [Johnston et al. 1997] and DAGs with structure sharing are used to capture rules for combination of multimodal inputs [Johnston 1998]. Instead, here the graph is a representation of the input streams and history and graph matching algorithms are used for both multimodal fusion and reference resolution.

The unification-based approaches to multimodal input described above in Section 1.4 specifically focused on a combination of semantic content from different modes within a single turn of input from the user. Alexandersson and Becker [2001] also adopt feature structure representations but instead address dialog phenomena across multiple turns of multimodal dialog. Alexandersson and Becker’s approach is applied in SmartKom, an extensive experimental multimodal system providing an interface supporting multiple domains including local search and searching for TV content in various interaction settings including mobile, set top box, and kiosk [Wahlster 2006]. The motivation for combining feature structures with a different operation than unification comes from dialog use cases including those where the user corrects or refines their input over multiple turns. For example, after a query like “what comedy movies are playing in Heidelberg?” if the user follows up with “what about action movies?,” the combination of feature structure representations of these utterances cannot be modeled as unification as they clash on the value of the type of movie genre and will not unify. Alexandersson and Becker [2001] model these cases using an operation related to unification called overlay that is similar to unification but allows for newly stated content to override contextual content from earlier utterances.

Ehlen and Johnston developed a related approach using *destructive unification* to model the combination of refinement queries with contextual information [Ehlen and Johnston 2012, 2013]. This approach was deployed in the Speak4It multimodal mobile application for local search. In their approach, natural language understanding is performed using a statistical intent classifier. Certain spoken and multimodal queries such as “what about korean food?” and “how about here <circle>?” are classified as *refine* queries by natural language understanding and the multimodal dialog manager in Speak4It uses a destructive unification algorithm to combine these queries with feature structures representing recent commands to the system held in the context. Destructive unification is an operation which combines graphs through unification but instead of failing in cases where there are feature paths with clashing values, the values in the second graph appear in the unified graph. The context and incoming commands are represented as typed feature structures and an ontology of types of entities and commands is used to control the combination of content. The same destructive unification mechanism is applied in the Interact multimodal pen/voice virtual assistant [Selfridge and Johnston 2015].

See Appendix 1.A for a detailed example of the application of destructive unification in a multimodal conversational system.

## 1.8 Applications of Machine Learning to Multimodal Integration

The multimodal integration techniques we have described so far are primarily rule-based or knowledge-based. Historically, one of the reasons for this has been the lack of sufficient quantities of data to explore the use of machine learning for multimodal integration and the high cost of collecting and labeling of streams of interaction with multimodal systems, a challenge that remains today. The nature of the task may also have played a role. The multimodal integration methods for conversational systems described here generally are not addressing overall classification of user intent or properties of users, rather they model the combination of semantic content from different modes into a structured semantic representation.

In recent years, a number of studies have explored the application of a variety of different machine learning approaches to intent classification (Maximum Entropy (MaxEnt), Support Vector Machines (SVMs) [[Cortes and Vapnik 1995](#)]) and sequence tagging (e.g., Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [[Lafferty et al. 2001](#)], and Recurrent Neural Networks (RNN) [[Ma and Hovy 2016](#)]) to tasks that involve fusion of information from different input modes. However many of these studies do not address user input to interactive conversational systems rather tasks such as automated classification and segmentation of meetings [[Al-Hames et al. 2006](#)] and classification of videos [[Chatterjee et al. 2015](#)]. There are a number of studies, however, that do apply machine learning to multimodal fusion in an interactive setting [[Morency et al. 2007](#), [Eisenstein and Davis 2004](#), [Chen and Di Eugenio 2013](#), [Ehlen and Johnston 2010](#), [Wu et al. 2002](#), [Vo 1998](#)]. [Morency et al. \[2007\]](#) use support vector machines trained on both visual features and spoken context features to more accurately detect head nods made by the user in a human-robot interaction scenario and an interactive browsing scenario. They found a significant improvement over using only visual features for head nod detection by including features from the speech modality.

[Eisenstein and Davis \[2004\]](#) explored the role of visual and linguistic evidence in classifying hand gestures into a series of four fairly broad categories. Following [McNeill \[1992\]](#), these categories are deictic gestures (pointing, indicating), iconic gestures (depicting features of action or event), beat gestures (small movements serving a pragmatic function such as emphasis), and metaphoric gestures (representation gestures capturing non-physical phenomena, e.g., a circling gesture to

indicate repetition).<sup>5</sup> The domain for the experiment involved a user standing at a white board, speaking, drawing, and gesturing to explain the operation of various mechanical devices. The work does not describe an interactive system but there are interactive applications this work could be applied to such as tutoring. Audio and video recordings of nine subjects explaining three different mechanical devices were collected. Human annotators were asked to manually classify gestures into three different categories under a number of different experimental conditions where they had access to audio, video, or both modalities. Given low occurrence in the corpus, metaphoric gestures were left out. The study found significantly higher agreement with the majority class assigned to each gesture for the multimodal condition with both audio and gesture providing evidence that both visual and linguistic features provide information that helps to discriminate among the different gesture types. Eisenstein and Davis also examined the ability of a classifier to predict gesture type, trained only on textual features extracted from the transcription for a time window around the gesture. The authors tested a series of different classifiers from the Weka package [[Witten and Frank 2009](#)]. The best performing approach (hyperpipes, a simple classifier based on fit to the attribute bounds for each category) achieved 65.9% accuracy, well in excess of the majority baseline (deictic, 48.7%) and above the accuracy of humans with audio only (45%), but below the performance of humans with audio and video (78%).

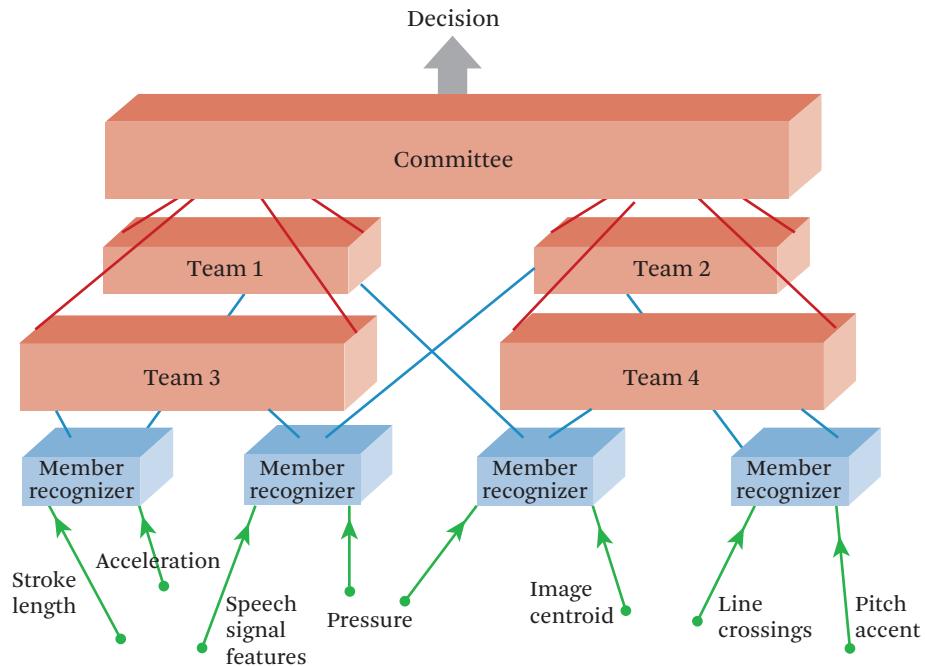
[Chen and Di Eugenio \[2013\]](#) describe a corpus of annotated multimodal data in the setting of eldercare in the home, where an adult helper assists an elderly person with day-to-day activities. A corpus of human-human activities was annotated in detail including spoken utterances and gestures and actions including pointing and holding. They performed a detailed investigation of the effectiveness of a number of different classifiers, including Maximum Entropy, CRF, and Naïve Bayes for the task of classifying utterances into a series of dialog act categories. The dialog act categories are drawn from the scheme developed for the task-oriented MapTask corpus [[Anderson et al. 1991](#)] with some additions to support gesture or action only contributions. In all cases, the addition of gesture modality-related features improved dialog act classification performance. Overall, the MaxEnt classifier provided the best performance. This experiment is not itself an interactive system but is part of a project (RoboHelper) aimed at developing the technologies to enable a robot helper for eldercare.

---

5. See [Cassell \[1998\]](#) for a detailed description of these different gesture types.

In the context of mobile applications, other factors beyond gesture need to be considered, including the physical location of the display and manipulation of maps and other complex displays. Ehlen and Johnston [2010] develop an approach for predicting which location is salient over time during interaction with a mobile multimodal system. Through examination of live field data from a mobile local search application [Johnston and Ehlen 2010] they observed that multiple different locations are potentially salient to a spoken command such as “sushi restaurants near there.” These include gestures made on the display (the system examined supports drawings made on a touch screen), the location that the map on the device currently displays, the physical location of the device (GPS), and locations referred to in previous turns of the dialog. While Kehler et al. [1998] (see Section 1.7) found reference to the current map location to be an outlier, this was a common pattern in the Speak4It application [Johnston and Ehlen 2010]. This possibly relates to changes in touch screen interfaces to maps, specifically the swipe to pan, pinch to zoom paradigm that has become established. Drawing on ideas from the work of Huls et al. [1995], Ehlen and Johnston initially developed an approach using context factors and deployed it in the production Speak4It application. Explicit confirmation of the location the user considered salient was then collected in a percentage of the live traffic to the fielded application. Based on this field data, they compared the initial heuristic strategy to modeling location salience to a machine learning approach using boosted decision trees [Witten and Frank 2009]. This classification-based approach yielded a significant improvement in performance over the heuristic strategy, achieving 85.1% concept accuracy, a 3.8% absolute improvement over the heuristic strategy (81.32%).

In the unification-based approach to multimodal integration [Johnston et al. 1997], in order to disambiguate in cases where more than one multimodal combination was licensed through unification, a simple weighted linear combination of scores from the speech recognition and gesture recognition algorithms was used. Wu et al. [1999], [2002] also utilize unification for the combination of content from multiple modalities but augment the approach with a trainable approach for assigning co-occurrence probabilities to multimodal combinations. They adopt a Member-Team-Committee (MTC) architecture (Figure 1.11) for ranking multimodal integration results. Individual member recognizers in the architecture include speech recognizers and gesture recognizers and these may use different kinds of models. At the *team* level, weighted combinations of the posterior probabilities from speech recognition and gesture recognition models are calculated and normalized in order to determine probability estimates for combined multimodal commands. Critically, the combination of speech and gesture posteriors is made



**Figure 1.11** Member-team-committee fusion architecture [Wu et al. 1999, 2002]. (Adapted from Oviatt and Cohen [2000], used with permission)

sensitive to the particular command type. At the *committee* level, the outputs from the teams are examined and a sorted list of potential multimodal combinations is generated for input combinations for which the probability estimate exceeds a confidence threshold. The surviving paired N-best lists of speech and gesture are then combined by feature structure unification [Johnston et al. 1997], determining which pairs form meaningful combinations. Empirical tests with the QuickSet Pen-Voice system [Cohen et al. 1997, Oviatt 1999] demonstrated a 5.93% absolute reduction in error rate compared to unweighted multiplication of speech and gesture probabilities (4.74% vs. 10.67% error). Limitations to this approach include it only addressing combination of a single spoken utterance with a single gesture and the assumption that a researcher has first conducted exploratory data analysis to find relevant features with which to train the member classifiers. The hierarchical structure of the MTC approach is similar in appearance to a deep neural network (DNN). There are important differences however. In the MTC architecture the outputs of the lower level recognizers are classifications of the gesture and speech.

The teams combine these to suggest their own outputs and the committee selects among those. In a prototypical DNN architecture for image or gesture recognition, lower level networks would expose combinations of features derived from the input features (hidden units), that are then combined through the operation of higher layers. Also, unlike a DNN there is no training of the overall network. Error is not propagated from the higher levels down through to the member recognizers during training.

One early application of neural networks to fusion of multiple modalities such as speech and gesture is described by [Vo and Waibel \[1997\]](#) and laid out in more detail in Vo's thesis work [[Vo 1998](#)]. The task they address is combination of continuous speech with pen inputs. They consider a series of different application domains, including multimodal interaction with maps and with calendar displays. More specifically, they develop an approach to aligning and segmenting streams of speech and gesture symbols using Multi-State Mutual Information Networks (MS-MIN). These networks determine the optimal alignment of series of words in the speech input and a series of events from pen gesture recognition with a series of high level semantic labels. The approach to assignment of semantic labels draws on work using mutual information [[Gorin et al. 1991](#)] to classify input sequences and is implemented in a type of connectionist network.<sup>6</sup> The mutual information network contains intermediate nodes between the input tokens and output labels, but these are not hidden layers per se, rather they represent bigram and trigram combinations of adjacent words in the input string. As such they serve a similar function to bigram and trigram lexical features in other approaches to text classification, using for example, logistic regression or support vector machines. Also, unlike other neural networks, the network parameters are not trained through back propagation of error on training data, rather node weights are assigned directly based on counts of the appearance of different N-grams in training examples. In fact, in the speech and pen applications Vo considers, very little data was available and so a significant part of the work was to develop a graphical authoring environment in which application authors could specify multimodal grammar patterns and the weights for the Multi-state Mutual Information Network are calculated directly from hand-crafted rule templates. In principle, the approach should be trainable based on realistic user interactions, but in [Vo \[1998\]](#) the evaluation of incremental training

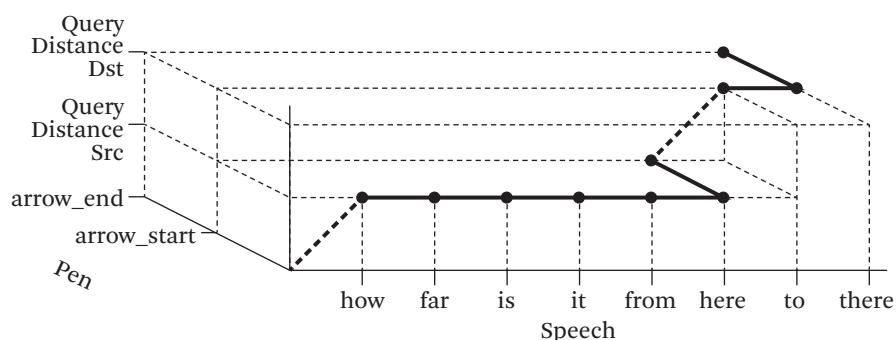
---

6. Gorin was one of Vo's thesis advisors and applied similar techniques using mutual information to the development of the How May I Help You? (HMIHY) system using stochastic understanding at AT&T for call classification [[Gorin et al., 1997](#)].

of the MS-MIN is with respect to synthetic data generated from the handcrafted multimodal grammar model (MMGL).

In order to assign not just a single label to the input, but rather a sequence of semantic labels, the outputs of the mutual information network feed into a *dynamic programming* algorithm similar to Viterbi search [Viterbi 1967] or Dynamic Time Warping [Rabiner et al. 1978] which determines the sequence of labels and alignment which optimize the overall score of the winning path. This use of Viterbi search on top of a classifier is similar to other approaches to sequence tagging used in natural language processing such as maximum entropy Viterbi approaches [McCallum 2000]. One of the key innovations of Vo's approach was to view the alignment and tagging of the input as operating over multiple token streams, which can be visualized in as in Figure 1.12. The multimodal aspects of the approach are captured by having the path scoring algorithm operate over multiple dimensions where each input stream is represented in a single dimension.

It is important to note that in this approach the MS-MIN network does not perform the actual combination of content, the semantic integration itself. The network instead uses the mutual information between events in the two channels to align them and associate them with a segmentation of the input into a series of semantic categories (labels). For example, in Figure 1.12 the words “to there” and the recognition of the end of an arrow are assigned to the semantic category *QueryDistanceDst*. The derivation from that alignment of a combined actionable semantic representation is performed by a post processing step using handlers authored in Java for each of the semantic categories.



**Figure 1.12** Example of multimodal alignment and segmentation of speech and gesture tokens.  
(From Vo [1998], used with permission)

Like [Johnston et al. \[2002b\]](#) and [Bangalore and Johnston \[2009\]](#), the approach does not directly use fine-grained temporal constraints in the multimodal segmentation and alignment, instead temporal constraints (based on findings from [Oviatt \[1997a\]](#)) are used to group gestures and speech into multimodal utterances and the alignment and segmentation relies purely on sequential order and the tokens themselves rather than fine-grained temporal constraints. Temporal constraints are used outside of the multimodal fusion mechanism to determine what constitutes a multimodal utterance.

## 1.9

### Conclusion

This chapter examined a number of different approaches to the problem of integrating and understanding multimodal inputs to interactive systems. Some approaches are more procedural, while others are more declarative. Some are focused on mode integration within a single user turn [[Johnston et al. 1997](#), [Johnston 1998](#), [Johnston and Bangalore 2005](#)] while others address the interplay of multimodal integration with reference resolution and discourse [[Kehler et al. 1998](#), [Chai et al. 2004](#), [Alexandersson and Becker 2001](#), [Ehlen and Johnston 2012](#)]. In some, multimodal integration is separated from speech understanding: two-level approaches [[Johnston et al. 1997](#), [Johnston 1998](#), [Chai et al. 2004](#)] in others it is part of the same process: one-level approaches [[Johnston and Bangalore 2001](#), [2005](#), [Bangalore and Johnston 2009](#), [Mehlmann and André 2012](#)]. Some of the approaches are rule or knowledge-driven, while others apply machine learning techniques, and others are hybrids incorporating both knowledge-based techniques and machine learning. Different approaches focus on different aspects of the multimodal fusion problem. In Table 1.1 we characterize different aspects of the different approaches.

The strengths of unification-based integration are in providing a declarative mechanism for integration and supporting mutual compensation. The unification-based parsing approach provides more expressive power enabling support for multiple gestures and gesture parsing and provides a declarative mechanism for statement of temporal and spatial constraints. The finite-state approach to multimodal integration attempts to model multimodal integration as part of understanding and parsing by applying directly to streams of words and gestures. It also addresses mutual compensation and enables direct application of the multimodal understanding mechanism to lattice inputs from speech and gesture recognition. The event logic approach [[Mehlmann and André 2012](#)] is also a one-level approach applying directly to words and gestures, and takes declarative representation a step further by providing a statechart representation to capture incremental interaction

**Table 1.1** Comparison of different approaches to multimodal fusion for interactive systems.

	Combines Speech Understanding and Multimodal Integration	Level at Which Integration Applies	Semantic representations of spoken utterances and gestures as feature structures	Addresses Dialog Phenomena	Addresses Incremental Interaction	Knowledge-based vs. Learning
Approach to Multimodal Fusion						
Unification-based Multimodal Integration/Parsing [Johnston et al. 1997, 1998]	No			No	No	Knowledge-based
Finite-state multimodal integration [Johnston and Bangalore 2001, 2005, Bangalore and Johnston 2009]	Yes		Words and gestures represented as lattices	No	No	Knowledge-based
Integration with event-logic statecharts [Mehlmann and André 2012]	Yes		Words and gestures represented as events	Yes	Yes	Knowledge-based
Multimodal graph matching [Chai et al. 2004]	No	Discourse level		Yes	No	Knowledge-based
Overlay, Destructive unification [Alexandersson and Becker 2001, Ehlen and Johnston 2012]	No	Discourse level		Yes	No	Knowledge-based
Multimodal Location Grounding [Ehlen and Johnston 2010]	No	Discourse level		Yes	No	Supervised classification
Multimodal Gesture Classification [Eisenstein and Davis 2004]	No	Linguistic features used to predict gesture type		No	No	Supervised classification
Multimodal Dialog Act Prediction [Chen and Di Eugenio 2013]	No	Words and gestures		Yes	No	Supervised classification
Multi-state Mutual Information Networks [Yo and Waibel 1997]	Yes	Words and gesture elements		No	No	Hybrid
Member-Team-Committee [Wu et al. 1999, 2002]	No	Semantic representations of spoken utterances and gestures as feature structures		No	No	Hybrid

management. It is less clear though how to support highly indeterminate input such as speech and gesture lattices in this approach and how it scales to a broader range of language input. [Chai et al. \[2004\]](#) focus on the interplay of multimodal integration with reference resolution and how to handle realistic situations on map displays where gestures are vague and may refer to multiple different entities. [Alexandersson and Becker \[2001\]](#) and [Ehlen and Johnston \[2013\]](#) extend unification to handle multimodal discourse phenomena.

Substantial challenges remain in the development of multimodal integration mechanisms. [Mehlmann and André \[2012\]](#) have started to address issues concerning incrementality although there is more work to explore in this area in order to robustly handle broad coverage language input. Another significant challenge for the future is enabling more broad application of machine learning to rich multimodal fusion tasks. There have been several successful applications of supervised classification on multimodal features to tasks such as gesture type prediction [[Eisenstein and Davis 2004](#)], dialog act tagging [[Chen and Di Eugenio 2013](#)], and location salience grounding [[Ehlen and Johnston 2010](#)]. [Vo and Waibel \[1997\]](#) was an early application of a connectionist network to the task of segmenting, aligning, and labeling multiple inputs over multiple dimensions. [Wu et al. \[1999\], \[2002\]](#) employ a multi-level approach to the combination of recognition hypotheses from multiple different recognizers, speech, and gesture recognizers in the multimodal case.

The majority of work on taking content from multiple modes and composing it into a combined semantic representation has used knowledge-based techniques with rules of various kinds. In the MTC approach [[Wu et al. 2002](#)], the possible recognition hypotheses are ranked through the MTC mechanism, but the semantic combination itself is achieved through unification of typed feature structures [[Johnston et al. 1997](#)]. In the MS-MIN approach [[Vo and Waibel 1997](#), [Vo 1998](#)] the combination of a mutual information-based classifier and path-scoring algorithm is used to segment, align, and label the multimodal input streams, but the process of creating a combined semantic representation is achieved procedurally using post-processing handlers authored in Java.

Unification-based approaches involve authoring of feature structure schemata for combining inputs. For the finite-state approach, rules are authored as multimodal phrase structure rules and then compiled into finite-state transducers. Potentially with availability of sufficient data the multimodal alignment and meaning assignment transducers could be induced from data. One of the factors that has historically impacted the application of machine learning for fusion is the lack of availability of sufficient quantities of annotated data for training. Another key future challenge is in online learning and adaptation of multimodal integration so

that systems can optimize their performance, learning personalized temporal and spatial constraints and adapting to the user's specific language and gestures.

Neural networks have become exponentially more popular in recent years with the resurgence of deep neural networks (DNNs), and we have seen significant gains in many tasks that can be attributed to the application of DNNs, including image recognition and speech recognition tasks. One risk for multimodal integration strategies that combine content from multiple modes after earlier recognition components have applied, is that interpretations of each signal that would be preferred given evidence from the other modes are prematurely removed from consideration. Inspired by the recent successes of end-to-end DNN approaches to speech recognition [Amodei et al. 2016] and text-to-speech [Wang et al. 2017] it should be possible to explore the creation of end-to-end multimodal understanding systems where, for example, speech recognition, gesture recognition, fusion, and understanding mechanisms are composed into a single multi-layer deep neural network which can be trained to optimally assign a combined interpretation to multiple streams of input.

## **1.A Explanation of Finite-state Multimodal Understanding Mechanism**

### **1.A.1 Simulating 3-tape Finite-state Device with Transducers**

In order to simulate the behavior of the three-tape device the multimodal grammar is compiled into two transducers  $R$  and  $T$ . The first of these  $R : G \rightarrow W$  maps from gesture symbols on the input tape to words on the output tape, and is used to perform an alignment of the speech and gesture lattices. The second of these  $T : G \times W \rightarrow M$  maps from an extended alphabet of multimodal tokens on the input tape into the meaning representation on the output tape. The  $R$  machine is composed with the speech and gesture lattices  $G'$  and  $W'$ , respectively  $((G' \circ G : W) \circ W')$ , yielding a transducer  $G' : W'$ . This resulting transducer is then factored onto a single tape finite state machine  $G'_-W'$  where the symbols are from a multimodal alphabet that combines gesture and speech tokens, a multimodal token language. This finite state machine is composed with the  $T$  transducer  $(G'_-W' \circ G_- W : M)$  resulting in a  $G'_-W' : M$  transducer which captures the relationship between the possible word and gesture interpretations in the input streams and their combined meaning. In order to read off the meaning representation, the  $M$  symbols are concatenated along the best path in the  $G'_-W' : M$  machine. The  $R : G \rightarrow W$

and  $T : G \times W \rightarrow M$  are approximated directly from the grammar using standard approximation techniques [Nederhof 1997]. The non-terminals in the  $W:G:M$  machine are mapped to  $G\_W : M$  transitions in  $T : G \times W \rightarrow M$  and the  $G\_W$  patterns are mapped to  $G : W$  transitions in the alignment transducer  $R : G \rightarrow W$ .

### **1.A.2 Representing Specific Content**

One of the characteristics of gesture input such as ink gestures made on a map display is that they convey specific semantic content such as map coordinates or the identities of selected items, such as the restaurants circled on an interactive map display. This semantic content needs to be incorporated into the resulting combined semantic representation for the multimodal input. In a more powerful formalism, such as feature structure grammars, this can be achieved through structure sharing in the underlying graph or by some kind of copying operation. When modeling multimodal integration as finite-state composition it is not practical and unwieldy to encode all of the possible different specific items, such as coordinates on a map or all IDs of entities on a complex display, in the finite-state transducers used to model integration (the  $R$  and  $T$  transducers). This problem can be addressed by storing specific contents from gesture in a series of buffers  $b_1 \dots b_n$  outside of the finite-state processing cascade, and not modeling these contents directly in the transducer. Instead in the transducer, a finite set of buffer references from the gesture are “copied” into the meaning stream with arcs such as  $\epsilon : b_1 : b_1, \epsilon : b_2 : b_2, \dots \epsilon : b_n : b_n$ . [Johnston and Bangalore 2001, Johnston and Bangalore 2005].

Bangalore and Johnston [2009] describe a more flexible approach to specific gesture content that is not limited by the number of buffers. In their approach, the gesture input is represented as a transducer  $I : G'$  which has the gesture symbols with specific content on the input tape and gesture symbols with specific content replaced with a reserved symbol  $SEM$  on the output tape. The finite-state composition of gesture with the alignment transducer  $G : W$  is with a projection  $G'$  from the output of  $I : G'$ . The multimodal grammar indicates locations where specific gesture content should go into the resulting meaning with terminal symbols of form  $\epsilon : SEM : SEM$ . After application of the  $T$  transducer, the resulting  $G'\_W' : M$  machine is factored into a  $G' : M$  machine, a move back from the multimodal  $G'\_W'$  input symbols to the  $G'$  symbols. This machine can then be composed with the original  $I : G'$  transducer,  $I : G' \rightarrow G' : M = I : M$ . This operation re-aligns the  $SEM$  symbols with the appropriate specific contents. In order to extract the meaning for the top path in the  $I : M$  transducer, symbols are concatenated from the  $M$  tape,

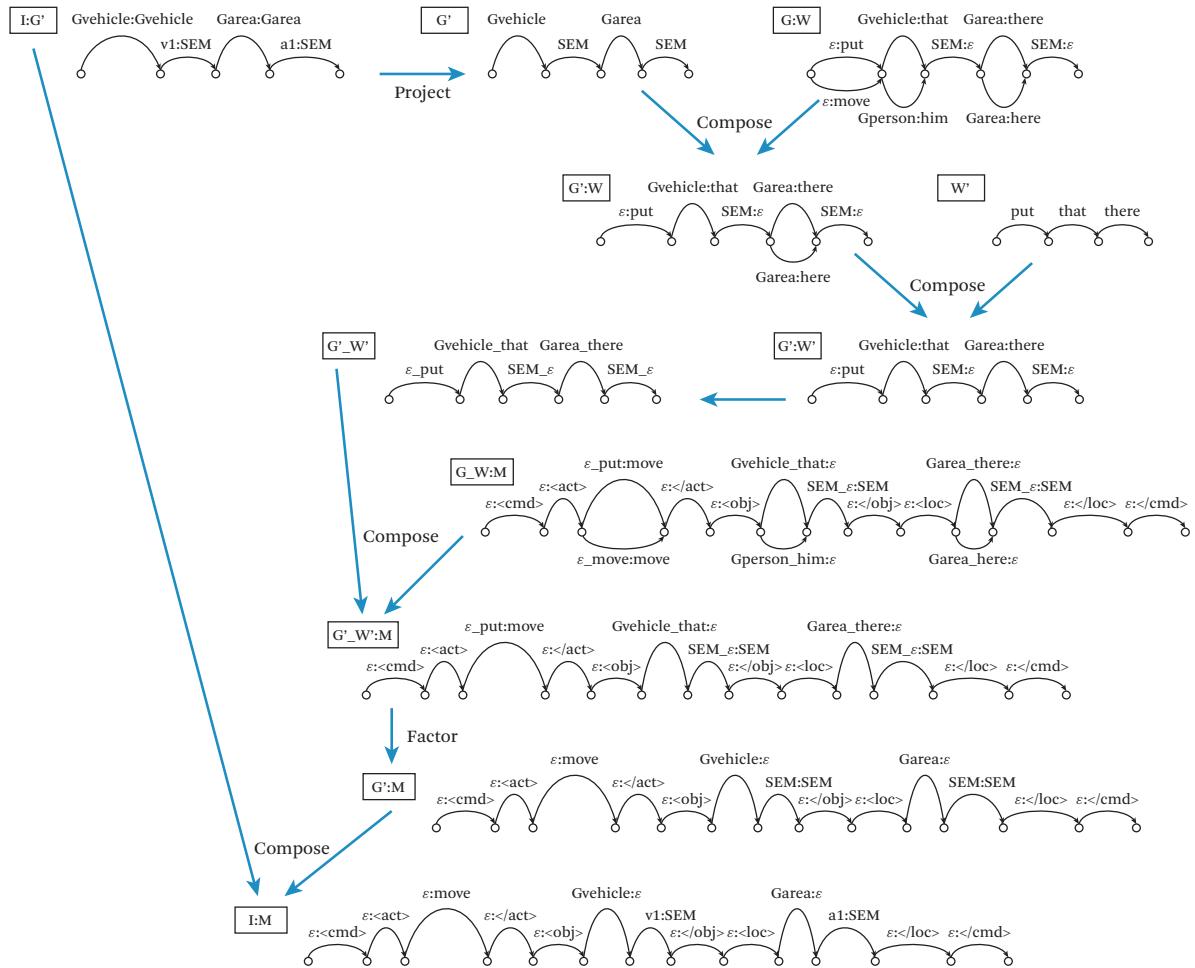
unless the symbol on  $M$  is *SEM*, in which case the content from  $I$  tape is taken instead.

### **1.A.3 Example of Finite-state Language Processing Cascade on Multimodal Inputs**

Figure 1.13 summarizes the cascade of finite-state operations in the approach to multimodal integration described in [Bangalore and Johnston \[2009\]](#) and illustrates the application of the cascade with an example that corresponds to the multimodal grammar example of “put that there” in Figure 1.8 in the chapter. The gesture lattice  $I : G'$  captures a sequence of gestures selecting a vehicle and then an area. A projection is taken on the output to abstract over the specific content. This is then composed with the alignment transducer  $G : W$  (top right) yielding  $G' : W$ . Note, for example, that while in the alignment transducer the first gesture could be *Gperson*, a gesture at a person, composition removes this possibility given the lack of that gesture in the gesture lattice. The  $G' : W$  transducer is next composed through finite-state composition with  $W'$ , the speech lattice, here capturing the sequence of words “put that there.” Now we have  $G' : W'$  an aligned representation of the speech and gesture input streams. This is factored from transducer to automaton by combining the speech and gesture symbols onto one tape. The resulting  $G'_-W'$  is composed with the  $T : G \times W \rightarrow M$ , the meaning mapping transducer yielding  $G'_-W' : M$ . This is factored into the transducer  $G' : M$ , which is of the correct form to compose with the original  $I : G'$  gesture transducer from the start. This operation essentially uses the  $G'$  paths to re-index back to the specific contents. The meaning is read from the output of the  $I : M$  transducer unless the meaning symbol is *SEM* in which case the input (the specific content) is taken instead.

### **1.A.4 Example of Destructive Unification for Contextual Resolution**

In Figure 1.14 we show an example dialog from Speak4It [[Ehlen and Johnston 2013](#)] demonstrating at each step how destructive unification of typed feature structures is used to model the fusion of refine queries with content from the multimodal dialog context, and determine which parts of the query should persist and which are updated by the new query. In the bottom right of the figure, there is a segment of the type hierarchy referenced by the destructive unification operation. The user’s first query 1. results in display of the locations of Japanese restaurants near Santa Monica. The user then makes a refinement 3. “How about chinese?”. This is classified as refinement and the multimodal dialog manager destructively unifies the resulting feature structure against the previous



**Figure 1.13** Finite-state multimodal language processing cascade.

command. *category* is a super type of *restaurant* in the type hierarchy so the destructive unification succeeds and in the case of the clashing value of *cuisine*: the new value *chinese* is taken. The user next asks “what about McDonalds?” 5. Since the type of the refinement *business\_name* clashes with *restaurant* the whole *term:* is replaced in the destructive unification. This captures the fact that the result of 5. should be a search for McDonalds, not Chinese restaurants called McDonalds.

1. User: "japanese restaurants near Santa Monica California"
2. System: [returns japanese restaurants near santa monica]

$\left[ \begin{array}{l} \text{term: } \left[ \begin{array}{l} \text{cuisine: japanese} \\ \text{type: restaurant} \end{array} \right] \\ \text{location: } \left[ \begin{array}{l} \text{city: santa\_monica} \\ \text{state: california} \\ \text{type: city\_state} \end{array} \right] \end{array} \right]$

3. User: "how about chinese?"

$\left[ \begin{array}{l} \text{term: } \left[ \begin{array}{l} \text{cuisine: chinese} \\ \text{type: category} \end{array} \right] \end{array} \right]$

**DESTRUCTIVE UNIFICATION**

$\left[ \begin{array}{l} \text{term: } \left[ \begin{array}{l} \text{cuisine: chinese} \\ \text{type: restaurant} \end{array} \right] \\ \text{location: } \left[ \begin{array}{l} \text{city: santa\_monica} \\ \text{state: california} \\ \text{type: city\_state} \end{array} \right] \end{array} \right]$

5. User: "what about McDonalds?"

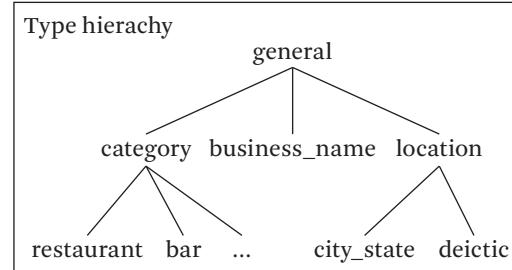
$\left[ \begin{array}{l} \text{term: } \left[ \begin{array}{l} \text{name: mcdonalds} \\ \text{type: business\_name} \end{array} \right] \end{array} \right]$

4. System: [returns chinese restaurants near Santa Monica]

**DESTRUCTIVE UNIFICATION**

$\left[ \begin{array}{l} \text{term: } \left[ \begin{array}{l} \text{name: mcdonalds} \\ \text{type: business\_name} \end{array} \right] \\ \text{location: } \left[ \begin{array}{l} \text{city: santa\_monica} \\ \text{state: california} \\ \text{type: city\_state} \end{array} \right] \end{array} \right]$

6. System: [returns McDonalds near Santa Monica]



**Figure 1.14** Example dialog with destructive unification from the Speak4it multimodal application.

### Focus questions

- 1.1. What is the role of multimodal fusion or integration in an interactive conversational dialog system?
- 1.2. What are the motivations for supporting multimodal input for interactive systems?

- 1.3.** What were the characteristics of the earliest approaches to multimodal fusion?
- 1.4.** How can combination of multimodal inputs be modeled as unification of feature structures?
- 1.5.** What are the differences between one-level and two-level approaches to processing multimodal inputs?
- 1.6.** How can multimodal inputs be combined and interpreted using finite state methods?
- 1.7.** What does it mean to process multimodal input incrementally, and which approaches attempt to address this aspect?
- 1.8.** What is the interplay between multimodal integration and multimodal reference resolution across turns?
- 1.9.** How has machine learning been applied to multimodal integration for interactive conversational systems?
- 1.10.** What are the benefits and limitations of different approaches to multimodal integration?

## References

- S. P. Abney. 1991. Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny, editors *Principle-based parsing*. IEEE, Los Alamitos, CA. pp. 257–278. DOI: [10.1007/978-94-011-3474-3\\_10](https://doi.org/10.1007/978-94-011-3474-3_10). 44
- J. Alexandersson and T. Becker. 2001. Overlay as the basic operation for discourse processing in a multimodal dialogue system. In *Proceedings of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. pp. 1–7. DOI: [10.1007/3-540-36678-4\\_17](https://doi.org/10.1007/3-540-36678-4_17). 36, 52, 59, 60, 61
- J. Alexandersson, T. Becker, and N. Pfleger. 2004. Scoring for overlay based on informational distance. In *Proceedings of KONVENS-04*. Vienna, Austria. pp. 1–4.
- C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata*, (CIAA 2007). Lecture Notes in Computer Science Vol. 4783, pp. 11–23. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-540-76336-9\\_3](https://doi.org/10.1007/978-3-540-76336-9_3). 45
- J. Allgayer, R. M. Jansen-Winkel, C. Reddig, N. Reithinger. 1989. Bidirectional use of knowledge in the multimodal NL access system XTRA. In *Proceedings of IJCAI 1989*, pp. 1492–1497. 32
- H. Alshawi. 1987. *Memory and Context for Language Interpretation*. Cambridge, UK. 50

- M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoli, and D. Zhang. 2006. Multimodal integration for meeting group action segmentation and recognition. In S. Renals and S. Bengio, editors, *MLMI 2005*, LNCS 3869, pp. 52–63. DOI: [10.1007/11677482\\_5](https://doi.org/10.1007/11677482_5). 53
- D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu. 2016. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, New York. 62
- A. H. Anderson, M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, and H. S. Thompson. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4). 54
- E. André. 2002. Natural language in multimedia/multimodal systems. In Ruslan Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, New York. 33
- O. Aran and D. Gatica-Perez. 2010. Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In *Proceedings of 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey. pp. 3687–3690. DOI: [10.1109/ICPR.2010.898](https://doi.org/10.1109/ICPR.2010.898). 25
- T. Baltrušaitis, C. Ahuja, and L.-P. Morency. 2018. Challenges and applications in multimodal machine learning. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 23
- S. Bangalore and G. Riccardi. 2002. Stochastic finite-state models of spoken language machine translation. *Machine Translation*. 17(3): 165–184. DOI: [10.1023/B:COAT.0000010804.12581.96](https://doi.org/10.1023/B:COAT.0000010804.12581.96). 44
- S. Bangalore and M. Johnston. 2004. Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of the North American Association for Computational Linguistics/Human Language Technology (NAACL/SLT)*, pp. 33–40. Boston, MA. DOI: [10.1109/ASRU.2003.1318444](https://doi.org/10.1109/ASRU.2003.1318444). 45
- S. Bangalore and M. Johnston. 2000. Tight-coupling of multimodal language processing with speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing. pp. 126–129.
- S. Bangalore and M. Johnston. 2009. Robust understanding in multimodal interfaces. *Computational Linguistics* 35(3): 345–397. DOI: [10.1162/coli.08-022-R2-06-26](https://doi.org/10.1162/coli.08-022-R2-06-26). 41, 42, 44, 45, 59, 60, 63, 64
- Bolt, R. A. 1980. “Put-that-there”: voice and gesture at the graphics interface. *Computer Graphics* 14(3): 262–270. 24, 32

- R. J. Brachman, D. L. McGuiness, P. F. Patel-Schneider, and L. A. Resnick. 1991. Living with CLASSIC: When and how to use a KL-ONE-like language. In J. Sowa, editor, *Principles of Semantic Networks*. Morgan Kaufmann, San Mateo, CA. DOI: [10.1.1.31.9028.33](https://doi.org/10.1.1.31.9028.33)
- R. Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge, UK. [30, 33, 34, 788](#)
- Cassell, J. 1998. A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland, editors, *Computer Vision in Human-Machine Interaction*, pp. 191–215. Cambridge University Press, Cambridge, UK. [54](#)
- M. Chatterjee, S. Park, L-P. Morency, and S. Scherer. 2015. Combining two perspectives on classifying multimodal data for recognizing speaker traits. In *Proceedings of ICMI 2015*, pp. 7–14. Seattle, WA. [25, 53](#)  
DOI: [10.1145/2818346.2820747](https://doi.org/10.1145/2818346.2820747). [25, 53](#)
- J. Chai, P. Hong, and M. Zhou. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of 9th International Conference on Intelligent User Interfaces (IUI)*, Madeira, Portugal. pp. 70–77. DOI: [10.1145/964442.964457](https://doi.org/10.1145/964442.964457). [26, 51, 59, 60, 61, 762](#)
- C. Chao and A. L. Thomaz. 2012. Timed petri nets for multimodal interaction modeling. In *Proceedings of ICMI 2012 Workshop on Speech and Gesture Production in Virtually and Physically Embodied Conversational Agents*, Santa Monica, CA. [49](#)
- L. Chen and B. Di Eugenio 2013. Multimodality and dialog act classification in the RoboHelper project. In *Proceedings of SigDial Conference*, pp. 183–192. Association for Computational Linguistics. Metz, France. DOI: [10.1.1.377.1919](https://doi.org/10.1.1.377.1919). [53, 54, 60, 61](#)
- J. Cocke and J. T. Schwartz. 1970. *Programming languages and their compilers: Preliminary notes* (Technical report) (2nd revised ed.). Courant Institute of Mathematical Sciences. New York University, New York. [38](#)
- P. R. Cohen, M. Dalrymple, D. B. Moran, F. C. N. Pereira, J. W. Sullivan, R. A. Gargan, J. L. Schlossberg, and S. W. Tyler. 1989. Synergistic use of direct manipulation and natural language. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'89)*, 227-234. New York: ACM Press. (Reprinted in Maybury & Wahlster editors, 1998. Readings in Intelligent User Interfaces pp. 29-37. San Francisco: Morgan Kaufmann.) [32](#)
- P. R. Cohen. 1992. The role of natural language in a multimodal interface. In *Proceedings of the 5th Annual ACM Symposium on User Interface Software and Technology*. Monterey, CA. pp. 143–149. ACM Press. DOI: [10.1145/142621.142641](https://doi.org/10.1145/142621.142641). [31](#)
- P. R. Cohen, M. Johnston, D. McGee, S. L. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. 1997. Multimodal interaction for distributed interactive simulation. In *Proceedings of Innovative Applications of Artificial Intelligence Conference*. AAAI/MIT Press, Menlo Park, CA. DOI: [10.1145/266180.266328](https://doi.org/10.1145/266180.266328). [23, 34, 56](#)
- P. R. Cohen, M. Johnston, D. McGee, S. L. Oviatt, J. Clow, and I. Smith. 1998. The efficiency of multimodal interaction: A case study. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia. [31](#)

- P. R. Cohen, D. McGee, S. Oviatt, L. Wu, and J. Clow. 1999. Multimodal Interaction for 2D and 3D environments. L. Rosenblum and M. Macedonia, editors, *IEEE Computer Graphics and Applications*. IEEE Press, New York. DOI: [10.1109/38.773958](https://doi.org/10.1109/38.773958). 24
- P. R. Cohen, E. C. Kaiser, C. M. Buchanan, and S. Lind. 2015. Sketch-Thru-Plan: A multimodal interface for command and control. *Communications of ACM*. April 2015. 58(4): pp. 56–65. DOI: [10.1145/2735589](https://doi.org/10.1145/2735589). 24, 32, 34
- P. R. Cohen, and S. Oviatt. 2017. Multimodal speech and pen interfaces. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3015783.3015795](https://doi.org/10.1145/3015783.3015795). 24
- A. Corradini, R. M. Wesson, and P. R. Cohen. 2002. A map-based system using speech and 3D gestures for pervasive computing. In *Proceedings of International Conference on Multimodal Interfaces* (ICMI). pp. 191–196. DOI: [10.1109/ICMI.2002.1166991](https://doi.org/10.1109/ICMI.2002.1166991). 24
- C. Cortes, and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20.3, pp. 273–297. DOI: [10.1023/A:1022627411411](https://doi.org/10.1023/A:1022627411411). 53
- A. Crimi, A. Guercio, G. Nota, G. Pacini, G. Tortora, and M. Tucci. 1991. Relation grammars and their application to multi-dimensional languages. *Journal of Visual Languages and Computing*, 2:333–346. DOI: [10.1016/S1045-926X\(05\)80003-5](https://doi.org/10.1016/S1045-926X(05)80003-5). 39
- L. Duncan, W. Brown, C. Esposito, H. Holmback, and P. Xue. 1999. *Enhancing Virtual Maintenance Environments with Speech Understanding*. Boeing M&CTechNet. Seattle, WA. 24
- J. Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*. 13: pp. 94–102. DOI: [10.1145/362007.362035](https://doi.org/10.1145/362007.362035). 39
- P. Ehlen and M. Johnston. 2010. Location grounding in multimodal local search. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (ICMI-MLMI '10), Beijing, China. DOI: [10.1145/1891903.1891945](https://doi.org/10.1145/1891903.1891945). 53, 55, 60, 61
- P. Ehlen and M. Johnston. 2012. Multimodal dialogue in mobile local search. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, Santa Monica, CA, pp. 303–304. DOI: [10.1145/2388676.2388741](https://doi.org/10.1145/2388676.2388741). 52, 59, 60
- P. Ehlen and M. Johnston. 2013. A multimodal dialogue interface for mobile local search. In *Proceedings of the ACM Conference on Intelligent User Interfaces* (IUI), Santa Monica, CA. pp. 63–64. DOI: [10.1145/2451176.2451200](https://doi.org/10.1145/2451176.2451200). 52, 61, 64
- J. Eisenstein and R. Davis. 2004. Visual and linguistic information in gesture classification. In *Proceedings of the International Conference on Multimodal Interaction* (ICMI). State College, PA, USA. pp. 113–120. DOI: [10.1145/1027933.1027954](https://doi.org/10.1145/1027933.1027954). 53, 60, 61
- A. L. Gorin, S. Levinson, A. Gertner, E. Goldman. 1991. Adaptive acquisition of language. *Computer Speech and Language*. 5:2, pp. 101-132. 57

- A. L. Gorin, G. Riccardi, and J. H. Wright. 1997. How may I help you? *Speech Communication*. 23, pp. 113–127.
- D. Harel. 1987. STATECHARTS: A visual formalism for complex systems. *Science of Computer Programming*. 8. pp. 231–274. North Holland. DOI: [10.1016/0167-6423\(87\)90035-9](https://doi.org/10.1016/0167-6423(87)90035-9).
- A. Hauptmann. 1989. Speech and gesture for graphic image manipulation. In *Proceedings of CHI '89*. pp. 241–245, Austin, TX. DOI: [10.1.1.47.3281.31](https://doi.org/10.1.1.47.3281.31)
- R. Helm, K. Marriott, and M. Odersky. 1991. Building visual language parsers. In *Proceedings of the Conference on Human Factors in Computing Systems: CHI '91*, ACM Press, New York. pp. 105–112. DOI: [10.1145/108844.108860.39](https://doi.org/10.1145/108844.108860.39)
- L. Hetherington. 2004. The MIT finite-state transducer toolkit for speech and language processing. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea. DOI: [10.1.1.138.3474.45](https://doi.org/10.1.1.138.3474.45)
- C. Huls, E. Bos, and W. Classen. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics* 21: 59–79. [50](#), [55](#)
- M. Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada. pp. 624–630. DOI: [10.3115/980845.980949](https://doi.org/10.3115/980845.980949.30). [30](#), [37](#), [38](#), [47](#), [49](#), [52](#), [59](#), [60](#), [786](#)
- M. Johnston. 2000. Deixis and conjunction in multimodal systems. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany. pp. 362–368. DOI: [10.3115/990820.990873.41](https://doi.org/10.3115/990820.990873.41)
- M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. 1997. Unification-based multimodal integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 281–288. DOI: [10.3115/979617.979653](https://doi.org/10.3115/979617.979653). [34](#), [52](#), [55](#), [56](#), [59](#), [60](#), [61](#)
- M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. 2002a. MATCH: An architecture for multimodal dialog systems. In *Proceedings of the Association of Computational Linguistics*, Philadelphia, PA. pp. 376–383. DOI: [10.1.1.92.2460](https://doi.org/10.1.1.92.2460). [23](#), [42](#), [46](#)
- M. Johnston, S. Bangalore, A. Stent, G. Vasireddy, and P. Ehlen. 2002b. Multimodal language processing for mobile information access. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO. pp. 2237–2240. DOI: [10.1.1.7.9135](https://doi.org/10.1.1.7.9135). [41](#), [59](#)
- M. Johnston and S. Bangalore. 2005. Finite-state multimodal integration and understanding. *Journal of Natural Language Engineering*, 11(2): 159–187. DOI: [10.1017/S1351324904003572](https://doi.org/10.1017/S1351324904003572). [32](#), [41](#), [43](#), [44](#), [45](#), [46](#), [59](#), [60](#), [63](#)
- M. Johnston and S. Bangalore. 2001. Finite-state methods for multimodal parsing and integration. In *Proceedings of the ESSLLI Workshop on Finite-state Methods*, Helsinki, Finland. DOI: [10.1.1.23.914](https://doi.org/10.1.1.23.914). [41](#), [43](#), [44](#), [45](#), [59](#), [60](#), [63](#)

- M. Johnston and P. Ehlen. 2010. Speak4It<sup>TM</sup>: Multimodal interaction in the wild. In *Proceedings of the IEEE Spoken Language Technology workshop*, Berkeley, CA. pp. 59–60. DOI: [10.1109/SLT.2010.5700840](https://doi.org/10.1109/SLT.2010.5700840). 23, 24, 55
- A. Joshi and P. Hopely. 1997. A parser from antiquity. *Journal of Natural Language Engineering*, 2(4): 6–15. DOI: [10.1017/S1351324997001538](https://doi.org/10.1017/S1351324997001538). 44
- J. Lafferty, A. McCallum, and F. C. N. Pereira, 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Departmental Paper CIS, UPENN. June 2001. DOI: [10.1.1.120.9821](https://doi.org/10.1.1.120.9821). 53
- E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)*. New York. pp. 12–19. DOI: [10.1145/958432.958438](https://doi.org/10.1145/958432.958438). 24, 32, 51
- R. M. Kaplan and J. Bresnan. 1995. Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*, pp. 173–181. MIT Press, Cambridge, MA. DOI: [10.1.1.70.3002](https://doi.org/10.1.1.70.3002). 27, 771
- R. M. Kaplan and M. Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3): 331–378. 44
- L. Karttunen. 1991. Finite-state constraints. In *Proceedings of the International Conference on Current Issues in Computational Linguistics*, Universiti Sains Malaysia, Penang. 44
- T. Kasami. 1965. *An efficient recognition and syntax-analysis algorithm for context-free languages (Technical report)*. AFCRL. 65–758. 38
- A. Kehler, J. C. Martin, A. Cheyer, L. Julia, J. R. Hobbs, and J. Bear. 1998. On representing salience and reference in multimodal human-computer interaction. In *Proceedings of the AAAI-98 Workshop on Representations for Multimodal Human-Computer Interaction*, Madison, WI. 50, 55, 59
- A. Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the AAAI'00*. pp. 685–689. Austin TX. 50
- K. K. Koskenniemi. 1984. Two-level morphology: A general computation model for word-form recognition and production. Ph.D. thesis, University of Helsinki. 44
- D. B. Koons, C. J. Sparrell, and K. R. Thorisson. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. In M. T. Maybury, editor, *Intelligent Multimedia Interfaces*. AAAI Press/MIT Press, Cambridge, MA, pp. 257–276. 33
- F. Lakin. 1986. Spatial parsing for visual languages. In S. K. Chang, T. Ichikawa, and P. A. Ligomenides, editors, *Visual Languages*. Plenum Press. pp. 35–85. DOI: [10.1007/978-1-4613-1805-7\\_3](https://doi.org/10.1007/978-1-4613-1805-7_3). 39
- M. E. Latoschik. 2002. Designing transition networks for multimodal VR-interactions using a markup language. In *Proceedings of the Fourth ACM International Conference on Multimodal Interfaces (ICMI)*, Pittsburgh, PA. pp. 411–416. DOI: [10.1109/ICMI.2002.1167030](https://doi.org/10.1109/ICMI.2002.1167030). 32

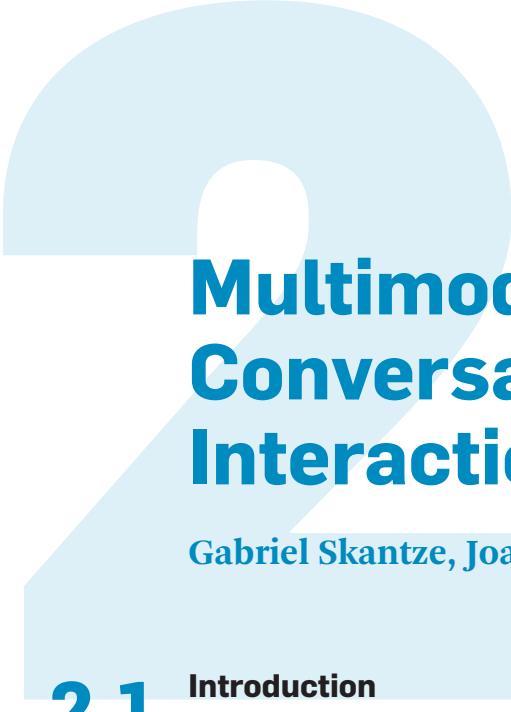
- X. Ma and E. Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the ACL*. pp. 1064–1074. Berlin, Germany. DOI: [10.18653/v1/P16-1101](https://doi.org/10.18653/v1/P16-1101). **53**
- A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the ICML 2000*, pp. 591–598. Stanford, CA. **58**
- D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago. **31, 53**
- G. Mehlmann. and E. André. 2012. Modeling multimodal integration with event logic charts. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. pp. 125–132. Santa Monica, CA. DOI: [10.1145/2070481.2070555](https://doi.org/10.1145/2070481.2070555). **47, 48, 59, 60, 61**
- M. Minsky. 1974. *A framework for representing knowledge*. MIT-AI Laboratory Memo 306. <http://web.media.mit.edu/~minsky/papers/Frames/frames.html>. Accessed June 17 2017. **28, 773**
- M. Mohri, F. C. N. Pereira, and M. Riley. 1998. A rational design for a weighted finite-state transducer library. *Lecture Notes in Computer Science*, 1436: 144–158. DOI: [10.1007/BFb0031388.pdf](https://doi.org/10.1007/BFb0031388.pdf). **45**
- L-P. Morency, C. Sidner, C. Lee, T. Darrell. 2007. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171: 568–585. DOI: [10.1016/j.artint.2007.04.003](https://doi.org/10.1016/j.artint.2007.04.003). **53**
- J. G. Neal and S. C. Shapiro. 1991. Intelligent multi-media interface technology. In J. W. Sullivan and S. W. Tyler, editors. *Intelligent User Interfaces*. Addison Wesley, New York. pp. 45–68. DOI: [10.1145/107215.128690](https://doi.org/10.1145/107215.128690). **32**
- M. J. Nederhof. 1997. Regular approximations of CFLs: A grammatical view. In *Proceedings of the International Workshop on Parsing Technology*. pp. 159–170, Boston, MA. **45, 63**
- T. Nishimoto, N. Shida, T. Kobayashi, and K. Shirai. 1995. Improving human interface in drawing tool using speech, mouse, and keyboard. In *Proceedings of the 4th IEEE International Workshop on Robot and Human Communication, ROMAN95*. pp. 107–112. Tokyo. DOI: [10.1109/ROMAN.1995.531944](https://doi.org/10.1109/ROMAN.1995.531944). **31**
- Openstream 2018, EVA:Enterprise Virtual Assistant. [www.openstream.com](http://www.openstream.com). Accessed August 31, 2018. **24**
- S. Oviatt and R. VanGent. 1996. Error resolution during multimodal human-computer interaction. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. pp. 204–207. Philadelphia, PA. DOI: [10.1109/ICSLP.1996.607077](https://doi.org/10.1109/ICSLP.1996.607077). **31, 606**
- S. L. Oviatt. 1997a. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*. 12(1): 93–129. DOI: [10.1207/s15327051hci1201&2\\_4](https://doi.org/10.1207/s15327051hci1201&2_4). **59**
- S. Oviatt, A. DeAngeli, and K. Kuhn. 1997b. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the SIGCHI*

*Conference on Human Factors in Computing Systems, CHI '97.* pp. 415–422, New York.  
 DOI: [10.1145/258549.258821](https://doi.org/10.1145/258549.258821). 38

- S. L. Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems: CHI'99*, Pittsburgh, PA. pp. 576–583. DOI: [10.1145/302979.303163](https://doi.org/10.1145/302979.303163). 31, 32, 36, 46, 56
- S. Oviatt and P. Cohen. 2000. Perceptual User Interfaces: Multimodal Interfaces that process what comes naturally. *Communications of the ACM* 43.3, pp. 45-53. 56
- F. C. N. Pereira and M. D. Riley. 1997. Speech recognition by composition of weighted finite automata. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge, MA. pp. 431–456. 44
- C. Pollard and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Center for the Study of Language and Information, University of Chicago Press, Chicago, IL. 27, 30, 33, 36, 39, 771, 786
- G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior. 2003. Recent advances in the automatic recognition of audio-visual speech. In *Proceedings of the IEEE* 91:9, pp. 1306–1326. DOI: [10.1109/JPROC.2003.817150](https://doi.org/10.1109/JPROC.2003.817150). 24
- G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Loumbaroulis, A. Vartholomaios, S. Thermos. 2017. Audio and visual modality combination in speech processing applications. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces: Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool Publishers, San Raphael, CA. DOI: [10.1145/3015783.3015797](https://doi.org/10.1145/3015783.3015797). 24
- L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson. 1978. Considerations in dynamic time-warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ICASSP-26, October 1978. DOI: [10.1109/TASSP.1978.1163164](https://doi.org/10.1109/TASSP.1978.1163164). 58
- O. Rambow, S. Bangalore, T. Butt, A. Nasr, and R. Sproat. 2002. Creating a finite-state parser with application semantics. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Taipei. pp. 1–5. DOI: [10.3115/1071884.1071910](https://doi.org/10.3115/1071884.1071910).
- G. Riccardi, R. Pieraccini, and E. Bocchieri. 1996. Stochastic automata for language modeling. *Computer Speech and Language*, 10:(4): 265–293. DOI: [10.1006/csla.1996.0014](https://doi.org/10.1006/csla.1996.0014). 44
- E. Roche. 1999. Finite-state transducers: parsing free and frozen sentences. In A. Kornai, editor, *Extended Finite-State Models of Language*. Cambridge University Press, Cambridge, UK. pp. 108–120. 44
- A. L. Rosenberg. 1967. Multi-tape finite automata with rewind instructions. *Journal of Computer and System Sciences*, 1(3): 299–315. 45
- A. Rudnicky, and A. Hauptman. 1992. Multimodal interactions in speech systems. In M. Blattner & R. Dannenberg, editors, *Multimedia Interface Design*. pp. 147–172. New York: ACM Press. 31

- E. Selfridge and M. Johnston. 2015. Interact: tightly coupling multimodal dialog with an interactive virtual assistant. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI)*, Seattle, WA. pp. 381–382. DOI: [10.1145/2818346.2823301](https://doi.org/10.1145/2818346.2823301). 23, 52
- R. Sharma, M. Yeasin, N. Krahnstover, I. Rauschert, G. Cai, I. Brewer, A. M. MacEachren, K. Sengupta. 2003. Speech-gesture driven multimodal interfaces for crisis management. In *Proceedings of the IEEE*. 91(9): 1327–1354. DOI: [10.1109/JPROC.2003.817145](https://doi.org/10.1109/JPROC.2003.817145). 24, 32
- M. Steedman. 1996. *Surface Structure and Interpretation*. MIT Press, Cambridge, MA. 39
- A. J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory* IT-13: 260–269. DOI: [10.1109/TIT.1967.1054010](https://doi.org/10.1109/TIT.1967.1054010). 58
- M. T. Vo and A. Waibel. 1997. *Modeling and Interpreting Multimodal Inputs: A Semantic Integration Approach*. CMU Technical Report. CMU-CS-97-192. 57, 60, 61
- M. T. Vo. 1998. *A Framework and Toolkit for the Construction of Multimodal Learning Interfaces*. Ph.D. Thesis, Carnegie Mellon University, CMU-CS-98-129. 53, 57, 58, 61
- Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomrygiannakis, R. Clark, and R. A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech 2017*. pp. 4006–4010. 62
- K. Wauchope. 1994. *Eucalyptus: Integrating Natural Language Input with a Graphical User Interface*. Naval Research Laboratory, Report NRL/FR/5510-94-9711.
- W. Wahlster. 2006. (editor) *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer. 23, 52
- A. Waibel, M. Vo, P. Duchnowski, S. Manke. 1996. Multimodal interfaces. *AI Review Journal*, pp. 299–319. 33
- S. Watt, T. Underhill, Y-M. Chee, K. Franke, M. Froumentin, S. Madhvanath, J-A. Magana, G. Pakosz, G. Russell, M. Selvaraj, G. Seni, C. Tremblay, L. Yaeger. September 2011. *Ink Markup Language (InkML)*. W3C Recommendation. <https://www.w3.org/TR/2011/REC-InkML-20110920/>. 42
- I. H. Witten and E. Frank. 2009. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann. 54, 55
- K. Wittenburg, L. Weitzman, and J. Talley. 1991. Unification-based grammars and tabular parsing for graphical languages. *Journal of Visual Languages and Computing*, 2:347–370. DOI: [10.1016/S1045-926X\(05\)80004-7](https://doi.org/10.1016/S1045-926X(05)80004-7). 39
- K. Wittenburg. 1993. F-PATR: Functional constraints for unification-based grammars. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pp. 216–223. DOI: [10.3115/981823.981843](https://doi.org/10.3115/981823.981843).

- W. A. Woods. 1970. Transition network grammars for natural language analysis. *Communications of the ACM*, Columbus, OH. 13 (10): 591–606. DOI: [10.1145/355598.362773](https://doi.org/10.1145/355598.362773). [26](#), [49](#), [762](#)
- M. Worsley and M. Johnston. 2010. Multimodal interactive spaces: MagicTV and MagicMAP. In *Proceedings of the IEEE Spoken Language Technology Workshop*, Berkeley, CA. pp. 161–162. DOI: [10.1109/SLT.2010.5700841](https://doi.org/10.1109/SLT.2010.5700841). [24](#)
- L. Wu, S. L. Oviatt, and P. R. Cohen. 1999. Multimodal integration—A statistical view. *IEEE Transactions on Multimedia*, 1(4): 334–341. DOI: [10.1109/6046.807953](https://doi.org/10.1109/6046.807953). [36](#), [55](#), [56](#), [60](#), [61](#)
- L. Wu, S. L. Oviatt, and P. R. Cohen. 2002. From members to teams to committee—A robust approach to gestural and multimodal recognition. In *Proceedings of the IEEE Transactions on Neural Networks* 13(40): 72–82. DOI: [10.1109/TNN.2002.1021897](https://doi.org/10.1109/TNN.2002.1021897). [36](#), [53](#), [55](#), [56](#), [60](#), [61](#)
- D. H. Younger. 1967. Recognition and parsing of context-free languages in time n<sup>3</sup>. *Information Control*, 10(2): 189–208. DOI: [10.1016/S0019-9958\(67\)80007-X](https://doi.org/10.1016/S0019-9958(67)80007-X). [38](#)



# Multimodal Conversational Interaction with Robots

Gabriel Skantze, Joakim Gustafson, Jonas Beskow

## 2.1

### Introduction

Being able to communicate with machines through spoken interaction has been a long-standing vision in both science fiction and research labs. In the 1968 movie *2001: A Space Odyssey*, the crew on board a spaceship on its way to Jupiter communicates with the ship's computer, called *Hal*, by means of spoken language. Another example is the 2013 movie *Her* in which a man develops a relationship with his intelligent operating system, personified through a female voice called *Samantha*. It is interesting to note that none of these intelligent agents are equipped with a face or a body. *Hal* only stares at the interlocutor with his (now emblematic) red eye. For a long time, spoken dialogue systems developed in research labs and employed in the industry also lacked any physical embodiment. One reason for this was of course that they were intended to be used over telephone, but it was perhaps also not clear what a physical body would actually add to the spoken interaction. Even today's smartphone assistants such as Apple's Siri and Google Now, are typically faceless.

In this chapter we focus on an important emerging application area for conversational interfaces: human-robot interaction. Whereas robots used to be found in factories and doing "3D tasks" that were seen as Dangerous, Dull, Dirty [Braybrook 2004], there is now a lot of effort in developing social robots that are able to perform "3C tasks" that require human-robot communication, coordination, or collaboration [Takayama et al, 2008]. These robots will be found in supermarkets, schools, small-scale manufacturing industry, and the homes of people. These robots should

be able to help humans in their everyday lives, and collaborate with them on complex tasks. It is quite clear that face-to-face conversation will be the most natural means for this interaction. To accomplish this, we must understand how this type of interaction differs from traditional speech interfaces. That is what we will discuss in this chapter.

One clear difference between the traditional speech interface and the emerging face-to-face interface is of course the importance of the visual channel, which is needed for detecting the faces of the users, but possibly also for other things, such as objects that might be under discussion. This points toward another important aspect that needs to be taken into account, namely the *situation* in which the interaction takes place. Unlike traditional speech interfaces, robots may interact with several users at the same time (so-called *multi-party* interaction), and the discussion might involve objects in the shared space. In such settings, it is important to keep track of where potential interlocutors are located, and whether they are currently involved in the conversation.

If we want to advance the spoken interaction toward a conversation and beyond mere command-and-control, it is also important to not only model the *verbal* aspect of the interaction (i.e., the words being spoken), but also the *non-verbal* aspects. These are obviously present in the visual channel (such as gaze and facial expressions), but also in the auditory channel, in the form of intonation, stress, laughter, etc. In this chapter, we will focus on how verbal and non-verbal signals are used in conversation to *coordinate* the interaction. We will put less focus on aspects that are common to all speech interfaces. For a comprehensive overview of conversational system in general, we refer to McTear et al. [2016]. We will start with a discussion of why the face is important in spoken interaction, and what the consequences are for the design of the robot's face. Using a concrete interaction scenario as an example, we will then discuss which sensors and processing steps are needed to build a system for human-robot interaction. Next we will discuss in more depth three important aspects of the coordination that needs to be achieved, where both the face and the voice play important roles. The first such important aspect is *turn-taking*, that is, how the robot should process the multimodal signals from the users to decide when it is and is not supposed to speak, and how it should display appropriate signals to facilitate smooth turn-taking. The second aspect is *grounding*, that is, to establish mutual understanding. This involves both picking up multimodal feedback from the users (in the form of speech and facial expressions), but also to give such feedback back to the users. Finally, we discuss the coordination of *joint attention*, for which the face also plays an important role.

**Glossary**

**Addressee detection** is the detection of who a user is addressing when speaking (the system/robot or another user).

**Backchannel** is a brief feedback (a very short utterance like “mhm” or a gesture such as a head nod) that the listener gives without intending to take the floor.

**Beat gestures** refers to rhythmic gestures, which could for example be used to emphasize a particular word or syllable.

**Deictic gestures** are pointing gestures which single out an object of interest (usually with the index finger, but also with gaze). These may be accompanied with a *deictic expression* (or *deixis*), such as “this,” “that,” “these,” or “those.”

**Echo cancellation** is the signal processing step of removing the system’s (robot’s) own speech (and its room-acoustic echoes) from the audio that comes into the microphones.

**Emblem** is a symbolic gesture, such as the thumbs up gesture, where the gesture bears no direct resemblance to what it signifies.

**Grounding**, as defined by [Clark \[1996\]](#), is the process by which speakers reach a common understanding (adding to their *common ground*). This should not be confused with the notion of *symbol grounding*, which is sometimes used in robotics, and denotes the problem of how words should be linked to objects in the real world.

**Iconic gesture** is a gesture which bears a resemblance to what it signifies, such as showing a round shape by tracing a circle in the air.

**Joint attention** means that the speakers are attending to the same object and are mutually aware of it.

**Microphone array** is a technology where several microphones are used, which allows for *sound source localization* (determine where the sound is coming from) and *beamforming* (focusing the signal on sound from a certain direction).

**Mona Lisa effect** is the phenomenon that when a 3D object (such as a face) is projected on a 2D display, all observers in the room will perceive the object to have the same rotation relative to them, no matter where they are located.

**Noise suppression** is a signal processing method which reduces unwanted sound (such as traffic noise) from the signal.

**Non-verbal communication** is mediated by other signals than words. In spoken language, this refers to aspects such as prosody, breathing, and laughter. In face-to-face interaction, non-verbal signals are also conveyed in the visual channel, such as facial expressions and gaze.

**Glossary (continued)**

**Prosody** refers to the elements of speech that are not the individual phonetic segments (that make up syllables and words). The three main prosodic elements are intonation (fundamental frequency), energy (loudness), and duration. Prosody may reflect many different things, including the emotional state of the speaker, the form of the utterance (statement or question), emphasis, contrast, and focus.

**Situation modeling** is the modeling of the current physical situation: which speakers are involved and which roles they have, as well as any objects or spatial elements that might be of interest.

**Statecharts**, or *Harel statecharts* [Harel 1987], is an extension of finite state machines, which allows for hierarchical and orthogonal states, as well as parallel execution, which is especially suitable for complex event-driven real-time applications. Examples of applications include the SCXML standard, developed by W3C (see also Chapter 1 and Chapter 9), as well as the IrisTK framework [[Skantze and Moubayed 2012a](#)].

**Turn-taking** is the process by which speakers take turns (speaking and listening), i.e., sequencing utterances and managing the floor.

**Verbal** communication is mediated by words. For example, a written chat is purely verbal (with the possible exception of emoticons and the like). Spoken language has a verbal and a non-verbal component.

**Voice activity detection** (VAD) is the binary detection of speech activity vs. silence or noise.

By coordinating their attention, the speakers can more easily refer to objects in the shared physical space and collaborate.

## 2.2

### The Importance of the Face in Interaction

The human face is one of the most important channels for social communication cues, and humans have specialized neural processes and areas of the brain specifically dedicated to the processing of faces [[Kanwisher et al. 1997](#)]. The face (and the upper body) serves a large range of functions in social communication, which are relevant for human-robot interaction. First of all, the face provides the speaker with an *identity*. Determining the identity of a person from the face is typically more reliable than other sources such as voice or gait [[Bruce 1996](#)]. This helps the speaker to keep track of different interlocutors, both during one interaction session and across sessions. A second important function of the face, that will not be discussed

in more detail here, is to express *emotions*. Studies have shown that facial expressions of basic emotions—such as anger, disgust, fear, happiness, sadness, and surprise—are similar across all cultures, and thus deeply rooted in our evolution [Ekman and Friesen 1971].

A third important feature of the face is *gaze*. Since the eyes are always directed toward the speaker's current visual focus of attention, they allow the interlocutor to get immediate access to where the attention is directed, and thereby get insight into an important aspect of the speaker's mind. The gaze direction can be used by the listener to infer the target of referring expressions, but it is also used for regulating the interaction, and for signaling status and intimacy [Argyle and Cook 1976]. Tomasello et al. [2007] has argued that this signal is so important for human interaction that the whiteness of the sclera in the human eye has evolved to facilitate this mind-reading. Humans are able to detect the eye gaze direction of another human with remarkable precision—down to a visual angle of few degrees [Bock et al. 2008]. Most easily detected are direct-gaze cues (i.e. eyes-forward, direct gaze). People are especially sensitive to eye contact since this cue has particular meaning—the so-called *stare in the crowd-effect* states that eyes-forward, direct gaze is more easily detected than averted gaze [von Grünau and Anston 1995]. It is important to note, however, that the gaze target is not only determined by observing the eyes—it is the combined vector of posture, head pose, and eye direction. Other cues, such as pointing, can also be taken into account [Langton et al. 2000].

A fourth feature of the face is the *visual articulation*. When we speak, we continuously move the articulators in order to produce sound. While a large portion of the speech production mechanism is hidden, there is still substantial information provided about the speech that can be deduced from observing the motion of the visual articulators (most importantly the lips, jaw, and tongue). The visual speech information will increase intelligibility of the speech signal, especially when the auditory signal is degraded. This is of course well known for hearing impaired people, who often deliberately rely on this information, but it can also be used by people with normal hearing under noisy conditions [Sumby and Pollack 1954]. Another testament to the strong influence of visual speech and the multimodal nature of speech perception is the McGurk effect [McGurk and MacDonald 1976]: conflicting audio-visual speech stimuli can give rise to a percept that is present in neither of the modalities. For example, an auditory /ba/ presented with a visual /ga/ was perceived by a majority of the subjects as /da/.

Finally, natural speech is also typically accompanied by a range of gestures which involve motion of the head, facial motion (e.g., eyebrows), eye-blanks, as well as upper-body gestures and body posture shifts. **Beat** gestures refer to rhythmic

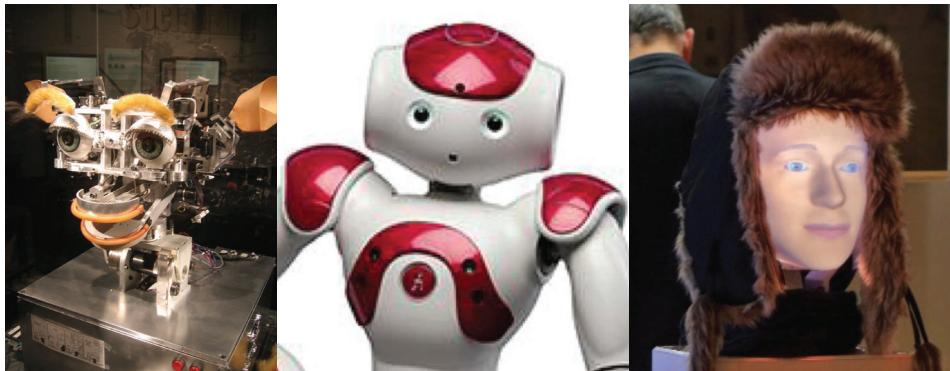
motion that occurs in synchrony with the syllabic structure and sometimes have the effect of emphasizing a particular word or syllable. In a study by Al Moubayed et al. [2010], it was shown that adding head nods or eyebrow motion to prominent syllables in an animated talking head led to increased intelligibility of the sentence. *Deictic* gestures are used when referring to entities in the physical space and can be performed with hands, head, gaze, or a combination. They are also synchronized with speech, but are typically looser than the speech gestures. Finally, *emblem* and *iconic gesture* refer to the semantic content (the former in terms of lexicalized signs, such as the thumbs-up gesture, and the latter in descriptive terms such as showing a round shape by tracing a circle in the air). These can be used to complement, emphasize, or replace part of the spoken message.

## 2.3

### Giving the Robot a Face

The simplest form of embodiment that can still exploit many of the important cues in face-to-face interaction discussed above is a virtual animated character, or embodied conversational agent (ECA), shown on a display [Cassell et al. 2000]. Since screens are readily available everywhere, it is often the most low-cost solution. It is also extremely flexible in that characters are rendered using computer animation, which places no restrictions on physical design or expressiveness of the character. However, a main drawback of using a screen-based agent is that it does not share the same physical space as the user, which means that it is impossible to infer the agent's gaze target in the user's physical space—the so-called *Mona Lisa effect* [Al Moubayed et al. 2012]. Either everyone in the room will think that the agent is looking at them, or nobody will. Consequently, it will not handle multi-party interaction or references to objects in physical space in a graceful way. Thus, interacting with such agents can be thought of as having a video conversation, which is typically perceived as inferior to a physical meeting [Nguyen and Canny 2005].

A physical robot, as opposed to a virtual character, will not suffer from the Mona Lisa effect, at least if the robot has a face with articulated eye and head, such as the *Kismet* robot head [Breazeal 2003]; see Figure 2.1. However, many of the other visual cues of the face are difficult, if not impossible, to implement with a mechatronic system. Lip motion is one example of this, where designing a mechatronic system capable of motion rapid enough for lip synchronization has yet to be done. In addition, the servos used in robotic faces make noise when they move (in contrast to the muscles in a human face), so as a side effect the visual cues will compete with the auditory cues. Many robot heads, such as the NAO robot shown in Figure 2.1, typically lack any human-like eye gaze, lip movement, or other important facial



**Figure 2.1** Examples of robot heads in research and industry: Kismet, NAO, and Furhat.

features. Apart from the challenge of implementing these features in mechatronic faces, designers sometimes also want to avoid the *uncanny valley* [Mori 1970]—the phenomenon that nearly (but not perfectly) human-like faces might be perceived as creepy.

An elegant way to circumvent both the Mona Lisa effect and the need for mechatronic solutions for facial expression is to use optically projected computer animation. Al Moubayed et al. [2012] carried out an experiment where a computer-animated avatar head was projected onto (1) a physical mannequin head and (2) a flat surface. The avatar would shift its gaze around, and a group of five subjects seated at different viewing angles were asked to indicate who the avatar was looking at. Results showed that projecting the face onto a 3D facial shape effectively alleviated the Mona Lisa effect and allowed gaze to be faithfully communicated by the avatar. This method can be exploited in the design of a robot head by placing a small projector behind a translucent face-shaped mask. Several systems have been built based on this technique, such as the Light Head [Delaunay et al. 2009], Mask Bot [Kuratake et al. 2011], and Furhat [Al Moubayed et al. 2013], as seen in Figure 2.1. These retro-projected facial displays are typically combined with a motorized neck that allows them to move like a human head, thus effectively combining the advantages of mechatronic and animated heads.

## 2.4

### Modeling Human-Robot Interaction

Numerous settings for spoken human-robot interaction have been studied where the task of the robot is to do surveys in public spaces [Skantze et al. 2012b], serve as bartender [Foster et al. 2012], take care of elderly people [Roy et al. 2000], and

act as a shopkeeper [Liu et al. 2014]. Here, we will briefly describe one such setting, which illustrates several of the key concepts discussed in this chapter. The system (described in detail in Skantze et al. [2015]), was exhibited during a week at the Swedish National Museum of Science and Technology in November 2014. As can be seen in Figure 2.2, two visitors at a time could play a collaborative game together with the Furhat robot head. On the touch table between the players, a set of cards are shown. The two visitors and Furhat are given the task of sorting the cards according to some criterion. For example, the task could be to sort a set of inventions in the order they were invented, or a set of animals by how fast they can run. This is a collaborative game, which means that the visitors have to discuss the solution together with Furhat. However, Furhat does not have perfect knowledge about the solution. Instead, Furhat's behavior is motivated by a randomized belief model. This means that the visitors have to determine whether they should trust Furhat's belief or not, just like they have to do with each other. Thus, Furhat's role in the interaction is similar to that of the visitors, as opposed to, for example, a tutor role which is often given to robots in similar settings.

For a robot to fully engage in face-to-face interaction, the underlying system must be able to perceive, interpret, and combine a number of different auditory and visual signals, as well as displaying these signals in the robot's voice and face. Building such systems is indeed a challenging task and involves not just different sensors, actuators, and software components, but also requires an interdisciplinary understanding of the problem. Thus, some kind of software framework is typically needed. In the robotics community, a well-established framework is Robot Operating System (ROS), although it does not specifically target conversational interaction. To implement the card-sorting game described previously, the open-source framework IrisTK [Skantze 2016, Skantze and Moubayed 2012a] was used, which is especially targeted toward modeling situated face-to-face interaction. Figure 2.3 schematically illustrates the necessary building blocks for a human-robot interaction system. In this the rest of this section, we will outline some of the fundamental steps in this process.

### 2.4.1 The Auditory Channel: Speech

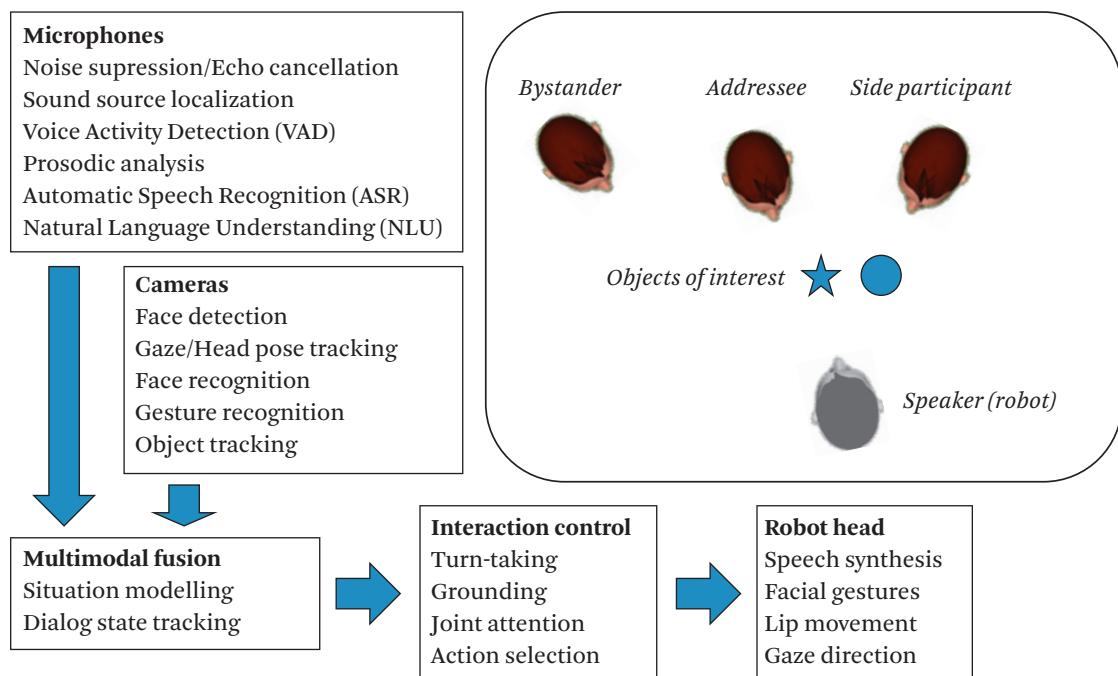
Compared to a voice assistant in a mobile phone, many human-robot interaction settings pose several additional challenges for the speech processing, as listed in Figure 2.3. The first issue is how to capture the speech from the users. In the card-sorting game, two close-talking microphones were used together with two parallel speech recognizers, allowing the robot to understand the users even when they are talking simultaneously. This worked fairly well since the users were sitting.



- U-1. I wonder which one is the fastest [looking at table]
- U-2. I think this one is fastest, what do you think? [looking at robot]
- R. I'm not sure about this, but I think the lion is the fastest animal
- U-1. Okay [moving the lion]
- R. Now it looks better
- U-2. Yeah . . . How about the zebra?
- R. I think the zebra is slower than the horse. What do you think? [looking at U-1]
- U-1. I agree

**Figure 2.2** Two children playing a card-sorting game with the robot Furhat (U-1 and U-2 denote the two users). A video showing the interaction can be seen at <http://www.youtube.com/watch?v=5fhjuGu3d0I>.

However, in many other scenarios distance microphones would typically be needed. If there are several users, the system then needs to detect where the sound is coming from. This can either be achieved by placing binaural microphones in physical ears on the robot's head [Hornstein et al. 2006] or by using a fixed *microphone array* in the vicinity of the robot [Valin et al. 2003]. The most commonly used solution in recent years (and supported by IrisTK) is the Kinect sensor from Microsoft, which combines a four-microphone array with a depth sensor and color camera. These sensors enable full-body 3D motion capture, face detection, and assigning sound sources to detected user bodies. The Kinect also features on-board *noise suppression* and *echo cancellation* that provide a relatively clean signal.



**Figure 2.3** Important processing steps in a system for face-to-face interaction.

The next thing an artificial conversational partner needs to do is to distinguish between speech and non-speech sounds. A push-to-talk mechanism, often used in smartphone applications, is typically not viable for human-robot interaction. Thus, **voice activity detection** (VAD) is needed, which is typically based on features like energy, periodicity, dynamics, and zero-crossing rates. In noisy environments, the detection can be improved by adding visual features that capture mouth movements [Almajai and Milner et al. 2008]. In a multi-party setting, not all spoken utterances are directed to the robot. Today's commercial virtual assistants (such as Apple's Siri, Google Now, and Amazon echo) have solved this by only listening to utterances that begin with certain keywords such as "Hey Siri" or "Alexa" [Sun et al. 2017]. This works in question-answering systems, but it is not really usable in more conversational settings. We will come back to this issue of *addressee detection* in Section 2.5.

In more conversational settings, such as the card-sorting game, automatic speech recognition (ASR) is a challenging task, partly because the speech is often disfluent (with truncated words, etc.), and partly because the users speak faster

with more phonetic reductions. Typically, ASR systems are trained on read speech, or command-based interaction. Thus, although the cloud-based recognizer used in the system in other settings performs very well, the performance in this more conversational setting was sometimes quite poor. Therefore, deeper syntactic parsing was not applicable here. Instead, during the discussion phase, the system used robust phrase spotting, for example, to discover whether one or more cards were mentioned. Thanks to the multimodal setting, it was possible for the system to combine this information with the movement of the cards on the table, which provided redundant information about which objects were currently being discussed. Therefore, the system could most often provide appropriate contributions to the discussion [Skantze et al. 2015]. Another interesting observation was that utterances directed toward the robot contained fewer errors, on average. This is most likely because users tended to speak more clearly toward the robot.

As discussed in the beginning of this chapter, the voice can be used to carry both verbal and non-verbal information. Examples of non-verbal communication include laughter and sighs that convey the speaker's emotional state. In spoken interaction, the meaning is a combination of the actual words that are uttered and the way these are spoken. *Prosody* is concerned with the way units beyond individual phonetic segments are spoken, and it contributes to functions such as intonation, stress, and rhythm. The voice quality and utterance melody also convey personality traits, affectional state, and willingness to talk. These non-verbal aspects are also important for coordinating the interaction, as will be seen in Section 2.5. Most speech recognizers only analyze the words that are spoken, and ignore non-verbal aspects of the speech signal. Thus, prosodic analysis (pitch, duration, and energy) typically has to be done by a separate component. Since prosodic features, such as pitch, differ between individuals (depending on gender, age, etc.), the analysis has to be normalized for the user.

## 2.4.2 The Visual Channel: Faces

In recent years, computer vision and sensing techniques for facial processing have reached a level of robustness and performance that makes them useful in interactive systems. This is partly thanks to a rapid evolution in pure vision-based approaches that take advantage of new machine learning techniques and large amounts of data. There has also been something of a revolution on the sensor side, with inexpensive depth sensors, as pioneered by the Kinect sensor from Microsoft in 2010. Regardless of the type of sensor used, there are five main categories of technologies that are relevant and that expand on the facial functions used in social interaction as discussed in Section 2.2 above. Open-source implementations

exist, for example in the OpenCV [[Itseez 2016](#)] image processing toolkit, as well as in commercial software packages.

*Face detection* algorithms have been commonplace for approximately the past decade. The method described by [Viola and Jones \[2001\]](#) was the first to be fast and robust enough to be implemented on consumer devices and face detection is now a standard feature in such technologies as digital cameras and smartphones. The facial detection is typically the first step in the facial processing chain as it defines the region of interest in the image for further processing.

*Head-pose tracking* and *gaze tracking* are important techniques for deducing the visual attention of the user and are a prerequisite for accurate handling of turn-taking by the system. In estimating focus of visual attention, head pose may work well as a proxy for gaze for practical purposes, but one should be aware that this is not always accurate. For gaze-tracking, several options exist with different accuracy and hardware demands. Vision-only methods can work under favorable circumstances (lighting conditions, etc.), but more accurate measurements require dedicated gaze tracking-hardware. Such hardware is available both in the form of head-worn glasses and stationary devices placed in front of the user. One caveat is that gaze trackers have to be calibrated for the user, making them difficult to employ in applications where one expects different users to interact with the system, and are therefore typically mostly used in research for controlled lab experiments.

*Face recognition* refers to identification or verification of a face from an image. The technique has been developed primarily with security applications in mind, but is also commonly found in technology such as photo album software. This can be an important feature in social human-robot interaction since it allows for an effective and seamless way for the system to keep a memory of who it has been interacting with, and about what.

*Facial feature tracking* refers to methods for tracking the location and shape of individual features of the face such as the contour of the lips, the jaw line, eyes, nose, and eyebrows. This type of feature can be used as input to systems detecting emotion, affective state, or facial gestures (such as brow raise).

For an in-depth description of how the auditory and visual channels can be combined to enhance speech processing, see Chapter 11, [[Potamianos et al. 2017](#)].

### **2.4.3 Modeling the Situation and Controlling the Interaction**

As illustrated in Figure 2.3, the information from the auditory and visual channels needs to be fused into a model of the users' behavior and the current situation. An example of this is the card-sorting game previously presented. A Kinect camera is used to track the location and rotation of the two users' heads, as well as their

hands. This data, together with the position of the five cards on the touch table, are sent to a *situation model*. The Situation model takes low-level events from the different sensors (Kinect, Touch table, and ASR), creates a 3D representation of the situation, and then generates high-level events for the combined sensory data. The individual sensors themselves only provide information relative to their own coordinate systems, but since the Situation model keeps information about the position of the different sensors, it can then translate these coordinates into a common 3D space. Also, if there are several sensors tracking the same users and objects, the Situation model can merge these streams into one coherent model, and map sensory events to a common set of user IDs. This way, speech recognition results from the microphones can be mapped to the right users based on their location, regardless of whether it is a microphone array or a close-talking microphone. Another task of the Situation model is to keep track of when users enter and leave the interaction spaces of system agents, and generate appropriate engagement and disengagement events.

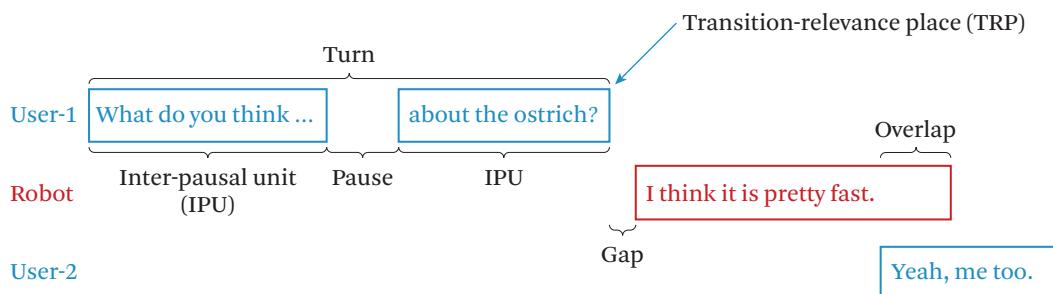
There are many different models for dialogue management. For simpler interactions, finite state machines or form-filling models may suffice, but there are also more complex planning-based accounts [McTear et al. 2016]. To optimize the system's policies, reinforcement learning is often used [Rieser and Lemon 2012]. In the card sorting game, a dialogue model based on *statecharts*, called IrisFlow, was used. Two such behavior modules were used in parallel: one for dialogue management and one for maintaining Furhat's attention. The Dialogue Flow module orchestrates the spoken interaction, based on events from the Situation model, such as someone speaking, shifting attention, entering or leaving the interaction, or moving cards on the table. The Attention Flow keeps Furhat's attention to a specified target (a user or a card), even when the target is moving, by consulting the Situation model. The 3D position of the target is then transformed into neck and gaze movement of Furhat (again taking Furhat's position in the 3D space into account).

Next, we will discuss three important issues related to the control of the interaction, all of which need to take the multimodal nature of the face-to-face setting into account: turn-taking, grounding, and joint attention.

## 2.5

### Turn-Taking

Many human social activities require some kind of *turn-taking* protocol, which determines the order in which the different actions are supposed to take place, and by whom. This is obvious when, for example, playing a game of chess (where



**Figure 2.4** Important concepts when modeling turn-taking.

the protocol is very simple), but it also applies to spoken interaction. Since it is difficult to speak and listen at the same time, speakers in dialogue have to somehow coordinate who is currently speaking and who is listening. In a seminal article, [Sacks et al. \[1974\]](#) describe a protocol for this. In their view, speakers try to minimize the amount of gaps and overlaps between turns. At certain points in the speech, there are *Transition-Relevance Places* (TRPs), where a shift in turn could potentially take place. At these places, the current speaker may select a next speaker, who then “has the right and is obliged” to take the next turn to speak, whereas no other participants are supposed to do so. If the current speaker does not select a next speaker, any participant has the opportunity to “self-select”, or the current speaker may continue. In a two-party (*dyadic*) conversation, the selection of the next speaker is trivial, but if there are several speakers (*multi-party* interaction), the current speaker typically gazes at the selected next speaker, or may use some other indicator, such as the person’s name or a pointing gesture.

Some important concepts in this process are illustrated in Figure 2.4. From a computational perspective, a useful term is *Inter-pausal unit* (IPU), which is a stretch of audio from one speaker without any silence exceeding a certain amount (such as 200 ms). These can relatively easily be identified using voice activity detection. A *turn* is then defined as a sequence of IPUs from a speaker, which are not interrupted by IPUs from another speaker. The term *utterance* is used often also, but its definition is often much vaguer, since it is typically linked to the actual semantic content of the speech (see [Traum and Heeman \[1997\]](#) for a discussion).

Traditionally, spoken dialogue systems have rested on a very simplistic model of turn-taking, where a certain amount of silence (say 700 ms) is used as an indicator that the user has stopped speaking, and that the turn is yielded to the system. Essentially, each user turn is considered as one IPU with a fairly long pause thresh-

**Table 2.1** Turn-yielding and Turn-holding cues

	Turn-yielding cue	Turn-holding cue
Syntax	Complete	Incomplete, Filled pause
Prosody - Pitch	Rising or Falling	Flat
Prosody - Intensity	Lower	Higher
Prosody - Duration	Shorter	Longer
Breathing	Breathe out	Breathe in
Gaze	Looking at addressee	Looking away
Gesture	Terminated	Non-terminated

old. The problem with this model is that turn-shifts often are supposed to be much more rapid than this, and that pauses within a turn often might be longer. Thus, the system will sometimes appear to give sluggish responses, and sometimes interrupt the user. A more accurate model would break up the user's speech into several IPUs using much shorter pause thresholds (such as 200 ms), and then try to identify whether the user is *yielding* or *holding* the turn. But what should this decision be based on?

Several studies have found that speakers use their voice and face to give *turn-holding* and *turn-yielding cues* [Duncan 1972, Koiso et al. 1998, Gravano and Hirschberg 2011]. For example, an IPU ending with an incomplete syntactic clause or a filled pause (such as "ehm") typically indicates that the speaker is not yielding the turn. Prosodically, a rising or falling pitch at the end of the IPU tend to be turn-yielding, whereas a flat pitch is turn-holding. The intensity of the voice tends be lower when yielding the turn, and the duration of the last phoneme tends to be shorter. By breathing in, the speaker may also signal that she/he is about to speak (thus not yielding the turn) [Ishii et al. 2014]. Gaze has also been found to be an important cue—speakers tend to not look at the addressee during an utterance, but then shift the gaze toward the addressee when yielding the turn [Kendon 1967]. Gestures can also be used as an indicator, where a non-terminated gesture may signal that the turn is not finished yet. A summary of these cues is presented in Table 2.1.

It is important to note that these are just typical patterns, and that all these cues do not conform to these principles all the time. However, studies on human-human dialogue have shown that the more turn-yielding cues are presented together, the more likely it is that the other speaker will take the turn [Duncan 1972, Koiso et al. 1998, Gravano and Hirschberg 2011]. Thus, one way of deciding whether the

robot should take the turn or not after an IPU, is to use machine learning in order to weigh these different cues together [Meena et al. 2014, Johansson and Skantze 2015, Roddy et al. 2018].

As we previously discussed, in multi-party interaction the robot does not only need to find out whether a user has yielded the turn or not, but also whether it was yielded to the robot, or to somebody else. This problem is often referred to as *addressee detection*. Again, several different multimodal cues can be used for this, and be combined using machine learning [Katzenmaier et al. 2004, Vinyals et al. 2012]. The most obvious cue is perhaps gaze (or head pose as proxy), but studies have also found the user's voice characteristics to be different when talking to a machine, compared to talking to another human [Shriberg et al. 2013].

In dyadic interaction, the participants can either take on the role as *speaker* or *addressee*. As can be seen in Figure 2.3, multi-party interaction also involves *side participants*, who are neither currently speaking or being addressed. But there might also be non-participants in the vicinity, such as *bystanders* and *overhearers* [Clark 1996]. Thus, the robot will need to be able to detect the roles of humans appearing in the shared space, and be able to engage and disengage in the interaction with them [Bohus and Horvitz 2009, 2014]. In a controlled experiment, Mutlu et al. [2012] found that a robot can effectively use gaze signals to shape the roles of the participants and how they perceive the robot's role. Similarly, several studies have found that robots can actively select the next speaker using gaze, which the users typically conform to [Bohus and Horvitz 2010, Skantze 2017]. In order to avoid turn-taking confusion, especially when there are processing delays, it might also help to signal that the robot is about to speak. Skantze et al. [2015] investigated different human-like multimodal cues for this (including gaze aversion and hesitation sounds), and found that a combination of cues was most effective.

The turn-taking model of Sacks et al. [1974] has also been challenged by other researchers, who argue that speakers do not always try to minimize gaps and overlaps, but that the criteria for successful interaction is highly dependent on the kind of interaction taking place [O'Connell et al. 1990]. In this view, overlaps can be a sign of engagement, and it is possible that robots should not necessarily always avoid overlaps. Another phenomenon that poses some problems for the IPU-model outlined above is *backchannels*—short utterances (such as “mhm” or “aha”) or head nods, which the listener provides to show continued attention [Yngve 1970]. Different models have been proposed for the timing of backchannels, based on similar cues as turn-taking [Koiso et al. 1998]. However, since backchannels are so brief and unobtrusive (especially head nods), they are generally not considered to constitute a turn. Also, it is quite common that they overlap with the interlocutor's

speech. Thus, timing models for backchannels should perhaps not be based on IPUs, but rather be made as a continuous decision [Ward and Tsukahara 2000, Morency et al. 2008].

## 2.6 Grounding and Feedback

Communication can be described as the process by which we make our knowledge and beliefs *common*, adding to our *common ground*. Clark [1996] defines the notion of *common ground* between two speakers as “the sum of their mutual, common, or joint knowledge, beliefs, and suppositions.” When engaging in a dialogue, two people may have more or less in their common ground to start with. During the conversation, they try to share their private knowledge and beliefs in an attempt to add them to the common ground. As Clark [1996] points out, however, the process by which speakers add to the common ground is really a joint project, in which the speakers have to cooperatively ensure mutual understanding. A speaker cannot simply deliver a message and hope that the listener will receive, comprehend, and accept it as correct. They have to constantly send and pick up signals about the reception, comprehension, and acceptance of the information that is communicated. This is the process of *grounding*, and the signals they use in this process are often referred to as *feedback* [Allwood et al. 1992].

Clark [1996] makes a distinction between four *levels of action* that take place when a speaker is trying to communicate something to a listener. Suppose Speaker A proposes an activity for Listener B, such as answering a question or executing a command. For communication to be “successful”, all these levels of action must succeed (listed from higher to lower).

- *Acceptance*: B must accept A’s proposal.
- *Understanding*: B must understand what A is proposing.
- *Perception*: B must perceive the signal (e.g., hear the words spoken).
- *Contact*: B must attend to A.

The order of the levels is important; in order to succeed on one level, all the levels below it must be completed. Clark calls this the *principle of upward completion*. Thus, we cannot understand what a person is saying without hearing the words spoken, we cannot hear the words without attending, and so on. Now, a speaker can use feedback to signal success or problems on these different levels, as shown in Table 2.2. By giving positive feedback on one level, all levels below it are assumed to also have succeeded (due to the principle of upward completion). Thus, if a person replies “I agree”, we can assume that he (thinks that he) heard and understood

**Table 2.2** Examples of positive and negative feedback on different levels using voice and face.

Level	Positive	Negative
Contact	Backchannel (“mhm”, nod)	“are you there?”
Perception	Backchannel	Repair initiator (“huh?”), Frown
Understanding	Reprise fragment (“blue”)	Clarification request (“blue?”), Frown
Acceptance	Acknowledgment (“okay”), Agreement (“I agree”), Smile	“I don’t agree”, “I cannot find that”

what was said. In the same way, if negative feedback is given on one level, all levels above it are assumed to have failed, whereas all levels below it are assumed to have succeeded. As an example, if a person says “huh?”, he has heard that something was said (positive Contact), but has not heard what was said (negative Perception, and thereby negative Understanding and Acceptance).

Speakers frequently use (auditory or visual) backchannels to show that the communication channel is open (positive Contact). Thus, the best way to make someone stop speaking (over the telephone at least) is to be completely silent (it will not take long before the other speaker will say “are you there?”). However, backchannels can be ambiguous since it is not obvious on what level they are actually committing. Small differences in prosody can have a big effect on their perceived meaning, and they can even have a negative function (like a prolonged “yeah . . . ” with a falling pitch). In a perception experiment, [Lai \[2010\]](#) found that different intonation contours of cue words (e.g., “yeah”, “right”, “really”) influence listeners’ perception of uncertainty. [Gravano et al. \[2007\]](#) did a similar analysis of the word “okay”, and found that both prosody and dialogue context affected the interpretation of the word as either a backchannel, acknowledgment, or a beginning of a new discourse segment. In a study on human-robot interaction, [Skantze et al. \[2014\]](#) found that both the lexical choice and prosody in the users’ feedback are correlated with uncertainty, and built a logistic regression classifier to combine these features. It has also been shown how these different functions can be achieved by varying the prosodic realization when synthesizing short feedback utterances [[Waller et al. 2006](#)].

The visual channel is of course also very important to convey feedback. In an experiment on negative and positive feedback in animated agents, [Granström et al. \[2002\]](#) showed that subjects are sensitive to both acoustic and visual cues. Not surprisingly, smile, brow raise, and nodding were perceived as positive feedback whereas brow frown was interpreted as negative. However, it should be noted that

some of these gestures can also have other functions. For example, brow raise and nodding can also be used to mark prominence in the speech [Beskow et al. 2006]. It is also important that robots can not only produce such feedback, but also recognize it from the user. Sidner et al. [2006] did a study in a human-robot interaction setting where the robot could recognize head nods from the user. It was found that when the robot gave gestural feedback back to the user, the users nodded more themselves.

## 2.7

### Joint Attention

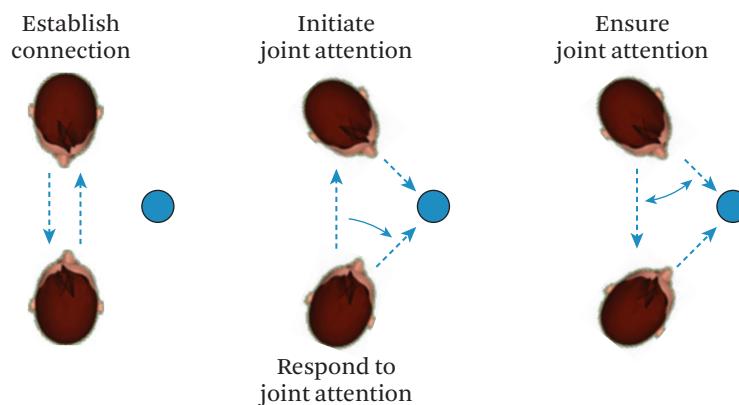
*Joint attention* is the process by which speakers coordinate their attention, in order to make sure that they attend to the same object of interest. In other words, it means that the speakers are attending to the same object and are mutually aware of it [Clark and Marshall 1981]. To achieve joint attention, a speaker first needs to alert the other speaker to an object by means of gaze, pointing gestures, or other verbal or non-verbal indicators. The other speaker then needs to detect this signal and attend to the same object. In situated interaction, speakers naturally look at the objects which are under discussion. The speaker's gaze can therefore be used by the listener as a cue to the speaker's current focus of attention. Speakers seem to be aware of this fact, since they naturally use deictic expressions (like "here" and "there") accompanied by a glance toward the object that is being referred to [Clark and Krych 2004]. In the same way, listeners naturally look at the object being referred to during speech comprehension [Allopenna et al. 1998], and their gaze can therefore be used as a cue by the speaker to verify common ground. These skills gradually develop in humans during the first two years after birth. However, some children do not develop this ability fully, which is one of the key factors in the Autism Spectrum Disorder. This is one explanation for why people on this spectrum often have difficulties in communicating with others [Baron-Cohen 1995]. Thus, failure to model joint attention in human-robot interaction could also be expected to result in communication problems.

On the most fundamental level, the robot's gaze could be used as a cue to facilitate the user's comprehension of the robot's speech (such as referring expressions). Here, it should be noted that the design of the robot puts constraints on how the robot's attention will be displayed. If the robot's cameras are located in the eyes, the eyes will naturally follow the robot's actual focus of attention. It is far from certain that this attentional behavior will be similar to that of a human, and this could potentially lead to confusion, if the human interlocutor expects it to be so. On the other hand, if the camera is not located in the eyes, the actual "attention" of the

system is decoupled from its signaled attention. If a wide-angle or multi-camera setup is used, the system could potentially “attend” to the whole visual scene all the time. In that case, the direction of the robot’s eye gaze should still be designed in order to facilitate the interaction as much as possible.

Several experiments have shown that subject can utilize the robot’s gaze to infer the target of referring expressions [Staudte and Crocker 2011, Boucher et al. 2012, Skantze et al. 2014]. However, in a study on infants, Okumura et al. [2013] found that while 12-month-olds understand the referential nature of human gaze, this is not the case for robot gaze. This suggests that humans do not naturally interpret robot gaze in the same way as the gaze of other humans, but that this anthropomorphism might be learned. The extent to which gaze following comes natural is also likely to depend on the design of the robot.

Even if the user can utilize the robot’s gaze to identify referents in the shared space, this is not in itself enough to achieve joint attention. This also requires that the robot is aware of the user’s attention and that it is able to coordinate its attention with the user. Huang and Thomas [2010] describe four necessary behaviors related to joint attention that a robot should be able to recognize and perform, illustrated in Figure 2.5. First, it must be able to establish a connection through mutual gaze, or recognize that the user is trying to establish a connection. Second, the robot should be able to either Initiate Joint Attention (IJA)—through gaze, pointing gestures, or verbal actions—or Respond to Joint Attention (RJA), depending on its role in the process. Then, it should be able to Ensure Joint Attention (EJA) by monitoring the user’s gaze.



**Figure 2.5** The process of establishing joint attention.

In Rich et al. [2010], a system capable of recognizing some of these behaviors in humans was presented. The model was used to allow a robot to play a physical puzzle game together with a human subject. Huang and Thomas [2010] explored the importance of responding to joint attention in a human-robot interaction scenario. A comparative study showed that robots responding to joint attention are experienced as more competent and socially interactive. The study also showed the importance of ensuring joint attention, which led to better task success and was perceived as more natural. The importance of establishing connection was explored by Imai et al. [2003], in a scenario where a robot and a human were discussing a poster. The robot could establish eye-contact with a human, draw the human's attention to a certain object of interest, and then adapt the utterance generation depending on the human's attention.

Another task where joint attention is important is hand-over of objects [Strabala et al. 2013]. Moon et al. [2014] did a comparative study on how the robot's gaze behavior affects the efficiency with which the robot can hand over an object to a human. It was found that the most human-like behavior—where the robot first gazes at the location where the hand-over will take place, and then up at the human when the human reaches for the object—was more efficient. In this condition, the human moved the hand to the hand-over location even before the robot's hand reached there.

It should also be noted that attention to objects in the shared space affects how gaze is used in turn-taking (for yielding the turn and selecting the next speaker). This has been shown to clearly affect the extent to which humans otherwise gaze at each other to yield the turn. In a study on modeling turn-taking in three-party poster conversations, it was found that the participants almost always looked at the shared poster [Kawahara et al. 2012]. In a Wizard-of-Oz study on multi-party human-robot interaction, where the participants discussed objects on a table between them, Johansson et al. [2013] found that turn shifts often occurred without the speakers looking at each other. It is also important to note that objects not only attract attention in conversation but that placing or moving an object also can be regarded as a communicative act in itself [Clark 2005].

## 2.8

### Conclusions

To conclude, we have discussed the importance of acoustic and visual signals in face-to-face interaction for coordinating turn-taking, giving feedback, and achieving joint attention. As we have seen, the same signal can serve different functions in different settings, and different signals can be combined to achieve various goals.

Thus, systems for human-robot interaction must not only be able to pick up these signals, but also fuse them in an intelligent way, taking the current physical situation and dialogue context into account. Also, if we want the robot to be able to convey these signals, and thereby utilize a protocol humans already know how to interpret, careful design of the robot body and the speech synthesis are required.

## References

- S. Al Moubayed, J. Beskow, and B. Granström. 2010. Auditory-visual prominence: From intelligibility to behavior. *Journal on Multimodal User Interfaces*, 3(4): 299–311. DOI: [10.1007/s12193-010-0054-0](https://doi.org/10.1007/s12193-010-0054-0). 82
- S. Al Moubayed, J. Edlund, and J. Beskow. 2012. Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems*, 1(2): 25. DOI: [10.1145/2070719.2070724](https://doi.org/10.1145/2070719.2070724). 82, 83
- S. Al Moubayed, G. Skantze, and J. Beskow. 2013. The Furhat Back-Projected Humanoid Head—Lip reading, gaze and multiparty interaction. *International Journal of Humanoid Robotics*, 10(1). DOI: [10.1142/S0219843613500059](https://doi.org/10.1142/S0219843613500059). 83
- P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4): 419–439. DOI: [10.1006/jmla.1997.2558](https://doi.org/10.1006/jmla.1997.2558). 95
- J. Allwood, J. Nivre, and E. Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1): 1–26. DOI: [10.1093/jos/9.1.1](https://doi.org/10.1093/jos/9.1.1). 93
- I. Almajai and B. Milner. 2008. Using audio-visual features for robust voice activity detection in clean and noisy speech. In *Proceedings of the 16th European Signal Processing Conference*, pp. 1–5. 86
- M. Argyle and M. Cook. 1976. *Gaze and mutual gaze*. 81
- S. Baron-Cohen. (1995). The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In C. Moore, and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development* (pp. 41–60). Hillsdale, NJ: Erlbaum. 95
- J. Beskow, B. Granström, and D. House. 2006. Visual correlates to prominence in several expressive modes. In *Proceedings of Interspeech 2006*, pp. 1272–1275. Pittsburg, PA. DOI: [10.1.1.158.9076](https://doi.org/10.1.1.158.9076). 95
- S. Bock, P. Dicke, and P. Thier. 2008. How precise is gaze following in humans? *Vision Research*, 48(7): 946–957. DOI: [10.1016/j.visres.2008.01.011](https://doi.org/10.1016/j.visres.2008.01.011). 81
- D. Bohus, and E. Horvitz. 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. In *Proceedings of SIGdial*. London, UK. DOI: [10.3115/1708376.1708411](https://doi.org/10.3115/1708376.1708411). 92

- D. Bohus and E. Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of ICMI*. Beijing, China. DOI: [10.1145/1891903.1891910](https://doi.org/10.1145/1891903.1891910). 92
- D. Bohus and E. Horvitz. 2014. Managing Human-Robot Engagement with Forecasts and . . . um . . . Hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction* pp. 2–9. 92
- J. D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey. 2012. I reach faster when I see you look: Gaze effects in human-human and human-robot face-to-face cooperation. *Frontiers in Neurorobotics*, 6. DOI: [10.3389/fnbot.2012.00003/full](https://doi.org/10.3389/fnbot.2012.00003/full). 96
- R. Braybrook. 2004. Three “d” missions—dull, dirty and dangerous. *Armada International*, 28(1): 10–12. 77
- C. Breazeal. 2003. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3): 167–175. DOI: [10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1). 82
- V. Bruce. 1996. The role of the face in communication: Implications for videophone design. *Interacting with Computers*, 8(2): 166–176. DOI: [10.1016/0953-5438\(96\)01026-0](https://doi.org/10.1016/0953-5438(96)01026-0). 80
- J. Cassell, J. Sullivan, S. Prevost, and E. F. Churchill 2000. *Embodied Conversational Agents*. Boston, MA: MIT Press. 82
- H. H. Clark and M. A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50, 62–81. DOI: [10.1016/j.jml.2003.08.004](https://doi.org/10.1016/j.jml.2003.08.004). 95
- H. H. Clark and C. R. Marshall. 1981. Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, and I. A. Sag, editors, *Elements of Discourse Understanding* (pp. 10–63). Cambridge, UK: Cambridge University Press. 95
- H. H. Clark. 1996. *Using Language*. Cambridge, UK: Cambridge University Press. 79, 92, 93, 774
- H. H. Clark. 2005. Coordinating with each other in a material world. *Discourse Studies*, 7(4–5): 507–525. DOI: [10.1177/1461445605054404](https://doi.org/10.1177/1461445605054404). 97
- F. Delaunay, J. De Greeff, and T. Belpaeme. 2009. Towards retro-projected robot faces: an alternative to mechatronic and android faces. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 306–311). DOI: [10.1109/ROMAN.2009.5326314](https://doi.org/10.1109/ROMAN.2009.5326314). 83
- S. Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23(2): 283–292. DOI: [10.1037/h0033031](https://doi.org/10.1037/h0033031). 91
- P. Ekman, and W. V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17, 124–129. DOI: [10.1037/h0030377](https://doi.org/10.1037/h0030377). 81
- M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. Petrick. 2012. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 3–10). DOI: [10.1145/2388676.2388680](https://doi.org/10.1145/2388676.2388680). 83

- B. Granström, D. House, and M. G. Swerts. (2002). Multimodal feedback cues in human-machine interactions. In B. Bel, and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 Conference* (pp. 347–350). Aix-en-Provence: Laboratoire Parole et Langage. [94](#)
- A. Gravano and J. Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3): 601–634. DOI: [10.1016/j.csl.2010.10.003](https://doi.org/10.1016/j.csl.2010.10.003). [91](#)
- A. Gravano, S. Benus, H. Chavés, J. Hirschberg, and L. Wilcox. 2007. On the role of context and prosody in the interpretation of ‘okey’. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 800–807). Prague, Czech Republic. [94](#)
- D. Harel. 1987. Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, 8, 231–274. DOI: [10.1016/0167-6423\(87\)90035-9](https://doi.org/10.1016/0167-6423(87)90035-9). [80](#)
- J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda. 2006. Sound localization for humanoid robots-building audio-motor maps based on the HRTF. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1170–1176. DOI: [10.1109/IROS.2006.281849](https://doi.org/10.1109/IROS.2006.281849). [85](#)
- C-M. Huang and A. Thomaz. 2010. Joint attention in human-robot interaction. In *Papers from the AAAI Fall Symposium on Dialog with Robots*, pp. 32–37. Arlington, VA. [96](#), [97](#)
- M. Imai, T. Ono, and H. Ishiguro. 2003. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transaction on Industrial Electronics*, 50(4): 636–643. DOI: [10.1109/TIE.2003.814769](https://doi.org/10.1109/TIE.2003.814769). [97](#)
- R. Ishii, K. Otsuka, S. Kumano, and J. Yamato. 2014. Analysis of Respiration for Prediction of “Who will be next speaker and when?” in multi-party meetings. In *Proceedings of ICMI*, pp. 18–25. New York: ACM. DOI: [10.1145/2663204.2663271](https://doi.org/10.1145/2663204.2663271). [91](#)
- Itseez. 2016. Open source computer vision library. <https://github.com/itseez/opencv>. Accessed September 2016. [88](#)
- M. Johansson and G. Skantze. 2015. Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. In *Proceedings of SIGDIAL*. Prague, Czech Republic. DOI: [10.18653/v1/W15-4642](https://doi.org/10.18653/v1/W15-4642). [92](#)
- M. Johansson, G. Skantze, and J. Gustafson. 2013. Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In *International Conference on Social Robotics-ICSR 2013*. Bristol, UK. DOI: [10.1007/978-3-319-02675-6\\_35](https://doi.org/10.1007/978-3-319-02675-6_35). [97](#)
- N. Kanwisher, J. McDermott, and M. Chun. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11): 4302–4311. [80](#)
- M. Katzenmaier, R. Stiefelhagen, T. Schultz, I. Rogina, and A. Waibel. 2004. Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In *Proceedings of International Conference on Multimodal Interfaces ICMI 2004*. State College, PA. DOI: [10.1145/1027933.1027959](https://doi.org/10.1145/1027933.1027959). [92](#)

- T. Kawahara, T. Iwatate, and K. Takanashi. 2012. Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations . . . In *Interspeech 2012*. DOI: [10.1.1.258.8592](https://doi.org/10.1.1.258.8592). 97
- A. Kendon. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26, 22–63. DOI: [10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4). 91
- H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41, 295–321. DOI: [10.1177/002383099804100404](https://doi.org/10.1177/002383099804100404). 91, 92
- T. Kuratake, Y. Matsusaka, B. Pierce, and G. Cheng. 2011. Mask-bot: a life-size robot head using talking head animation for human-robot communication. In *Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)* pp. 99–104. DOI: [10.1109/Humanoids.2011.6100842](https://doi.org/10.1109/Humanoids.2011.6100842). 83
- C. Lai. 2010. What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. In *Proceedings of Interspeech*. Makuhari, Japan. DOI: [10.1.1.185.2411](https://doi.org/10.1.1.185.2411). 94
- S. Langton, R. Watt, and V. Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences*, 4(2): 50–59. DOI: [10.1016/S1364-6613\(99\)01436-9](https://doi.org/10.1016/S1364-6613(99)01436-9). 81
- P. Liu, D. Glas, T. Kanda, H. Ishiguro, and N. Hagita. 2014. How to train your robot-teaching service robots to reproduce human social behavior. In *Proceedings of Robot and Human Interactive Communication (RO-MAN)* pp. 961–968. DOI: [10.1109/ROMAN.2014.6926377](https://doi.org/10.1109/ROMAN.2014.6926377). 84
- H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264 (5588), 746–748. DOI: [10.1038/264746a0](https://doi.org/10.1038/264746a0). 81
- M. McTear, Z. Callejas, and D. Griol. 2016. *The Conversational Interface*. Springer. DOI: [10.1007/978-3-319-32967-3\\_78](https://doi.org/10.1007/978-3-319-32967-3_78), 89
- R. Meena, G. Skantze, and J. Gustafson. 2014. Data-driven Models for timing feedback responses in a Map Task dialogue system. *Computer Speech and Language*, 28(4): 903–922. DOI: [10.1016/j.csl.2014.02.002](https://doi.org/10.1016/j.csl.2014.02.002). 92
- A. Moon, D. Troniak, B. Gleeson, M. Pan, M. Zheng, B. Blumer, K. MacLean, and E. Croft. 2014. Meet Me Where I'm Gazing: How Shared Attention Gaze Affects Human-robot Handover Timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, pp. 334–341. New York: ACM. DOI: [10.1145/2559636.2559656](https://doi.org/10.1145/2559636.2559656). 97
- L. P. Morency, I. de Kok, and J. Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of IVA*, pp. 176–190. Tokyo, Japan. DOI: [10.1007/s10458-009-9092-y](https://doi.org/10.1007/s10458-009-9092-y). 93
- M. Mori. 1970. The Uncanny Valley. *Energy*, 7(4): 33–35. 83

- B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro. 2012. Conversational Gaze Mechanisms for Humanlike Robots. *ACM Transactions on Interactive Intelligent Systems*, 1(2), 12: 1–12:33. DOI: [10.1145/2070719.2070725](https://doi.org/10.1145/2070719.2070725). 92
- D. Nguyen, and J. Canny. 2005. MultiView: spatially faithful group video conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 799–808. DOI: [10.1145/1054972.1055084](https://doi.org/10.1145/1054972.1055084). 82
- D. C. O'Connell, S. Kowal, and E. Kaltenbacher. 1990. Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6): 345–373. DOI: [10.1007/BF01068884.pdf](https://doi.org/10.1007/BF01068884.pdf). 92
- Y. Okumura, Y. Kanakogi, T. Kanda, H. Ishiguro, and S. Itakura. 2013. Infants understand the referential nature of human gaze but not robot gaze. *Journal of Experimental Child Psychology*, 116, 86–95. DOI: [10.1016/j.jecp.2013.02.007](https://doi.org/10.1016/j.jecp.2013.02.007). 96
- G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos. 2017. Audio and Visual Modality Combination in Speech Processing Applications. *Association for Computing Machinery*, pp. 489–543. Morgan & Claypool New York, NY. DOI: [10.1145/3015783.3015797](https://doi.org/10.1145/3015783.3015797). 88
- C. Rich, B. Ponsler, A. Holroyd, and C. Sidner. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 375–382. DOI: [10.1109/HRI.2010.5453163](https://doi.org/10.1109/HRI.2010.5453163). 97
- V. Rieser and O. Lemon. 2012. *Reinforcement Learning for Adaptive Dialogue Systems*. Berlin: Springer-Verlag. DOI: [10.1007/978-3-642-24942-6\\_89](https://doi.org/10.1007/978-3-642-24942-6_89)
- M. Roddy, G. Shantze, and N. Harte. 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, pp. 186–190. New York, New York, USA: ACM Press. DOI: [10.1145/3242969.3242997](https://doi.org/10.1145/3242969.3242997). 92
- N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Margaritis, M. Montemerlo, J. Pineau, and J. Schulte. 2000. Towards personal service robots for the elderly. In *Workshop on Interactive Robots and Entertainment (WIRE 2000)*, p. 184. 83
- H. Sacks, E. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696–735. DOI: [10.2307/412243](https://doi.org/10.2307/412243). 90, 92
- E. Shriberg, A. Stolcke, and S. Ravuri. 2013. Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Interspeech 2013*, pp. 2559–2563. 92
- C. Sidner, C. Lee, L.-P. Morency, and C. ForLines. 2006. The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction*, pp. 290–296. ACM Press. DOI: [10.1145/1121241.1121291](https://doi.org/10.1145/1121241.1121291). 95
- G. Skantze. 2017. Predicting and regulating participation equality in human-robot conversations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pp. 196–204. New York, New York, USA: ACM Press. DOI: [10.1145/2909824.3020210](https://doi.org/10.1145/2909824.3020210). 92

- G. Skantze, and S. Al Moubayed. 2012a. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA. DOI: [10.1145/2388676.2388698](https://doi.org/10.1145/2388676.2388698). 80, 84, 785
- G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, and B. Granström. 2012b. Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue. In *Proceedings of IVA-RCVA*. Santa Cruz, CA. DOI: [10.1.1.367.378.83](https://doi.org/10.1.1.367.378.83)
- G. Skantze, A. Hjalmarsson, and C. Oertel. 2014. Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication*, 65, 50–66. DOI: [10.1016/j.specom.2014.05.005](https://doi.org/10.1016/j.specom.2014.05.005). 94, 96
- G. Skantze, M. Johansson, and J. Beskow. 2015. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In *Proceedings of ICMI*. Seattle, WA. DOI: [10.1145/2818346.2820749](https://doi.org/10.1145/2818346.2820749). 84, 87, 92
- G. Skantze. 2016. IrisTK. <http://www.iristk.net> Accessed September 2016. 84
- M. Staudte and M. W. Crocker. 2011. Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, 120, 268–291. DOI: [10.1016/j.cognition.2011.05.005](https://doi.org/10.1016/j.cognition.2011.05.005). 96
- K. W. Strabala, M. K. Lee, A. D. Dragan, J. L Forlizzi, S. Srinivasa, M. Cakmak, V. Micelli. 2013. Towards seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1): 112–132. DOI: [10.5898/JHRI.2.1.Strabala](https://doi.org/10.5898/JHRI.2.1.Strabala). 97
- W. Sumby and I. Pollack. 1954. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26, 212–215. DOI: [10.1121/1.1907309](https://doi.org/10.1121/1.1907309). 81
- M. Sun, A. Schwarz, M. Wu, N. Strom, S. Matsoukas, and S. Vitaladevuni. December 2017. An empirical study of cross-lingual transfer learning techniques for small-footprint keyword spotting. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference*, on pp. 255–260. IEEE. DOI: [10.1109/ICMLA.2017.0-150](https://doi.org/10.1109/ICMLA.2017.0-150). 86
- L. Takayama, W. Ju, and C. Nass. Beyond dirty, dangerous and dull: what everyday people think robots should do. In The 3rd ACM/IEEE International Conference on Human-Robot Interaction, HRI'08, pages 25–32. 77
- M. Tomasello, B. Hare, H. Lehmann, and J. Call. 2007. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, 52(3): 314–320. DOI: [10.1016/j.jhevol.2006.10.001](https://doi.org/10.1016/j.jhevol.2006.10.001). 81
- D. Traum and P. Heeman. 1997. Utterance units in spoken dialogue. In *In Proceedings of ECAI Workshop on Dialogue Processing in Spoken Language Systems*, pp. 125–140. DOI: [10.1007/3-540-63175-5\\_42.90](https://doi.org/10.1007/3-540-63175-5_42.90)
- J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau. 2003. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings of Intelligent Robots and Systems (IROS)*, pp. 1228–1233. DOI: [10.1109/IROS.2003.1248813](https://doi.org/10.1109/IROS.2003.1248813). 85
- O. Vinyals, D. Bohus, and R. Caruana. 2012. Learning speaker, addressee and overlap detection models from multimodal streams. In *Proceedings of the 14th ACM*

*International Conference on Multimodal Interaction*, pp. 417–424. DOI: [10.1145/2388676.2388770](https://doi.org/10.1145/2388676.2388770). 92

P. Viola and M. Jones. 2001. Robust real-time object detection. *International Journal of Computer Vision*, 4. DOI: [10.1.1.110.4868](https://doi.org/10.1.1.110.4868). 88

Å Wallers, J. Edlund, and G. Skantze. 2006. The effects of prosodic features on the interpretation of synthesised backchannels. In E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, editors, *Proceedings of Perception and Interactive Technologies*, pp. 183–187. Springer. DOI: [10.1007/11768029\\_19](https://doi.org/10.1007/11768029_19). 94

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8): 1177–1207. DOI: [10.1016/S0378-2166\(99\)00109-5](https://doi.org/10.1016/S0378-2166(99)00109-5). 93

V. H. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–578. Chicago, IL. 92

M. von Grünau and C. Anston. 1995. The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, 24(11): 1297–1313. DOI: [10.1068/p241297](https://doi.org/10.1080/p241297). 81

# Situated Interaction

Dan Bohus, Eric Horvitz

## 3.1

### Introduction

Interacting with computers via natural language is an enduring aspiration in artificial intelligence. The earliest attempts at dialog between computers and people were *text-based dialog systems*, such as Eliza [Weizenbaum 1966], a pattern-matching chat-bot that emulated a psychotherapist, and SHRDLU [Winograd 1971], a natural language understanding system that allowed interactions with a blocks world. Over the years, progress in large vocabulary speech recognition, natural language understanding, and speech synthesis have enabled the addition of speech capabilities. Attention shifted toward nurturing a new class of task-oriented spoken dialog systems, in which users interact with a system via spoken language over multiple turns to accomplish specific tasks. Examples include the ATIS [Pallett et al. 1992] and Communicator projects [Walker et al. 2002], which aimed to provide information and make reservations for flights and hotels, the TRIPS system [Ferguson and Allen 1999] which explored collaborative planning tasks in various logistics domains, and CALO [Tur et al. 2010], which aimed to construct a personal assistant. The accumulated body of research and technical progress in *task-oriented spoken dialog systems* led to the wide-scale deployment of many telephony-based interfaces, and later to the personal assistants that are embedded in mobile phones, such as Siri, Google Assistant, and Cortana. More recently, these technologies have been deployed into fixed, standalone devices and smart speaker platforms running Alexa, Cortana, and Google Home.

While task-oriented spoken dialog systems have achieved commercial success and are deployed widely, their operation stands in stark contrast to the language interactions that people have in their daily lives. For instance, today's deployed systems assume that there is a single user who interacts solely via speech and who performs one task at a time. The interactions are rigidly controlled via push-to-talk buttons or triggered by wake-up words like "Alexa!", "Cortana!", or "Ok Google!"

The dialog follows a simple pattern of one question followed by one answer. *Dialog context*, i.e., the history of what has been said, is carried and leveraged across multiple turns in limited ways, and the systems do not have awareness of their *physical context* and surroundings.

In contrast, conversations among people are fluid, fast-paced, and context-sensitive. They often involve multiple participants who continuously and effortlessly monitor one another and coordinate their actions across multiple channels to establish and maintain mutual understanding. Beyond the spoken turns, other aspects of the physical context such as body pose, head and hand gestures, the direction or changing patterns of gaze, the locations and affordances of objects in the environment, and the spatial relationships among participants and objects are continuously monitored and interpreted, and these signals play a fundamental role in shaping the interactions.

Over the years progress in speech technologies has enabled task-oriented spoken dialog systems. In a similar manner, advances in vision, perception, and multimodal technologies are starting to make possible a new class of interactive systems that reason more deeply about their surroundings, and that can interact in a more natural fashion in physically situated settings. This area of research is referred to as situated spoken language interaction or, in short, interacting.

Situated interaction extends the scope of engagement with people, from simple, controlled back-and-forth dialog on well-defined tasks, to interactions with and among people in the more complex *open world*. Such a generalization brings numerous new challenges to the forefront. As a concrete illustration, consider the *Receptionist* system [Bohus and Horvitz 2009b], a research prototype built to assist people at Microsoft to make reservations for rides on a campus-wide shuttle service. An example interaction with the Receptionist is shown in Figure 3.1. The system employs a screen that displays a virtual avatar head and engages in spoken language dialog with people in its vicinity. The avatar head pivots to turn its gaze to one or more people with whom it is engaging. The task of booking shuttles is simple and straightforward; the system interacting needs only one piece of information from the user, i.e., the building they are going to. Nevertheless, performing this task in a physically situated setting, such as a building lobby, poses sets of important challenges that extend well beyond those raised by traditional task-oriented spoken dialog systems.

Like a human receptionist, the system must reason about people in its vicinity and understand who it should be interacting with. People sometimes arrive in the lobby in groups, and the groups may dynamically change over time. For exam-

ple, in an open space, people may simply congregate in the area in front of the Receptionist without a need to interact. Someone may peel off a group and come to make a shuttle reservation. In other cases, multiple people may queue up for service, and so forth. In the case of a human receptionist, the status and dynamics of spatial relationships and proximities of people provide signals with shared meaning about a desire to engage and to seek assistance from the receptionist. The Receptionist can engage in conversation about the task at hand, i.e., booking shuttles, but also in communication about the engagement, for example inquiring about who's together with whom and who needs the shuttles (see turns 5–6 in the example interaction from Figure 3.1). At any given point, the system must be able to understand who is engaged in the interaction, who is currently speaking, who they are speaking to, when it is its own time to speak, and to whom it should address the next utterance. The fundamental unit of service is not the conversation, but rather the task of helping the user get on a shuttle to a desired location. This task can span multiple conversations. For instance, in the example shown in Figure 3.1, the interaction with the first user spans four segments (turns 3–4, 7–13, 15–19, 29–32) that are interleaved with other conversations and tasks. The Receptionist must be able to maintain state over multiple interactions and interleave multiple conversational threads. Other aspects of the physical context can also help shape the interaction. For instance, in turn 15, as the Receptionist turns attention back to continue the conversation with  $A_1$ , it notices that his attention is elsewhere. Thus, it uses an interjection to re-engage, continuing only when mutual attention has been re-established. In turn 24, the system asks “And this is just for you, right?” because there is another person hovering in the background; given the proximity and potential relationship between the two people, the system does not assume that the shuttle is just for one person. A human receptionist might be expected to ask such a question to better understand the full context of the task at hand, given the uncertainty about the role and needs of the nearby person.

The example from Figure 3.1 illustrates several ways in which the physical context is relevant and can shape interactions. However, the list of challenges with situated interaction extends however well beyond the cases we have illustrated here. Key problems include the need to reason about the changing roles and relationships of multiple people detected by a system and to understand the meaning and intended targets of spoken utterances and non-verbal actions, within the context of the inferred goals and needs of multiple actors. We know that people continuously leverage shared information about the world in which they are jointly immersed, including the physical structure of spaces, the presence and locations of objects



**Figure 3.1** An illustrative interaction with Receptionist. Capture from Receptionist's point of view. Inset: Receptionist avatar. Opposite page: Transcript of interaction. (Figure adapted from [Bohus and Horvitz \[2009b\]](#) copyright ACM. Video of the interaction is available at <https://aka.ms/si-engv>)

and their affordances, the location, attention, and the likely intentions of others. People are physically and socially situated in their environments and leverage language in the context of shared understandings of each other and the world. To be fully engaged in a fluid and natural manner, interactive systems must similarly leverage both background knowledge and streaming evidence about the physical world and the overall setting of collaborations. Sensing and reasoning from rich, multimodal streams of information is key.

In this chapter, we review the state of the art and discuss the challenges and opportunities for situated spoken language interaction. We adopt a view of language interaction as a coordinated process among participants engaged in a conversation and we identify a minimal set of core competencies that are necessary for successful

A first user, detected and tracked as actor  $A_0$ , passes by in front of the Receptionist system  $R$ ; he makes gaze contact and greets as he passes by and exits the field of view;  $R$  returns the greeting but does not further continue the dialog.

- 1  $A_0 \rightarrow R$  Hi!
- 2  $R \rightarrow A_0$  Hi!

As the user approaches a second time, he is detected and tracked as  $A_1$ . This time he stands in front of the system and greets again;  $R$  opens the dialog.

- 3  $R \rightarrow A_1$  Hi! My name is Laura and today I'm here to help the receptionist with shuttle reservations.  
What's your name?

- 4  $A_1 \rightarrow R$  Dan Bohus.

While the above interaction has been taking place, a second user  $A_2$  also approaches and stands a bit to the right, behind  $A_1$ .  $R$  needs to determine whether the two need a shuttle together, so it engages with both of them.

- 5  $R \rightarrow A_1, A_2$  Are the two of you together?
- 6  $A_2$  No.

$R$  continues the conversation only with  $A_1$

- 7  $R \rightarrow A_1$  Okay, let me start with you first. Do you need a shuttle?
- 8  $A_1 \rightarrow R$  Yes.
- 9  $R \rightarrow A_1$  Which building are you going to?
- 10  $A_1 \rightarrow R$  Building 34.
- 11  $R \rightarrow A_1$  And this is just for you, right?
- 12  $A_1 \rightarrow R$  Right.
- 13  $R \rightarrow A_1$  I'm making a shuttle reservation for you for building 34. Let me know if you need anything else.

At this point  $A_1$ 's attention goes to his cell phone, and  $R$  turns momentarily toward  $A_2$  to let him know he's next.

- 14  $R \rightarrow A_2$  I'll be with you in just a moment.

When  $R$  turns its attention back to  $A_1$ ,  $A_1$  is still looking at his cell phone. To engage,  $R$  interjects a gentle 'Excuse me' and waits until  $A_1$  lifts his attention back to continue the conversation.

- 15  $R \rightarrow A_1$  Excuse me!
- 16  $R \rightarrow A_1$  You will be on shuttle 56. It should be here in two minutes.
- 17  $A_1 \rightarrow R$  All right thank you.
- 18  $R \rightarrow A_1$  Bye-bye!
- 19  $A_1 \rightarrow R$  Bye!

Once the conversation with  $A_1$  is completed,  $R$  turns attention back to  $A_2$ .

- 20  $R \rightarrow A_2$  Now, let's see . . . Do you need a shuttle?
- 21  $A_2 \rightarrow R$  Yes.
- 22  $R \rightarrow A_2$  Which building are you going to?
- 23  $A_2 \rightarrow R$  Building 9.
- 24  $R \rightarrow A_2$  And this is just for you, right?
- 25  $A_2 \rightarrow R$  Yes.
- 26  $R \rightarrow A_2$  I'm making a shuttle reservation for you for building 9 . . . Let me know if you need anything else . . . Let's see, I should get the shuttle number for you in a moment . . . You will be on shuttle 68. It should be here any second.
- 27  $A_2 \rightarrow R$  Thank you!
- 28  $R \rightarrow A_2$  Bye-bye!

At this time  $A_2$  leaves.  $A_1$  is still in the background somewhere further away, waiting for his shuttle.  $R$  understands this but does not initiate any interaction yet. However, at the moment  $A_1$  turns his attention from a distance toward  $R$ ,  $R$  provides some additional information about the shuttle.

- 29  $R \rightarrow A_1$  Let me check on your shuttle. You're still on shuttle 56, it should be here in one minute or so . . .
- 30  $A_1 \rightarrow R$  All right, great, thanks a lot!
- 31  $R \rightarrow A_1$  Bye-bye!
- 32  $A_1 \rightarrow R$  Bye!

**Figure 3.1** (continued)

communication. We also highlight and discuss key factors that characterize physically situated interactions that need to be accounted for as part of a broader solution to the problem. We then focus more deeply on one of these core competencies as a running example: the ability to reason about and manage *engagement*—the process by which people initiate, maintain, and terminate their connection with each other during the interactions that they undertake [Sidner et al. 2004]. We discuss multi-modal methods for making inferences and decisions about engagement, and we highlight various approaches in this area as well as opportunities for future work. While we focus attention on challenges with managing engagement, the issues and themes that we discuss are broader, and re-appear when developing computational models for other communicative processes involved in situated interaction such as turn taking, joint attention, situated language understanding, interaction planning, etc.

## 3.2

### Situated Spoken Language Interaction

Conversations between people are fast-paced, highly coordinated encounters that go far beyond spoken words. A careful look at communication reveals that participants in a conversation continuously coordinate their actions to establish and maintain mutual understanding in a process called grounding [Clark and Schaefer 1989, Clark and Wilkes-Gibbs 1990, Clark and Brennan 1991, Clark 1996]. Successful communication hinges on attending to one another and contributing verbally and non-verbally to solve several problems.

A fundamental problem that participants in an interaction must solve is that of creating and maintaining an open communication channel. Establishing and maintaining a channel involves non-verbal communication and body work. People enter into specific joint spatial orientations denoted as *F-formations* [Kendon 1990b], and signal their intentions to start, maintain, or terminate interactions by using verbal and non-verbal means, e.g., via greetings, salutations, patterns of head and body pose, gaze, and other indicators of attention. Given the serial nature of speech, as well as cognitive and attentional constraints, participants must also coordinate on the temporal production of various communicative signals. People take turns in conversation, and gestures are closely coordinated with speech and other non-verbal actions of themselves and others. Meaning can be ambiguous and context-dependent and, as such, participants must also coordinate to establish a mutual understanding of the meanings of the signals exchanged. Frequently, interactions are permeated with meta-communicative acts such as verbal and non-verbal confirmations and clarifications, aimed at resolving rising uncertainties.

Finally, coordination also happens at the higher level of the overall interaction, where joint goals and intents are negotiated, adopted, or abandoned. The work on coordination of communication enables participants to come to common understandings even when entering interactions with different goals, assumptions, and intentions.

While coordination and grounding span multiple levels (from attention and engagement, to taking turns, to language understanding and interaction planning), work to date on task-oriented spoken dialog systems has been focused largely on problems like speech recognition, natural language understanding, and dialog management. This is not surprising, given the disembodied, non-situated, and audio-only (unimodal) nature of traditional spoken dialog systems. In telephony-based systems, the channel problem can be easily resolved: if a call is received, it is safe to assume that the channel is open, and the user's attention is exclusively on the conversation at hand. Similarly, push-to-talk solutions are both sufficient and efficient for mobile phone assistants. Given the *dyadic* nature of the interactions, i.e., a single user interacting with a single system, simple turn-taking assumptions like you-speak-then-I-speak can often be sufficient. Voice activity detection algorithms are used in these settings to determine when a user's speech starts and ends, and each user turn triggers a system response. The conversation in traditional spoken dialog systems is a simple back-and-forth of contributions (see also discussion on turn taking in Chapter 2).

In contrast, once we move toward physically situated interactions, these solutions are neither sufficient nor appropriate. Situated interaction is characterized by several aspects that mark important departures from assumptions that are often made in task-oriented spoken dialog systems, framing new challenges for multimodal interaction research. We briefly outline them below.

**Physical Context.** For systems that interact in a physical setting, various aspects of the physical surroundings, including people, objects, and spaces, and their spatial configuration and relationships, play a significant role in shaping the interactions.

The configuration of people in the scene, and their non-verbal actions and signals represent an important constituent of the physical context. People involved in a conversation adopt a certain spatial stance while interacting and use it together with other non-verbal signals to mark and understand the *participation status* a person might have with respect to a given utterance and interaction [Goffman 1979, 1981, Clark 1996], e.g., participant vs. non-participant, addressee vs. side-participant, bystander, and eavesdropper (see the *Glossary*). As an example, consider turn 11 in the interaction shown in Figure 3.1. The simple presence of an

additional bystander in the scene can influence the course of the dialog, with the system asking “And this is just for you, right?”, rather than just assuming so. More broadly, and regardless of the task at hand, various non-verbal signals and actions such as the pose and orientation of the head and body, gaze direction, head and hand gestures, facial expressions, as well as their temporal dynamics, communicate information continuously about the level of engagement, understanding, and help to coordinate and regulate turn taking (see Chapter 2).

Beyond contextual cues about other participants, the presence, location and movement of objects in the environment can also provide important cues and can affect the course of an interaction. Participants continuously monitor and understand the location and affordances of relevant objects, as well as the relationships among people and objects. Object references are generated and understood by reasoning jointly about attention, gesture, speech, and physical locations. Finally, beyond people and objects, broader aspects such as the overall topologies of spaces (e.g., the shape and constraints of the open interaction space, presence of doorways, hallways, rooms, walls, whether and how participants are moving through space while interacting or simply in motion to get to a destination, etc.) can also influence the course and shape of interactions.

**Multiparty.** A second important aspect of physically situated interactions is that they are often *multiparty* in nature, i.e., they involve not just one, but rather multiple other people. Open-world settings typically contain more than one relevant person, each with a potentially distinct role, goals, intentions, and needs, all of which may vary over time. People may start, join, or leave interactions at any point in time. People engaged in a conversation may be coordinating in parallel with other people or interleaving their interactions with other tasks and activities. The multiparty nature of the interactions raises new challenges and renders traditional approaches for spoken dialog insufficient. Systems providing services in physically situated settings need robust competencies to work with multiple people. At a minimum, they need to reason about and decide with whom to engage and when, and they need to coordinate their actions and conversational turns with multiple other participants.

**Multimodal.** A third important aspect of physically situated interactions is that they involve multiple input and output channels. Participants coordinate with each other not only via spoken words, but also via other signals and actions. Non-verbal signals and gestures such as body pose, head orientation, gaze direction, attention, head and hand gestures, as well as spatial positioning and orientation play a fundamental role in coordination and grounding, and critical inferences are often made based on these non-linguistic signals. In addition, physical actions

### Glossary

**Decision-theoretic:** reasoning methodology for selecting ideal actions in accordance with the principles of probability and utility; decision-theoretic reasoning involves guiding actions by expected utility. Expected utilities of actions are computed by coupling probabilities, inferred about current or future states, with considerations of the value of outcomes (see [Horvitz et al. \[1988\]](#)).

**Dialog context:** information contained in previous utterances that is relevant to and can help determine how the conversation progresses.

**Dyadic:** a term denoting an interaction that involves two participants. Most work in spoken dialog systems has traditionally focused on dyadic settings, where a dialog system interacts with a single human user.

**Encoding dictionary:** captures a set of templates, i.e., transformations, which can be applied to a feature to better capture the relationship between this feature and events on other modalities [[Morency et al. 2008](#)]. Examples include temporally shifting the location of the feature activation by an offset to enable reasoning about potential coordinative delays between activities on different channels, or the use of ramp functions in cases where the influence on the target variable is expected to be changing over time.

**Engagement:** we adopt here the definition proposed by [Sidner et al. \[2004\]](#) as “the process by which two (or more) participants establish, maintain, and end their perceived connection.” Situated interaction systems generally need to reason about and manage engagement, i.e., who they are interacting with, and when.

**F-formation:** a term used to denote the spatial pattern in which participants arrange themselves during interactions. Per [[Kendon 1990b](#)], “*An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.*” The prototypical example is a circular pattern, with participants oriented toward a common center, although other configurations such as L-shaped, V-shaped, side-by-side, and vis-à-vis are common.

**Joint inference:** an inference model that reasons jointly about multiple entities and that produces a probability distribution over the joint space (cardinal product) of the variables of interest. This stands in contrast to inference models that reason independently about each entity, and that produce a separate probability distribution for each variable of interest.

**Multiparty interaction:** an interaction that may involve more than two participants. Situated interaction systems deployed in the open world need to be designed to handle multiparty interactions, as people may often arrive and interact with the system in groups.

### Glossary

**Participation status:** characterizes the alignment between people and a particular interaction. At the high level, people in a scene can be divided into participants (those who are involved in the interaction) and non-participants (those who are not). Participants can be further divided into speaker (the producer of an utterance), addressees (the people being addressed), and side-participants (the people that are not being addressed by the current utterance but are still ratified participants). Non-participants include bystanders (people nearby that hear the utterance but do not participate in the interaction and are known to the speaker), and eavesdroppers (other people that are listening, but are not known to the speaker) (see [Goffman \[1979\]](#), [1981](#), [Clark \[1996\]](#)).

**Physical context:** information contained in the physical environment that is relevant to and helps determine how the conversation progresses. Examples include information about people, such as how many people are around, their location and body pose, head and hand gestures, eye gaze and attention; information about the presence, location, and affordances provided by task-relevant objects; the overall topology and structure of spaces, etc.

**Proxemics:** a term coined by [Hall \[1966\]](#) denoting “the interrelated observations and theories of man’s use of space as a specialized elaboration of culture.” Proxemic (spatial) information is an important ingredient in reasoning about conversational engagement, participation status, and other communicative processes in physically situated interaction.

**Situated (spoken language) interaction:** an area of research investigating the development of computer systems that can reason about their surroundings and interact (via spoken language) in a more natural manner in open-world, physically situated settings. Such systems generally integrate information from speech and vision, reason about both verbal and non-verbal behaviors of potentially multiple participants, and leverage both the dialog and [physical context](#) when making decisions.

**System initiative, User initiative, Mixed initiative:** Terms that refer to the nature of decisions about action or initiative in human-computer interaction. With system initiative, the computer system takes primary control of the flow of the activity, including conversation or other kinds of collaborations. With user initiative, the human takes primary control of the flow of the activity. In mixed-initiative systems, the computer and human can each take primary control of one or more steps of the activity. Initiative in a mixed-initiative system can be guided by simple fixed-policies, heuristic procedures, or probabilistic and decision-theoretic inference (see [Horvitz \[1999\]](#)).

**Task-oriented spoken dialog system:** a computer system that allows the user to perform a certain task, like booking a flight or checking the weather, via spoken language input and output.

**Glossary** (*continued*)

**Text-based dialog system:** a computer system that interacts with a user via natural language text input and output.

**Tracking, smoothing, and forecasting:** Terms that refer to different hidden state estimation problems in a dynamical system. Let us assume a system which has an unknown but evolving state over time. The tracking problem (also sometimes referred to as filtering) amounts to estimating the state of the system, at the current time  $t$ , based on observations made up to the current time  $t$ , i.e.,  $P(s_t|o_1, o_2, \dots, o_t)$ . Smoothing refers to estimating the state of the system, at some past time  $p < t$ , based on observations made up to the current time  $t$ , i.e.,  $P(s_p|o_1, o_2, \dots, o_t)$ . Finally, forecasting or prediction refer to the task of estimating the state of the system, at some future time  $f > t$ , based on observations made up to the current time  $t$ , i.e.  $P(s_f|o_1, o_2, \dots, o_t)$ .

**Wait-vs.-act tradeoff:** denotes a tradeoff that often arises in systems that need to make decisions under uncertainty and latency constraints. In general, by choosing to wait rather than act, a system might accumulate more evidence about important state variables, reduce uncertainty, and therefore make a better decision about which action to take. However, as timing is also important, waiting might also lead to missed opportunities, or to increased costs of action, e.g., due to inappropriateness of delay.

such as manipulating objects, pointing, grasping, and handing over are also closely coordinated with the other linguistic and non-verbal actions. Multiple channels are used to provide redundancy and improve robustness in communication, but also to complement, nuance and sometimes entirely alter the meaning carried via the linguistic channel.

Creating systems that can engage in conversation in physically situated settings requires that we adopt a multimodal stance, and that we reconsider the various processes involved in communication and grounding from a multimodal perspective. Core competencies for managing engagement, coordinating on taking turns, conveying and understanding the meaning of utterances, and high-level interaction planning must be anchored in reasoning from continuous streams of multimodal evidence.

In the remainder of this chapter, we will focus on the single core competency of managing engagement as an illustrative example. We note that many of the considerations, themes, and challenges that we encounter and discuss around engagement are relevant for other competencies.

## 3.3 Engagement

The term engagement has been used in the research community in a variety of ways. Here, we adopt the definition introduced by [Sidner et al. \[2004\]](#), who define engagement as “the process by which two (or more) participants establish, maintain and end their perceived connection.” Prior research in anthropology, psycholinguistics, and sociolinguistics sheds light on various multimodal and coordinative aspects of this process in human-human interactions. For instance, [Kendon \[1990a\]](#) identifies several stages of action coordination that occur during initiation of engagement: pre-sighting, sighting, distance salutation, approach, and close salutation. [Kendon \[1990b\]](#) uses the term F-formation to describe the physical configuration pattern that “arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access,” and discusses various types of F-formations and the role they play in interactions. [Goffman \[1963\]](#), [Argyle and Cook \[1976\]](#), [Kendon \[1990a\]](#), as well as [Vertegaal et al. \[2001\]](#), among others, have discussed how various cues, including gaze, are used to initiate and maintain social and communicative engagement. Overall, the existing body of work in human-human communication shows that engagement is a highly-coordinated, ***mixed-initiative*** process that leverages both verbal and non-verbal signals, including proximity, body pose, eye gaze, head and hand gestures, as well as verbal salutations.

In single-user, non-situated systems, the management of engagement generally can be done via simple means. In traditional, telephony-based dialog systems, the system can safely assume it is engaged in a conversation from the moment the call is received until the user hangs up. Similarly, in mobile phone assistants, a push-to-talk button is used to signal engagement. These approaches are sufficient, appropriate, and efficient in these settings. In contrast, when dealing with a physically situated, embodied system such as a robot providing assistance in an open-world setting, the problem of managing engagement becomes more difficult, and the system cannot rely on simple, explicit signals. Instead, people will naturally rely on the use of such signals as approach behaviors, salutations, and establishing and breaking mutual gaze. Thus, a more sophisticated model that enables a system to reason about and manage the complexities arising from the situated nature of the interaction is needed.

Several key ingredients are required to manage engagement in open-world settings. First, we need an appropriate *representation*, i.e., a set of key variables that provide the basis for making engagement control decisions. Second, we need to construct the *perceptual* capabilities that enable the system to track these variables

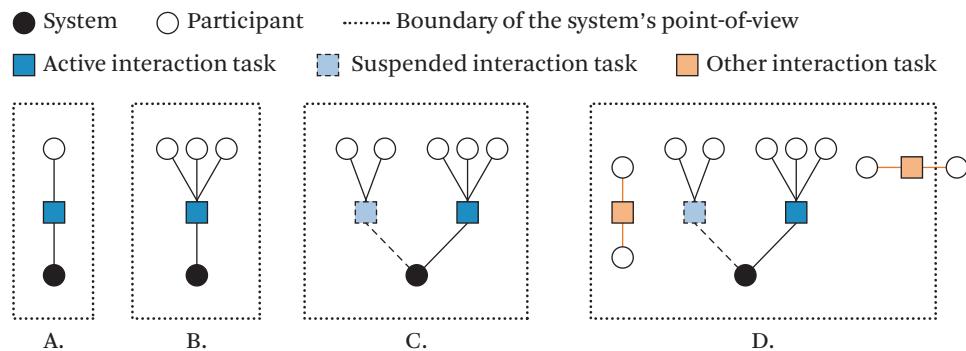
from lower level data streams collected by sensors, such as audio and video signals. Third, we need a *control policy* that indicates how the system should make engagement control decisions based on sensed or inferred state of the world. Finally, we need to *render* the high-level engagement control decisions made by the system, such as *engage-with-person-X*, into lower-level coordinated multimodal outputs, such as appropriate gestures and speech. In the following subsections, we discuss each of these challenges in more detail.

### 3.3.1 Representation

Managing engagement in physically situated settings requires an appropriate representation that can serve as a basis for authoring engagement control decisions. For instance, a situated system needs to explicitly track what interactions are in progress and who is engaged in which interaction. In [Bohus and Horvitz \[2009c\]](#), we articulate a representation and a model that specifically target situated interaction scenarios, where multiple participants might be attending to multiple tasks. We briefly review this model as a starting point for a broader discussion of the multimodal challenges that arise.

The proposed model leverages a refined notion of *interaction task* as a basic unit of sustained, interactive problem solving. For instance, in the receptionist domain, an interaction task may be targeted at eliciting information from the user and making a shuttle reservation. Another task may be notifying a user who has been waiting that he or she will be attended to momentarily, or that their shuttle should be arriving shortly. Each interaction task can involve two or more participants (where one of the participants may be the system itself), and this number may vary in time as new participants may join and current participants may leave an existing interaction. In the work and implementation reported in [Bohus and Horvitz \[2009c\]](#), we focus on the ability to model multiple engagements with the system, corresponding to the multiparty setting illustrated in Figure 3.2C. With the proposed model, the system can be actively engaged in one interaction task at a time, but it can keep track of additional, suspended interaction tasks. In some settings, it may be useful to go beyond this approach, and to also reason about engagements among others, as shown in Figure 3.2D, and do so even when some participants fall outside the sensing field of the system.

The representation in [Bohus and Horvitz \[2009c\]](#) subsumes a set of variables that track the *engagement state*, *engagement actions*, and *engagement intentions* for each detected person. The engagement state is a binary variable which indicates whether a person is engaged in an interaction with the system or not. The engagement intention is also a binary variable that reflects whether a person *intends* to be



**Figure 3.2** Schematic representation of different levels of capabilities for reasoning about and managing engagement in situated interaction: (A) dyadic setting: system can handle a single interaction at a time, with a single end user; (B) multiparticle setting: system can handle a single interaction at a time, with multiple end users; (C) multiparty setting: system can handle multiple multiparticle interactions at a given time, one of which may be active; and (D) open-world, multiparty setting: system can track multiple multiparty interactions (including ones in which the system is not involved, and where some participants may fall outside the known world, e.g., outside the field of view of the system). (Figure adapted from [Bohus and Horvitz \[2009d\]](#) copyright IJCAI)

engaged in an interaction with the system. Finally, at any given point in time, each person may perform one of four engagement actions, depending on the engagement state that they are in. A person in an *Engaged* state may perform a *Maintain* or a *Disengage* action, indicating they are maintaining or terminating the existing engagement with the system. A person in a *NotEngaged* state may perform an *Engage* action indicating that they are trying to start a new engagement with the system, or *NoAction* otherwise. The engagement variables are inferred and tracked based on lower-level sensor data. We review various approaches to making inferences about engagement variables in Section 3.2.

The detailed trace of the interaction with the Receptionist system displayed in the sidebar illustrates how the proposed representation supports making key inferences and enables a variety of engagement behaviors for the system. First, the separation between engagement state and actions allows reasoning about engagement as a collaborative process: the engagement state evolves as a result of the joint actions of the participants. For example, in order to transition from a *NotEngaged* to an *Engaged* state, both participants need to perform an *Engage* action. Similarly, to maintain the *Engaged* state, both participants need to continuously perform the

*Maintain* action. The separation between engagement actions and intentions allows us to model situations where a participant may want (intend) to engage in an interaction but may not be taking any directly visible engagement actions, such as when they are waiting patiently in the background for an engagement between the Receptionist and another person to end.

The open-world, multiparty engagement model operates in a sense-think-act loop and leverages these variables to make engagement control decisions. Based on the inferred values of these variables, as well as other higher-level evidence and constraints, an *engagement control policy* makes decisions about which engagement actions the system should take at any given time, that is, whom the system should engage or disengage with, and when. The engagement actions chosen by the control policy are then rendered into a set of coordinated low-level verbal and non-verbal behaviors, such as greetings, salutations, interjections, head and hand gestures, and establishing or breaking eye contact. The rendering also depends on context and on the affordances provided by the available embodiment.

The example interaction shown in Figure 3.1 and described in more detail in Sidebar 3.1 is a demonstration scripted to illustrate how the proposed representation and control policy can provide support for managing engagement in open-world interaction scenarios, even when using simple, heuristic inference models. However, creating systems that can robustly manage engagement with this type of fluidity in the real world remains a challenging task. While the proposed representation provides the basis for reasoning about multiparty engagement, acting correctly depends critically on accurate multimodal perception, on timely decision making, and on well-coordinated production of multimodal low-level actions. As decisions are often done under uncertainty, actions may be inappropriate or poorly timed. Thus, methods for error recognition and strategies for recovering from errors can be valuable. We now turn our attention to several of these challenges, review the state of the art, discuss central themes that emerge, and highlight challenges and opportunities for future research.

### 3.3.2 Perception

The starting point for making good decisions about engagement in physically situated settings is the ability to sense and correctly interpret the verbal and non-verbal cues that people naturally use. Accurate perception is therefore a key challenge. Several research efforts, such as [Michalowski et al. \[2006\]](#), [Bohus and Horvitz \[2009a\]](#), [\[2014\]](#), [Castellano et al. \[2009\]](#), [Rich et al. \[2010\]](#), [Sanghvi et al. \[2011\]](#), [Szafir and Mutlu \[2012\]](#), [Xu et al. \[2013\]](#), [Foster et al. \[2013\]](#), and [Leite et al. \[2015\]](#) have focused

**Sidebar 3.1** Detailed Trace of Interaction with Receptionist

We review in more detail the example interaction with the Receptionist system introduced earlier and discuss how the proposed representation and engagement control policy enable it to reason about and generate a variety of engagement behaviors. More information about the overall system software and hardware architecture, as well as its various situational awareness capabilities such as face detection, pose tracking, focus of attention, and additional inferences made by the system, can be found in [Bohus and Horvitz \[2009b\]](#).

The trace of the interaction, as well as image captures at key points and the various engagement actions taken by the Receptionist, are shown in Figure 3.3. At time  $t_0$ , as the first user (who is detected and tracked by the system as actor  $A_0$ ) is approaching, the Receptionist computes the probability that he has an intention to engage—see  $P(EI=Engaged)$  in Figure 3.4B. The user greets the Receptionist as he passes by. This is detected as an *Engage* action; see Figure 3.4C, and the system responds at  $t_1$  with its own *Engage* action, rendered as a greeting, and constructs an active interaction task  $I_0$  at time  $t_2$ . At about the same time, the participant exits the field of view and the tracked face is lost by the vision system. The Receptionist keeps reasoning about the missing participant for a few moments and the heuristic model that is used to compute the probability of disengagement indicates over time that it is likely that the participant is performing a *Disengage* rather than *Maintain* action (see Figure 3.4D). The system disengages in turn at time  $t_3$  and terminates the interaction task. Since no face is present, no verbal rendering of the system's disengagement action is created.

As the user approaches again, he is detected and tracked by the system as  $A_1$  (the system increments the actor index as it cannot discern that he is the same person) and the sequence repeats itself. This time, however, the user remains in front of the Receptionist, establishing an F-formation with the system. At  $t_5$ , the engagement action inference model recognizes this as an *Engage* action from  $A_1$  (see Figure 3.4G.) The Receptionist creates the interaction task  $I_1$  via a greeting and continues by introducing itself and asking for the participant's name.

Shortly after time  $t_6$ , a second user  $A_2$  approaches and places himself in line, somewhere behind the first participant, but in the field of view of the Receptionist. The engagement model infers that this user has an intention to engage but does not yet produce an engagement action; see  $P(EI=Engaged)$  and  $P(EA=Engage)$  for  $A_2$  in Figure 3.4H and 3.4I. The distinct modeling of intentions and actions allows the system to manage engagement in this complex situation. At time  $t_7$ , after allowing the first participant to respond to the original question, the Receptionist takes an engagement action toward the second participant by simply glancing toward him, and continues the interaction task  $I_1$ , now with both participants, by addressing the question “Are the two of you together?” to both of them. Once the second participant responds “No”, the Receptionist disengages with him at time  $t_9$ , and continues the interaction task with only the first participant. The particular phrase chosen to render the continuation, “Okay. Let me start with you first” is selected based on the temporal context and previous engagement activities.

**Sidebar 3.1** (*continued*)

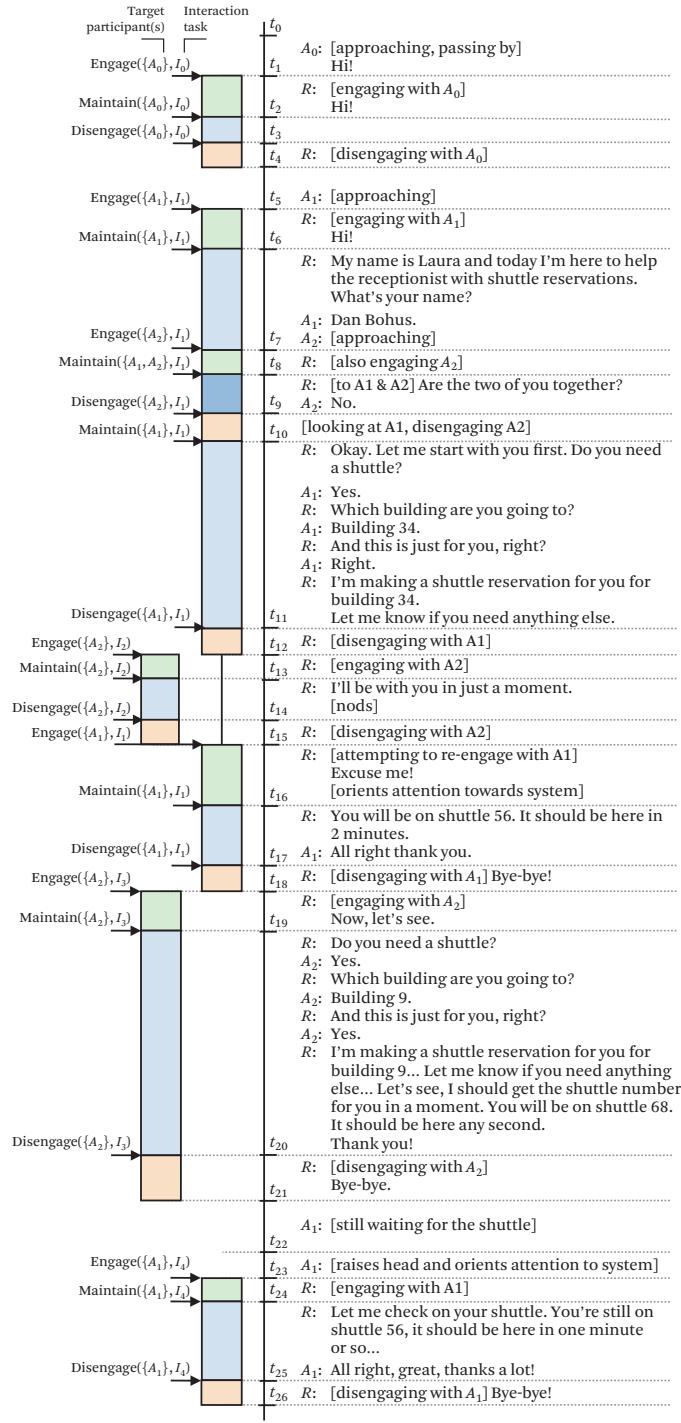
At time  $t_{11}$ , the Receptionist triggered a shuttle reservation request for the first participant and is waiting for a response from the shuttle service. Since this might take a while, the system decides to temporarily suspend the interaction task  $I_1$  with the first participant and to engage momentarily in a new interaction task  $I_2$  with the second participant, just to let them know that they will be attended to soon (see the right hand side in Figure 3.3, from time  $t_{12}$  to  $t_{15}$ ). Next, at  $t_{15}$ , as the Receptionist attempts to execute an action to re-engage in interaction task  $I_1$  with the first participant, it finds that his attention is not on the system, but rather on his mobile phone. Engagement actions are however rendered based on the available context. Given the situation, rather than simply continuing the dialog, the system renders its own Engage action as an interjection, i.e., “Excuse me!” to draw the participant’s attention. Once his attention returns to the system, the engagement action completes successfully, and the system continues the interaction.

Finally, at time  $t_{17}$  the Receptionist completes the current interaction and disengages with the first participant, and then turns and re-engages with the second participant, creating a new interaction task  $I_3$ . The rendering of the engagement action toward him from  $t_{18}$  is again contextualized by the history and the system re-opens with “Now, let’s see . . . ” At  $t_{21}$ , the interaction with the second participant completes. In the meantime, the first user has been standing in the background, waiting for the shuttle. At time  $t_{22}$ , while he is still somewhat at a distance, his attention turns back toward the system. Given that the Receptionist knows this user’s current activity is waiting for the shuttle, the engagement inference model indicates the user likely wants to engage, even though he is still at a distance and not in a clear F-formation with the system. Based on this, the engagement control policy initiates another engagement and creates interaction task  $I_4$  from  $t_{23}$  to  $t_{26}$  to let the participant know about the status of the shuttle.

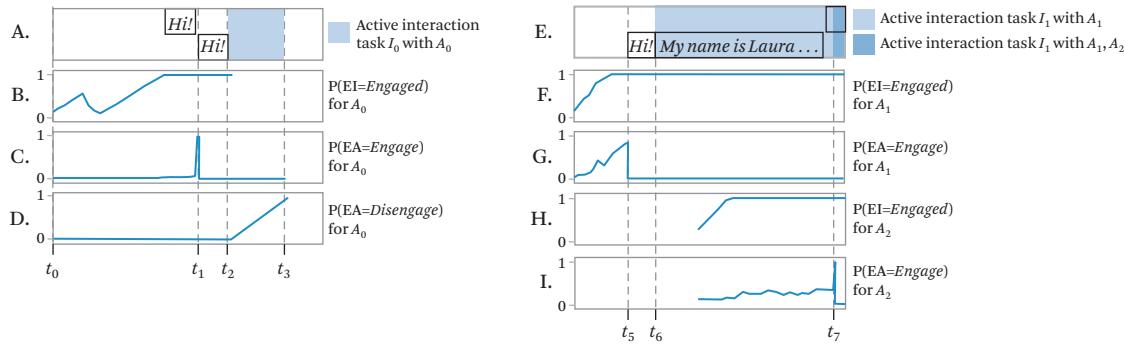
on detecting key variables related to engagement, and a set of common approaches and themes can be identified across this body of work.

### **3.3.2.1 Perception as Multimodal Inference Challenge**

A first observation is that inference of engagement variables is anchored in sensing a diverse set of verbal and non-verbal cues. As we have previously mentioned, a large set of signals and behaviors have been identified to be involved in regulating engagement, and several computational efforts have built upon this existing knowledge base. For instance, [Michałowski et al. \[2006\]](#) takes as a starting point the framework of *proxemics* introduced by [Hall \[1966\]](#) and proposes an approach anchored in the spatial domain for detecting engagement with a stationary robot. [Hall \[1966\]](#) proposed a classification of distances between individuals into a set



**Figure 3.3** Illustrative interaction with Receptionist. Left: Scene analysis results from various points in the interaction. Right: Transcript of interaction showing interaction tasks and system engagement actions. (Figure adapted from [Bohus and Horvitz \[2009b\]](#) copyright ACM)



**Figure 3.4** Inferences about key engagement variables during example Receptionist interaction. (Figure adapted from Bohus and Horvitz [2009b] copyright ACM)

of nested zones: intimate (0–18 in), personal (18 in–4 ft), social (4–12 ft), and public (>12 ft). The model from Michalowski et al. [2006] is loosely based on Hall’s idea of social distances, and classifies potential participants into four categories depending on their distance to the robot: present (people standing far), attending (people idling closer to the robot), engaged (people next to the robot’s booth), and interacting (people actively participating in an exchange with the robot). The robot then provides differential responses toward the potential participants, based on this classification. Another important cue is visual attention. Automatic tracking of eye gaze is generally difficult from a distance and in open-world settings, but alternative measures such as head pose tracking or frontal face detection features can be used to approximate visual focus of attention and relate that to measures of engagement [Bohus and Horvitz 2009a]. Besides spatiotemporal and attentional features, a number of other cues have been leveraged in various engagement detection tasks. These include body posture and motion features such as slouching, leaning forward or backward, quantity-of-motion [Sanghvi et al., 2011], dialog state information [Bohus and Horvitz 2014], facial expressions such as smiles [Castellano et al. 2009, Xu et al. 2013] and physiological sensors [Szafir and Mutlu 2012].

### 3.3.2.2 Multimodal Fusion via Heuristics

Given the diversity of cues involved, inference often relies on multimodal fusion (see the Glossary of Chapter 1 and a broader discussion on multimodal fusion in Wagner and Andre [2018] and Chapter 1). A simple class of approaches for the multimodal fusion problem is to construct heuristics that aggregate lower-level sensing and inferences. As an example, in Bohus and Horvitz [2014] we report

on a heuristic model for estimating whether a participant is disengaging with a stationary robot in open-world settings. The model was informed by accumulated observations and lessons learned with the robot and relies on three signals related to engagement: *proximity* (P), *stability* (S), and *attention persistence* (A). Based on these signals, the probability that a participant is disengaging is estimated as follows:

$$P(\text{Disengagement}) = 1 - P \cdot (1 - (1 - S) \cdot (1 - A)).$$

The scores for proximity (P), stability (S), and attention persistence (A) are computed by lower-level models which constrain the possible values to the 0–1 range via manually tuned sigmoid transforms. The proximity score is based on the size of the tracked faces. The stability score is based on the ratio between the maximum horizontal excursion of the face throughout the last second (i.e., the difference between the maximum and minimum value throughout this time interval) and the size of the face. Finally, the attention persistence score is based on the average (computed over the past 2 s) of the probability that the user is attending to the system, which was in turn estimated via a machine-learned model that leveraged face tracking and head pose information. The disengagement equation shown above captures an assumption that, for a participant to be engaged, the proximity score must be high and at least one of stability or attention persistence scores must be high. If the proximity score is low, this indicates that the user is far away, and the disengagement score increases. Alternatively, if the user is nearby but both the stability and attention persistence scores are low, the probability of disengagement again increases toward one, as these signals indicate the participant is moving and not attending to the system.

In another example of heuristic inference about engagement, Rich et al. [2010] proposed to recognize when participants maintain engagement by integrating information from four types of *connection events*: directed gaze, mutual gaze, conversational adjacency pairs, and backchannels. A directed gaze event captures the situation when one participant looks and optionally points to an object and a second participant directs her gaze to the same location. A mutual gaze event captures the situation when two participants are looking at each other. An adjacency pair event captures the situation when two utterances by two participants have minimal gap or overlap between them, and the first utterance provokes the second (e.g., a question-answer pair). Rich et al. [2010] extend this notion of adjacency pair to include both verbal and non-verbal communicative acts (e.g., a nod as an answer to a question). Finally, a backchannel event corresponds to a brief verbal (e.g., “uh

huh") or non-verbal acknowledgement (e.g., head nod) produced by a listener while the other participant is speaking. These multimodal events are detected via audio and visual processing, and the proposed model computes a measure of *pace* of the interaction as:

$$pace \propto \frac{1}{MBTCE},$$

where *MTBCE* is the Mean Time Between Connection Events. The proposed model computes pace over a sliding window and considers a decrease in pace as evidence of disengagement.

The advantage of heuristic methods, like the two we have reviewed above, is that they provide a simple and transparent approach for engagement inference, which is decoupled from and can be layered over lower-level sensing. The decoupling allows for the lower-level sensing components to be improved over time, while maintaining the high-level inference rules: for instance, in the model proposed by Rich et al. [2010], the detectors for various connection events as well as aspects of the high-level rule, such as a threshold on the pace, could be adjusted independently. These simple, layered models are also transparent, i.e., they can be inspected and enable debugging: for instance, with the Bohus and Horvitz [2014] disengagement model reviewed above, if the model commits an error, developers can inspect and diagnose which one of the proximity, stability, or attention persistence scores were inaccurate, or whether one of the modeling assumptions was being violated, e.g., are other variables, besides proximity, stability, and attention relevant in a given case? At the same time, these heuristic approaches rely on expert knowledge. The fine-tuning process requires manual effort and the methods cannot be easily scaled to take advantage of the large number of additional signals and data that is often available via sensing.

### **3.3.2.3 Machine Learning Approaches**

An alternative, common approach for developing multimodal inference models that aims to address these limitations is to use machine learning. With this approach, inference models are learned automatically from data, based on larger sets of extracted features and ground truth labels about the variable of interest. For instance, in Bohus and Horvitz [2009a], we leveraged a set of spatiotemporal and attentional features with a maximum entropy model to infer whether a person is about to initiate engagement with a situated system. In other work, Castellano et al. [2009] jointly leveraged non-verbal features, such as whether the user is looking at the robot and whether the user is smiling, task-level features, such as the

state of the game, and social-interaction based features, such as the nature of the robot's responses, to detect engagement in a game-based human-robot-interaction task. Their results showed that the task and social-interaction features can help increase robustness when challenges like poor illumination and noisy backgrounds affect the visual channel. Xu et al. [2013] harnessed a variety of spatiotemporal (distance, angle, motion), postural (upper body pose), attentional (face detected, gaze directed to agent), verbal (user speaking), and emotional (facial expression such as smile or scowl) features to assess engagement intentions. Foster et al. [2013] trained classifiers to estimate the engagement state of customers for a bar-tender robot. They used spatial, postural, and audio features, such as the x, y, z coordinates for the customer's head and hands, torso angle, as well as an estimate of whether the customer is speaking. Their experiments with a variety of machine learning techniques (e.g., regression, nearest neighbors, decision trees, SVMs, etc.) and an offline model evaluation showed that the learned models outperformed a rule-based classifier that assumed a user was seeking engagement when their head was close to the bar and they were roughly facing forward.

In contrast to heuristic approaches, multimodal machine learning approaches allow for the easy integration of large numbers of features and can generally achieve better performance given appropriate quantities of data. At the same time, they also bring to the forefront several important additional challenges. In the remainder of this section, we briefly outline some of the core issues and themes that arise when using machine learning approaches for making inferences about engagement and, more generally, about other perceptual tasks in support of *situated spoken language interaction*.

**Feature Engineering.** The performance of machine-learned models often hinges critically on the quality of the data and features used. As a result, a key challenge with any machine learning approach is constructing informative features for the task at hand.

As time and coordination play an important role in interaction, one key aspect with inference models for situated interaction is capturing the temporal dynamics of various signals. Across multiple modalities, various phenomena operate in a coordinated fashion and carry relevant information at different time scales, from hundreds of milliseconds all the way to minutes. Morency et al. [2008] describe an approach for engineering features that capture such temporal statistics and dependencies between modalities via an *encoding dictionary*. The encoding dictionary contains a set of templates, i.e., transformations, that can be applied to the raw features to generate derived features that can be more informative for the task at hand. For instance, the templates may temporally shift the location of the

feature activation on the original stream, with windows of various widths to allow for reasoning about potential coordinative delays between activities on different channels. Another example is the use of a set of ramp functions in cases where the influence on the target variable is expected to change over time, etc.

In our own work on detecting engagement [[Bohus and Horvitz 2009a](#)], we trained maximum entropy models using raw temporal signals like face location, size, and attention, as well as a set of derived temporal statistics over windows of different sizes (in that work we used windows of 5, 10, 20, and 30 frames). For continuous features, the temporal statistics included the min, max, mean, and variance of the feature in a given window, as well as parameters of linear and quadratic fits through the signal in that window. For binary features, such as the attention signal (i.e., is the user's attention on the system or not?), we included the number and proportion of times when the feature had a value of 1 in a given time window, and the number of frames since the feature last had a value of 1. Our results from [[Bohus and Horvitz 2009a](#)] indicated that the additional temporal features helped improve model performance.

This type of feature engineering is often a time-intensive and iterative process. In general, the process is guided by observations and the developer's knowledge or intuitions about the nature of the phenomena being modeled and can therefore be suboptimal. Recent work with deep neural networks shows that in a number of domains we can alleviate the need to manually perform detailed feature engineering. While deep learning approaches generally require large amounts of data, these techniques have proven to be effective and have generated significant improvements in areas such as speech and visual processing. In addition, the feature representations constructed by deep neural networks have been shown to be reusable for training models in new but related audio or visual domains, with relatively small amounts of data. This type of approach may also be useful in inference problems related to situated interaction, such as learning to predict engagement states, actions, and intentions. For a more in-depth look at the use of deep learning approaches in multimodal systems, see [[Keren et al. \[2018\]](#)] and [[Baltrušaitis et al. \[2018\]](#)].

***Labeling Data.*** A second important challenge with machine learning approaches is obtaining labeled data for training. Frequently, labels are constructed through an intensive process in which a trained annotator views audiovisual and other sensor data and constructs the necessary labels. In the realm of situated interaction, many variables capture important distinctions in the world or behaviors that are generally difficult to reify and therefore to annotate, such as the mental states and attitudes of participants. Developing good annotation schemes can take significant effort. Data

annotation often requires extensive training and is done via specialized multimodal annotation tools such as ELAN [Brugman and Russel 2004] and Anvil [Kipp 2001].

Given the challenges of working with video data and behavioral constructs, crowdsourcing approaches, which have been extensively and successfully used to label data in other domains like images and text, can be more challenging in this space [Cabrera-Quiros et al. 2018]. The set of available tools continues however to evolve and improve, with multiple video-centric annotation tools being explored in the vision community (for a brief review, see Bianco et al. [2015]), and with interesting approaches aiming at rapidly coding behavioral video via crowds [Lasecki et al. 2014].

Beyond on-site labeling via trained annotators and crowdsourcing approaches, situated interaction systems provide interesting opportunities for automatically acquiring labels online, through interaction. The users of the system can sometimes provide the required labels. This may happen explicitly, for instance, when the user provides an answer to a system question that elicits specific information about the current context. Labels can also be generated implicitly. As an example, a system may not be able to determine in a timely fashion whether a person is disengaging, but may discover this a bit later, as the person departs and is far away or disappears from the scene. In this case a useful “disengaging” label can be constructed by projecting this information backwards in time, to a moment when it may be relevant for the system (see Bohus and Horvitz [2014]).

**Forecasting.** So far, we have discussed problems where the goal is to infer the value of a hidden variable such as an engagement action or intention at the current time, from evidence available to the system via its sensors. Knowing the state of the world at the current moment is however not always enough. Timely decisions often hinge on the ability to anticipate what will be happening in the future, i.e., to make inferences about the value of a hidden variable at a future time: will this participant disengage in the next three seconds? Will this participant start talking in the next 500ms? We refer to this type of inference as *forecasting* or prediction, in contrast to *tracking* or filtering, which refers to inferring the current state from current evidence, or *smoothing*, which refers to inferring a past state from current evidence (see the [Glossary](#)). The same type of machine learning approaches we have discussed above can be used to develop forecasting models—the difference lies largely in how the labels are generated.

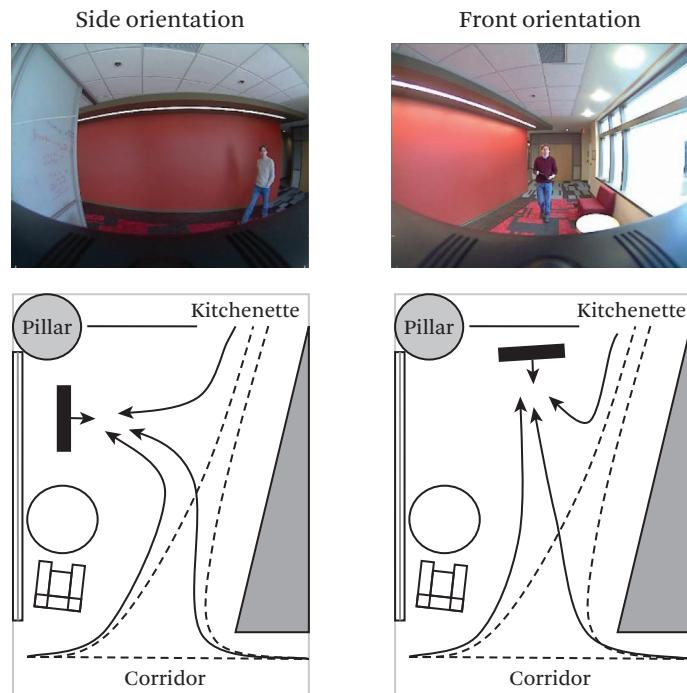
For instance, in Bohus and Horvitz [2014], we have shown how spatiotemporal, attentional, and dialog features can be used with logistic regression or boosted trees to anticipate whether a user will disengage with a system within a future time window. In another forecasting example, working from human-human interaction

data, [Skantze \[2017\]](#) constructed a model for predicting speech activity in a 3-s future time window in a dyadic setting, by using an LSTM-based, recurrent neural network approach (for an overview of deep learning approaches for multimodal learning; see also [Keren et al. \[2018\]](#)). He showed that the learned model can be applied to predict whether another speaker will take the turn when a pause is detected, or whether the upcoming turn will be a long utterance or a short backchannel when speech onset is detected. [Roddy et al. \[2018\]](#) further improve upon the approach described in [Skantze \[2017\]](#) by architecting the network to enable reasoning at multiple time scales.

In a number of prediction problems, the training labels can be automatically determined by the system—the challenge is that they become known too late for making decisions. As an example, in the model for predicting disengagement discussed above [[Bohus and Horvitz 2014](#)], the labels were automatically constructed by rolling back in time a conservative heuristic rule that was able to eventually detect disengagement. The availability of labels in the interaction can enable online learning approaches, where the system continues to learn and get better with time at anticipating the future. Training effective predictive models remains an important, open area of research.

***Joint Inference.*** Another important theme for systems designed to operate in physically situated settings is that of [\*joint inference\*](#). Inference models are commonly built to reason about one entity at a time. For instance, models for tracking engagement might be trained on and applied to each detected participant individually, in essence assuming independence in the engagement of participants. This type of independence assumptions may often be violated, leading to potential losses in modeling fidelity. Besides reasoning jointly about participants, different variables capturing information about the same participant, such as their visual focus of attention, engagement, and group relationships, are also related to each other. There are important research questions and opportunities for progress in the area of joint modeling, and advances in this space will lead to increases in accuracy and ultimately to more natural interactions.

***Robustness and Generalization.*** Finally, as system performance often critically depends on the accuracy of the perceptual pipeline, developing inference models that are robust and generalize well to a wide array of situations is another important and challenging problem. Machine learning approaches are susceptible to overfitting to the data used in training, and various types of mismatches between the training set and the data encountered at run-time can significantly impact performance. As an illustrative example, in [Bohus and Horvitz \[2009a\]](#), we trained models for predicting engagement intentions with a situated system in two different spatial



**Figure 3.5** Engagement experiments in two different orientations denoted Side and Front. Top: Image from system's viewpoint. Bottom: Top-down view of system placement (black rectangle is the system's screen; dotted lines are typical trajectories of movement for people in this space; continuous lines are typical approach trajectories). (Figures adapted from Bohus and Horvitz [2009a] copyright ACL)

orientations, denoted as *Side* and *Front*, as illustrated in Figure 3.5. While we successfully trained models for each situation, applying the model learned from one spatial orientation over the data collected in the other spatial orientation led to performance degradations.

In another example, Leite et al. [2015] investigated mismatches produced by the number of people that were interacting with a system. Specifically, they trained models for predicting disengagement based on data collected in interactions when participants interacted alone and when they were in a group. Not surprisingly, the best performance was attained under matched conditions. Their results indicated that training with instances from both types of interactions was a good compromise, and that training on the social setting alone was better than training with data from the single participants. Investigations into the generalization capabil-

ity of models remains an open research topic and will hopefully lead to a deeper understanding of robustness in open-world settings.

*Interactivity.* Besides situational mismatches between training and runtime data such as the ones described above, the interactive nature of situated interaction systems can also raise additional challenges for machine learning approaches. The typical development loop involves constructing an early version of the system based on simple heuristic inferences and using it to collect the data required to build more performant machine-learned models. The introduction of new machine-learned models may however alter the system's behavior, and in some cases, these changes might in turn trigger participants to start behaving differently, which may then lead over time to a shift in the distribution of the inputs that the system (and the trained machine-learned model) receive. Such mismatches introduced by the interactive loop are often ignored in practice. Online, continual, lifelong learning approaches (see, for instance, Kapoor and Horvitz [2007], Continual Learning Workshop [2018]) constitute an active area of research and hold promise for enabling interactive systems to continuously adapt their models to changing situations and behaviors.

In summary, multimodal perception and inference is a core capability for situated interaction systems. We have seen that both heuristic and machine-learned approaches can be used to track engagement related variables and we have discussed a number of issues that arise with the latter: challenges with feature engineering, obtaining data and ground truth labels, the need for anticipating the future and for reasoning jointly about coupled relevant variables, and finally the need for robust models that generalize and adapt across situations and over time. We have illustrated these themes in the context of constructing engagement inference models. The same themes resurface however in a variety of other inference tasks related to situated interaction, such as managing turn-taking, reasoning about attention and interruptability, situated language understanding, etc.

### 3.3.3 Action

While accurate perception provides the basis for acting, participating in interaction requires that the system makes timely decisions about which actions to take, and ultimately produces actions in a manner coordinated with the other participants in the interaction. The collaborative, real-time, and multiparty nature of situated language interactions brings to the forefront a number of important challenges with respect to the development of decision-making components.

First, decisions in situated interaction systems often require an integrative approach which leverages information from more than one perceptual system and reasons about multiple people in the scene. Consider, for example, the challenge

of managing engagement. The Receptionist tracks the engagement states, actions and intentions of multiple participants in the scene. While these variables provide the basis for the system's engagement decisions, the system also considers additional inferred variables such as whether the participants present are in a group together (see turns 5–6 in Figure 3.1) and their long-term goals and activities (see turns 29–32 in Figure 3.1).

As with perception, for the simpler cases, the control policies for making decisions can be handcrafted. For instance, the Receptionist uses a heuristic rule whereby the system can clarify whether two participants are together, and, depending on the answer, either continues the interaction with both participants or disengages with one and continues the interaction with the other (the latter is shown in the example displayed in Figure 3.1.) In other cases, the Receptionist can temporarily suspend an existing interaction and engage with a waiting participant to let them know they will be attended to momentarily (see turn 14 in Figure 3.1.)

In another example of heuristic decision-making for engagement, we demonstrated how a situated interaction system that plays a trivia questions game can draw bystanders into the interaction [Bohus and Horvitz 2009c]. In this case, in addition to key engagement variables, the system's heuristic engagement policy also considered the results of inferences about the possible goals (*Play*, *Watch*, *Other*) and the current activities (*Passing-by*, *Interacting*, *Playing*, *Watching*, *Departing*) of each person detected in the scene. The policy broadly worked as follows: if a bystander was detected while the system was interacting, and her activity was *Watching*, the system suspended the current interaction task and attempted to engage the bystander and convince her to join the existing game. The actual prompts toward the bystander were contextualized based on the state of the game at the moment of the interruption and on whether the system had made an earlier attempt at engaging the bystander (see Figure 3.6).

Heuristic approaches to decision making can be sufficient in simple cases. They provide modularity and transparency, and they can be authored, tested, and tuned independently of the perceptual pipeline. However, in the more general case, decision making in situated interaction requires solving complex tradeoffs among multiple courses of action under uncertainty and under time constraints. More principled approaches for decision making can be developed by casting the control problem as a *decision-theoretic* optimization problem. In this case, the controller will be optimized to maximize the expected utility of a series of decisions, under the uncertainty present in the state variables. Formalizing a decision-theoretic controller ultimately requires defining the costs of various possible courses of actions, under different states.

**If a watching bystander *B* is detected during the first three questions**

and the engaged participant *E* has not made any mistakes

- [*S* → *E*]: Hold on a second.
- [*S* → *B*]: Hi. Excuse me, would you like to join in?
- [*B*]: Okay.
- [*S* → *E&B*]: Great. So let's see how well you can do together. Next question [ . . . ]

and the engaged participant *E* has just made a mistake, and the system did not previously engage the bystander *B*.

- [*S* → *E*]: Perhaps you can use some help. Hold on a second.
- [*S* → *B*]: Hi. Do you think you could help out?
- [*B*]: No thanks.
- [*S* → *B*]: Okay. Well, I guess you can just watch for now and if you wait for a bit we'll be done and you can start a new game yourself.
- [*S* → *E*]: Guess you're still on your own. Next question [ . . . ]

and the engaged participant *E* has just made a mistake, and the system did previously attempt to engage the bystander *B*.

- [*S* → *E*]: I think you could really use some help. Hold on a second.
- [*S* → *B*]: Are you sure you don't want to help out? Come on, this is fun.
- [*B*]: Sure.
- [*S* → *E&B*]: Great. So let's see how well you can do together. Next question [ . . . ]

**If a watching bystander *B* is detected during the last question**

- [*S* → *E*]: Excuse me for one moment.
- [*S* → *B*]: We're almost done here. If you wait for a bit we can start a new game right after.
- [*S* → *E*]: Sorry about that [ . . . ]

**Figure 3.6** Engagement control policy for Trivia Questions game. *S* denotes the system, *E* denotes the engaged participant, and *B* denotes the bystander. Figure adapted from [Bohus and Horvitz \[2009c\]](#) copyright Association for Computational Linguistics.

A tradeoff that often arises when taking actions under uncertainty is between waiting to collect more evidence vs. acting immediately, based on the currently available evidence. This [wait-vs.-act tradeoff](#) arises because of the typical dynamics in the confidence of inferences made by perceptual systems: often, uncertainty can be reduced as more evidence is being accumulated over time; at the same time, important opportunities for acting may disappear, or action might have a much higher cost if the system waits too long.

A concrete example of this tradeoff in the realm of managing engagement is discussed in [Rosenthal et al 2013a]. The analysis centers on the Assistant [Bohus and Horvitz 2009b], a physically situated agent deployed outside an office that provides several administrative functions for its owner, including identifying good times for interruptions, capturing and sharing messages on behalf of the owner, and scheduling meetings when the owner is away. In a baseline version, the Assistant engages only with participants who approach and enter into an F-formation with it. With this type of *user-initiative* policy, the Assistant misses important opportunities to assist people who take a seat in a waiting area adjacent to the Assistant. If the system had the ability to initiate engagement, it could convey important information to the visitor, such as the case where the owner is running 10 min late, so as to avoid a situation where the visitor would leave in frustration a few minutes before the owner returns. Creating this type of mixed-initiative engagement policy requires solving this wait-vs.-act tradeoff. Ideally, the Assistant should initiate engagement if the person waiting has a scheduled meeting with the owner, but not otherwise (the system should not gratuitously engage everyone that sits nearby). The longer the Assistant waits, the more likely it is that the face recognition system may catch the visitor in a favorable pose and be able to detect with high reliability whether the visitor is indeed the person on the calendar. However, the longer the Assistant waits, the chances increase that the visitor leaves before the system has a chance to engage. Rosenthal et al [2013a] describe an approach for resolving this type of wait-vs.-act tradeoff based on developing models that allow the system to predict its future beliefs, in essence attempting to answer the question: If I wait longer, will I know better—and what is the cost of waiting?

A similar wait-vs.-act tradeoff arises when trying to manage disengagement. Given the challenges with perception, a situated interaction system may decide to maintain engagement and launch a spoken contribution right at the moment or very shortly before an engaged participant might decide to disengage and leave. We have observed this behavior in the *Directions Robot*, a system using the Nao humanoid robot that we developed to provide people with directions inside our building [Bohus and Horvitz 2014]. If the robot had waited a bit longer, it would perhaps have been correctly able to understand that the person is disengaging and avoid issuing a new contribution. However, waiting is inappropriate and problematic if the participant actually wants to maintain the engagement and continue the interaction. To mitigate this tradeoff, we trained forecasting models that anticipate whether the user is about to disengage in the next few seconds (as discussed in Section 3.3.2.3) and used them in conjunction with a policy that allowed the system to

inject linguistic hesitations such as filled pauses (e.g., “So . . .”, “Let’s see . . .”) to buy more time [Bohus and Horvitz 2014].

In resolving wait-vs.-act tradeoffs, the ability to anticipate future states and to develop forecasting models is important. As we discussed, labels for forecasting models can sometimes be collected automatically during the interaction; the system eventually finds out what happens in the future. This eases the burden of model development and provides interesting opportunities for online learning. We believe developing control policies that leverage hesitations and, more generally, that reason explicitly about the time involved in order to mitigate uncertainty and the cost of waiting, is an important area of research. Advances with handling the interplay between time and quality of inference will help to enhance the naturalness and fluidity of language interactions with people in physically situated, open-world settings.

We have presented key issues and themes that arise when making decisions in interactive systems. First, decisions must be made under perceptual uncertainty and time constraints. Heuristic approaches can be devised and used successfully, especially in cases where perceptual systems are accurate enough to provide a good basis for decision making. When it is difficult to resolve uncertainties with heuristics and the tradeoffs become more complex, more sophisticated probabilistic and decision-theoretic approaches can be leveraged. We have highlighted the importance of reasoning about time in making decisions and discussed wait-vs.-act tradeoffs that often arise in these systems. As with perception, the challenges that arise in managing engagement control decisions extend to other communicative processes. For instance, wait-vs.-act tradeoffs play an important role in managing turn-taking in multiparty settings.

## 3.4

### Conclusion

We provided an overview of situated interaction, marking a shift from audio-only, single-user, task-oriented dialog, to systems that reason about their physical surroundings and engage in conversation with people in open-world, multiparty settings. These systems are embodied and perceive the world around them via multiple sensors. Beyond spoken turns, situated interaction systems need the ability to leverage ongoing perceptions and inferences about participants’ body poses, head and hand gestures, gaze directions, as well as about the locations and affordances of objects in the environment, and the spatial relationships among participants and objects. Multimodal capabilities and associated enabling technologies lie at

their core. The development of situated interaction systems requires careful consideration of sets of conversational competencies, and the development of computational models for key processes, including engagement, turn-taking, language understanding, and interaction planning that are deeply anchored in the physical context.

To illustrate the state of the art, as well as challenges and opportunities ahead with situated interaction, we focused on the challenge of managing engagement in physically situated, open-world settings. We discussed the importance of choosing an appropriate representation as a basis for perception and action. We have reviewed heuristic and machine learning approaches for perception and highlighted several important themes that arise, including feature engineering, forecasting, joint modeling, and generalization. Similarly, with respect to decisions and actions, we presented simple yet powerful heuristic approaches and discussed important tradeoffs that arise, such as wait-vs.-act, and means for addressing them.

The key principles and themes we covered on engagement highlight the broader spectrum of challenges that arise in building situated interaction systems. Many fundamental problems are made salient with attempts to build systems that operate with streaming, uncertain sensory data, and that have to closely coordinate their actions with several participants. Consider the problem of managing turn taking in an open-world, multiparicipant interactive scenario. As with engagement, the traditional approaches to turn-taking, such as policies based on voice-activity-detection, are no longer sufficient. In multiparicipant settings, people may talk to each other, and multiple turns may occur among others before the system should contribute to an interaction. Novel representations are necessary to enable the system to reason not only about when speech is happening, but also about who is talking, whom they are talking to, and who should next take the conversational floor. Making inferences about turn-taking in multiparicipant settings requires considering both audio and visual streams of evidence, including non-verbal behaviors, such as gaze and visual focus of attention, gestures, head nods, which all play important roles in signaling whom utterances are addressed to and who is expected to talk next. As with engagement, maintaining responsiveness is essential. Making timely turn-taking decisions and, more generally, engaging in natural, fluid interactions hinges on the ability to recognize, understand and synthesize utterances incrementally e.g., [DeVault et al. 2009, Skantze and Schlangen 2009, Schlangen and Skantze 2011], and to forecast when someone will end or start talking e.g., [Skantze 2017, Roddy et al. 2018]. Finally, the type of wait-vs.-act tradeoffs we have discussed in the context of engagement are perhaps even more important with respect to turn-taking: waiting can help reduce uncertainty, but this may also

lead to important lost opportunities or to awkward pauses and delays in the dialog e.g., [Bohus and Horvitz 2011].

Moving ahead, we see opportunities to make progress in situated interaction via harnessing recent progress in natural language, perception, inference and prediction, and decision making, including advances in all these areas enabled by machine learning. We believe that situated interaction research can also benefit greatly from leveraging insights and understandings developed in multiple fields outside of computer science, including anthropology, psycholinguistics, sociolinguistics, microsociology, and conversation analysis. Beyond engagement and turn taking, other areas of multimodal research play a fundamental role in developing competent situated interaction systems. These include reasoning about space and proxemics e.g., [Michalowski et al. 2006, Mumm and Multu 2011], spatial language understanding e.g., [Tellex et al. 2011, Ma et al. 2012], situated reference resolution e.g., [Prasov and Chai 2010, Giuliani et al. 2010], issues of grounding e.g., [Clark and Brennan 1991, Clark and Schaefer 1989, Clark and Wilkes-Gibbs 1990, Clark 1996], short- and long-term memory models e.g., [Rosenthal et al 2013b], and affective issues [McDuff and Czerwinski 2018]. Adopting and solving challenges in physically situated language interaction will also continue to benefit from and push forward the state of the art in multimodal inference, reasoning, and decision making. Ultimately, we believe that solving key challenges with situated interaction will help us to achieve the long-standing dream of fluid and natural human-machine collaboration.

### Acknowledgments

We would like to thank Sean Andrast, the editors, and the anonymous reviewer for their helpful suggestions and feedback.

### Focus Questions

- 3.1. Describe some of the important aspects that distinguish situated interaction systems from traditional spoken dialog systems.
- 3.2. What is conversational engagement?
- 3.3. Describe why inferences and decisions about engagement can be challenging.
- 3.4. Enumerate some of the key variables useful in a representation for modeling engagement, and some of the signals, cues and features that can be used to make inferences about engagement.

- 3.5.** What tradeoffs are faced when using heuristic vs. machine learned models for making inferences about relevant variables (such as engagement) in physically situated interactive systems?
- 3.6.** Explain the difference between smoothing, tracking and forecasting.
- 3.7.** Consider the process of taking turns in multi-participant conversation and describe the wait-vs.-act tradeoff that a physically situated agent might encounter in this context.
- 3.8.** How would situated interaction systems provide benefits over existing approaches to human-computer interaction? Provide one or more examples or scenarios.

## References

- M. Argyle and M. Cook. 1976. *Gaze and Mutual Gaze*, Cambridge University Press, New York. [116](#)
- T. Baltrušaitis, C. Ahuja, L. P. Morency. 2018. Challenges and applications in multimodal machine learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107993](https://doi.org/10.1145/3107990.3107993). [127](#)
- S. Bianco, G. Ciocca, P. Napoletano, and R. Schettini. February 2015. An interactive tool for manual, semi-automatic and automatic video annotation. In *Computer Vision and Image Understanding*, vol. 131, pp. 88–99. DOI: [10.1016/j.cviu.2014.06.015](https://doi.org/10.1016/j.cviu.2014.06.015). [128](#)
- D. Bohus and E. Horvitz. 2009a. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGdial'2009, pp. 244–252. London, UK. [119, 123, 125, 127, 129, 130](#)
- D. Bohus and E. Horvitz. 2009b. Dialog in the open world: platform and applications. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI 2009, pp. 31–38, Boston, MA. DOI: [10.1145/1647314.1647323](https://doi.org/10.1145/1647314.1647323). [106, 108, 120, 122, 123, 134](#)
- D. Bohus and E. Horvitz. 2009c. Models for multiparty engagement in open-world dialog. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGdial'2009, pp. 225–234. London, UK. DOI: [10.3115/1708376.1708409](https://doi.org/10.3115/1708376.1708409). [117, 132, 133](#)
- D. Bohus and E. Horvitz. 2009d. Open-world dialog: challenges, directions and prototype. In *Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Pasadena, CA. [118](#)

- D. Bohus and E. Horvitz, 2011. Decision about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI 2013, pp. 153–160. Alicante, Spain. DOI: [10.1145/2070481.2070507](https://doi.org/10.1145/2070481.2070507). 137
- D. Bohus and E. Horvitz. 2014. Managing human-robot engagement with forecasts and . . . um . . . hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI'2014, pp. 2–9. Istanbul, Turkey. 119, 123, 125, 128, 129, 134, 135
- H. Brugman, and A. Russel. 2004. Annotating multi-media/multimodal resources with ELAN. In *Proceedings of the 4th Conference on Language Resources and Evaluation*, LREC 2004. Lisbon, Portugal. 128
- L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung. 2018. The MatchNMingle dataset: a novel multisensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. In *IEEE Transactions on Affective Computing*, 1-1. DOI: [10.1109/TAFFC.2018.2848914](https://doi.org/10.1109/TAFFC.2018.2848914). 128
- G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI 2009, pp. 119–126. Cambridge, MA. DOI: [10.1145/1647314.1647336](https://doi.org/10.1145/1647314.1647336). 119, 123, 125
- Continual Learning Workshop. 2018. <https://sites.google.com/view/continual2018/home>. Accessed December 2018. 131
- H. H. Clark, and S. A. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, Washington, DC. DOI: [10.1037/10096-006](https://doi.org/10.1037/10096-006). 110, 137
- H. H. Clark, and E. F. Schaefer. 1989. Contributing to discourse. In *Cognitive Science*, 13(2): 259–294. DOI: [10.1016/0364-0213\(89\)90008-6](https://doi.org/10.1016/0364-0213(89)90008-6). 110, 137
- H. H. Clark, and D. Wilkes-Gibbs. February 1990. Referring as a collaborative process. *Cognition*, 22(1): 1–39. Reprinted in: P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Massachusetts. 110, 137
- H. H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge. 110, 111, 114, 137, 782
- D. DeVault, K. Sagae, and D. Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue* (SIGDIAL 2009), pp. 11–20, London, UK. 136
- G. Ferguson, and J. Allen. 1999. TRIPS: The rochester interactive planning system. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference*, AAAI 1999/IAAI 1999, pp. 906–907, Orlando, FL. 105

- M. E. Foster, A. Gaschler, and M. Giuliani. 2013. How can I help you?: Comparing engagement classification strategies for a robot bartender. In *Proceedings of the 15th International Conference on Multimodal Interaction*, ICMI 2013, pp. 255–262, Sydney, Australia. DOI: [10.1145/2522848.2522879](https://doi.org/10.1145/2522848.2522879). [119](#), [126](#)
- M. Giuliani, M. E. Foster, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. 2010. Situated reference in a hybrid human-robot interaction system. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG 2010, pp. 67–75, Dublin, Ireland. [137](#)
- E. Goffman. 1963. *Behaviour in Public Places: Notes on the Social Order of Gatherings*, The Free Press, New York. [116](#)
- E. Goffman. 1979. Footing. In *Semiotica*, 25(1-2): 1–29. [111](#), [114](#), [782](#)
- E. Goffman. 1981. Footing. In E. Goffman, *Forms of Talk*, pp. 124–159. University of Philadelphia Press, Philadelphia. [111](#), [114](#), [782](#)
- E. T. Hall. 1966. *The Hidden Dimension: Man's use of Space in Public and Private*. Doubleday New, York. [114](#), [121](#), [782](#)
- E. Horvitz, J. S. Breese, M. Henrion. 1988. Decision theory in expert systems and artificial intelligence. *Journal of Approximate Reasoning*, 2: 247–302. DOI: [10.1016/0888-613X\(88\)90120-X](https://doi.org/10.1016/0888-613X(88)90120-X). [113](#), [766](#)
- E. Horvitz. May 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the CHI '99, ACM SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA. DOI: [10.1145/302979.303030](https://doi.org/10.1145/302979.303030). [114](#), [787](#)
- A. Kapoor and E. Horvitz. 2007. Principles of lifelong learning for predictive user modeling. In *Proceedings of the 11th International Conference on User Modeling*, UM 2007, pp. 37–46, Corfu, Greece. DOI: [10.1007/978-3-540-73078-1\\_7](https://doi.org/10.1007/978-3-540-73078-1_7). [131](#)
- A. Kendon. 1990a. A description of some human greetings. In *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Studies in International Sociolinguistics, Cambridge University Press, Cambridge. [116](#)
- A. Kendon. 1990b. Spatial organization in social encounters: The F-formation system. In A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Studies in International Sociolinguistics, Cambridge University Press, Cambridge. [110](#), [113](#), [116](#), [771](#)
- G. Keren, A. El-Desoky Mousa, O. Pietquin, S. Zafeiriou, and B. Schuller. 2018. Deep learning for multisensorial and multimodal interaction. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces*, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107996](https://doi.org/10.1145/3107990.3107996). [127](#), [129](#)
- M. Kipp. 2001. Anvil—A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, Eurospeech, pp. 1367–1370, Aalborg, Denmark. [128](#)

- W. S. Lasecki, M. Gordon, D. Koutra, M. F. Jung, S. P. Dow, and J. P. Bigham. 2014. Glance: rapidly coding behavioral video with the crowd. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST 2014, pp. 551–562, Honolulu, HI. DOI: [10.1145/2642918.2647367](https://doi.org/10.1145/2642918.2647367). 128
- I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati. 2015. Comparing models of disengagement in individual and group interactions. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI 2015, pp. 99–105. Portland, OR. 119, 130
- Y. Ma, A. Raux, D. Ramachandran, and R. Gupta. 2012. Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGdial 2012, pp. 169–178, Seoul, South Korea. 137
- D. McDuff, and M. Czerwinski. December 2018. Designing emotionally sentient agents. In *Communications of the ACM*, 61(12): 74–83. 137
- M. P. Michalowski, S. Sabanovic, and R. Simmons. 2006. A spatial model of engagement for a social robot. In *Proceedings of the 9th IEEE Workshop on Advanced Motion Control*, AMC 2006, pp. 762–767. Istanbul, Turkey. DOI: [10.1109/AMC.2006.1631755](https://doi.org/10.1109/AMC.2006.1631755). 119, 121, 123, 137
- L. P. Morency, I. de Kok, and J. Gratch. 2008. Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI 2008, pp. 181–188, Chania, Crete, Greece. DOI: [10.1145/1452392.1452426](https://doi.org/10.1145/1452392.1452426). 113, 126, 770
- J. Mumm and B. Multu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI 2011, pp. 331–33 137
- D. S. Pallett, N. L. Dahlgren, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, and B. C. Tjaden. 1992. DARPA February 1992 ATIS Benchmark Test Results. In *Proceedings of the Workshop on Speech and Natural Language*, pp. 15–27, Harriman, NY. DOI: [10.3115/1075527.1075532](https://doi.org/10.3115/1075527.1075532). 105
- Z. Prasov and J. Y. Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2010, pp. 471–481, Cambridge, MA. 137
- C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, HRI 2010, pp. 375–382, Osaka, Japan. DOI: [10.1109/HRI.2010.5453163](https://doi.org/10.1109/HRI.2010.5453163). 119, 124, 125
- M. Roddy, G. Skantze, and N. Harte. 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI 2018, pp. 186–190. Boulder, CO. DOI: [10.1145/3242997](https://doi.org/10.1145/3242997). 129, 136

- S. Rosenthal, D. Bohus, E. Kamar and E. Horvitz. 2013a. Look vs. leap: computing value of information with high-dimensional streaming evidence. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, IJCAI 2013, pp. 2561–2567. Beijing, China. [134](#)
- S. Rosenthal, S. Skaff, M. Veloso, D. Bohus, and E. Horvitz. 2013b. Execution memory for grounding and coordination. Late-breaking report. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI 2013, pp. 213–214, Tokyo, Japan. [137](#)
- J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI 2011, pp. 305–312, Lausanne, Switzerland. DOI: [10.1145/1957656.1957781](https://doi.org/10.1145/1957656.1957781). [119](#), [123](#)
- D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. In *Dialogue and Discourse*, 2(1): pp. 83–111. [136](#)
- C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pp. 78–84, Madeira, Portugal. DOI: [10.1145/964442.964458](https://doi.org/10.1145/964442.964458). [110](#), [113](#), [116](#), [771](#)
- G. Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SigDial 2017, Saarbrucken, Germany. DOI: [10.18653/v1/W17-5527](https://doi.org/10.18653/v1/W17-5527). [129](#), [136](#)
- G. Skantze, J. Gustafson, and J. Beskow. 2018. Modelling face-to-face conversational interaction with robots. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool, San Raphael, CA.
- G. Skantze and D. Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2009, pp. 745–753, Athens, Greece. [136](#)
- D. Szafir and B. Mutlu. 2012. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2012, pp. 11–20, Austin, TX. DOI: [10.1145/2207676.2207679](https://doi.org/10.1145/2207676.2207679). [119](#), [123](#)
- S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI 2011, pp. 1507–1517, San Francisco, CA. [137](#)
- G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tur, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S.

- Mason, J. Niekrasz, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang. August 2010. The CALO meeting assistant system. In *IEEE Transactions on Audio, Speech and Language Processing*, 18(6). DOI: [10.1109/TASL.2009.2038810](https://doi.org/10.1109/TASL.2009.2038810). [105](#)
- R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2001, pp. 301–308, Seattle, WA. : [10.1145/365024.365119](https://doi.org/10.1145/365024.365119). [116](#)
- Q. Xu, L. Li, and G. Wang. 2013. Designing engagement-aware agents for multiparty conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2013, pp. 2233–2242, Paris, France. DOI: [10.1145/2470654.2481308](https://doi.org/10.1145/2470654.2481308). [119](#), [123](#), [126](#)
- M. A. Walker, A. I. Rudnicky, J. Aberdeen, E. O. Bratt, J. S. Garofolo, H. Hastie, A. N. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, S. Roukos, G. A. Sanders, S. Seneff and D. Stallard. 2002. DARPA communicator evaluation: progress from 2000 to 2001. In *Proceedings of the 7th International Conference on Spoken Language Processing*, ICSLP 2002, Denver, CO. [105](#)
- G. Wagner, and E. Andre. 2018. Real-time sensing of affect and social signals in a multimodal framework: a practical approach. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Kruger, editors, *Handbook of Multimodal-Multisensor Interfaces*, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition. Morgan and Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108000](https://doi.org/10.1145/3107990.3108000). [123](#)
- J. Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. In *Communications of the ACM*, 9: 36–45. DOI: [10.1145/365153.365168](https://doi.org/10.1145/365153.365168). [105](#)
- T. Winograd. February 1971. Procedures as representation for data in a computer program for understanding natural language. *MIT AI Technical Report 235*. [105](#)





# **Software Platforms and Toolkits for Building Multimodal Systems and Applications**

**Michael Feld, Robert Neßelrath, Tim Schwartz**

## **4.1**

### **Introduction**

This chapter introduces various concepts needed for the realization of multimodal systems. Alongside an overview of the evolution of multimodal dialogue platform architectures, we give an overview of the major components found in most of today's architectures: input and output processing; fusion and discourse processing; dialogue management; fission and presentation planning; and middleware. We compare several different dialogue management approaches, look in more detail at how the fusion component works, and introduce dialogue act annotation with communicative functions. We will explain the multimodal reference resolution process and consider the special case of cross-modal references. Finally, we present SiAM-dp, an actual multimodal dialogue platform used in a number of research projects and prototypes and highlight some of its particular features.

## **4.2**

### **Definitions**

A dialogue takes place when more than one agent (human or machine) communicate with each other in a structured and multidirectional way, i.e., a request in one direction may trigger a response in another.

A *dialogue system* is a software agent that allows users to converse with a machine in a coherent structure. The origins of dialogue systems are in spoken dialogue but over time, many additional modalities have been employed in multimodal dialogue systems, where text, speech, haptics, graphics, gestures, and other modes are combined. Some systems are purely informational, while others serve as an interface to services, and still others allow manipulation of devices, such as controlling robots. They can be distinguished, for example, by the way they accept user input, how they allow conversations to flow, by the number of simultaneously supported users, and by the ability to incorporate, preserve, and refer to context (including discourse).

A *dialogue platform* is an underlying framework that is used to execute a dialogue system, and that is often accompanied by tools employed by dialogue system designers to create dialogue applications. The advantage of using a dialogue platform is that it is not linked to a particular dialogue system, but its core components and services can be consumed by any number of dialogue applications.

In this chapter, we will use the term *dialogue application* in order to highlight the fact that a dialogue system was written for a specific application use-case and is backed by a dialogue platform, such that the application-specific part consists only of those features and models not already present in the framework.

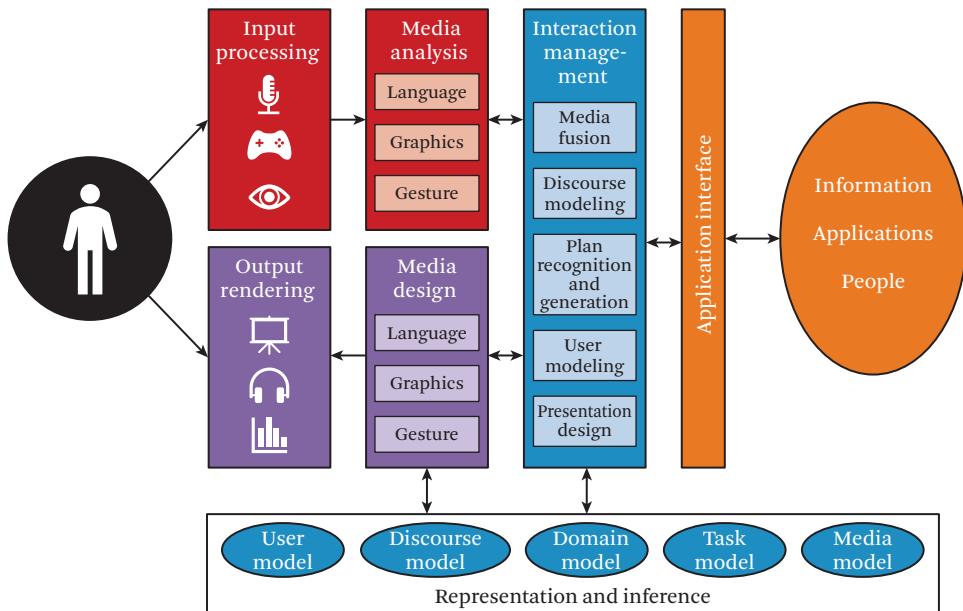
## 4.3

### Architecture of Dialogue Systems

Despite the wide range of available tools for modeling dialogue systems, we can identify a set of core architecture components that appear in many of these utilities in one or another shape. We present these basic dialogue platform “building blocks” in this section and introduce their function. We will also point out to special interpretations of these components found in certain tools, if applicable.

Figure 4.1 shows a reference architecture for intelligent multimedia user interfaces as introduced by Maybury and Wahlster [1998]. The architecture has later been extended by Bunt et al. [2005]. The typical architecture comprises three sequential processing phases.

1. Analyzing and understanding user input. This includes modality specific recognizers and analyzers, and fusion as well as discourse processing.
2. Dialogue management and action planning.
3. Planning and generation of system output by modality fission and modality specific generation and realization.

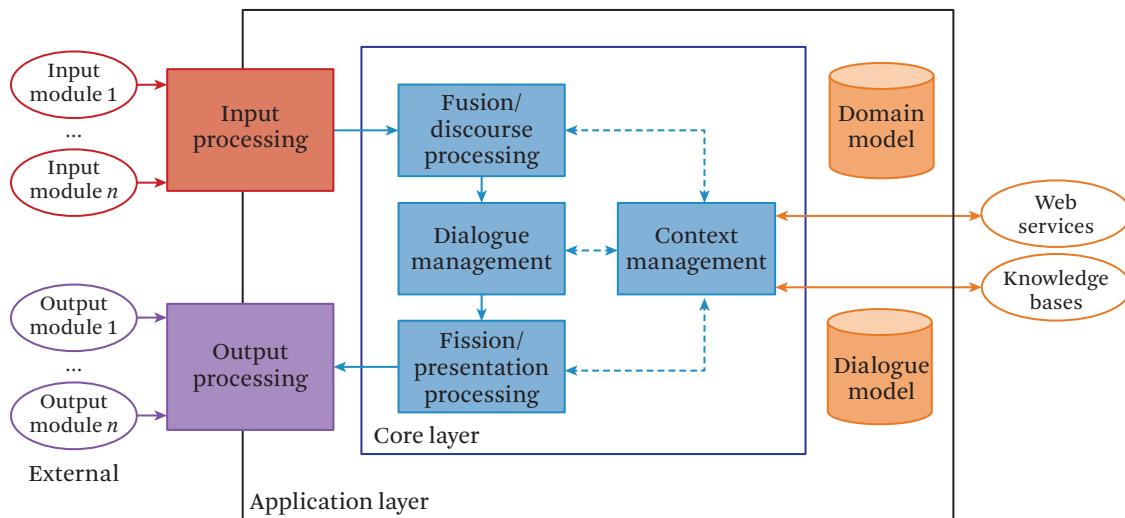


**Figure 4.1** Early reference architecture for intelligent user interfaces (c.f. [Maybury and Wahlster \[1998\]](#)).

Furthermore, the architecture includes models for users, discourse, context, domain, task, media, and applications that are accessible by all components.

We are slightly generalizing this model and move some of the pre-processing logic, such as the Media Mode analysis, into what we call the Input and Output Processing layer. Thus, the resulting dialogue platform architecture consists of six components, which are depicted in Figure 4.2 and explained in further detail in this chapter. The same figure also introduces a categorization of these components into three layers.

- The *core layer* contains the components that are essential to the platform and are provided out-of-the-box. They are usually tightly coupled and are connected via an internal communication mechanism or via middleware components.
- The *application layer* refers to components that extend the platform in order to realize particular dialogue applications or advanced dialogue features.



**Figure 4.2** Generalized architecture for multimodal dialogue platforms with added layers.

These components are not generally considered part of the dialogue platform, but may be added by users of the platform to add features without having to modify the core platform.

- The *input/output processing layer* describes, which modalities are available, how the information is structured and obtained etc. It is separate from the rest of the platform since it should be possible to introduce new modalities without affecting other parts.

### 4.3.1 Input and Output Processing

In a strict sense, input refers to interaction performed by the user to explicitly trigger a reaction from the system, while output is a way of the system to present information to the user. With the increasing pervasiveness of cyber-physical environments hosting a large number of sensors and actuators [Wahlster 2014], this concept is extended to also include sensor information as part of the input to the dialogue and actuator control as part of the output. This allows a proactive dialogue system to also implicitly trigger a reaction based solely on sensor information, such as the user being present in a certain spot. It also allows modeling the operation of a device as an integral interaction that directly affects context and turn management.

Input encompasses a more or less abstract description of the information from the device or sensor that captures it. There can be more than one syntactical way to

express some information. It is the task of input processing to transform low-level or raw information into higher-level and possibly annotated information, which usually reduces and abstracts data. This process, however, is independent of the dialog state and context. Consider a generic gesture recognition input module. The camera sensor records a series of images recording the user's movement. This data stream could be considered the lowest level of representing the user input. A segmentation component would then split the stream into meaningful blocks, producing a higher level of information. Next, an image analysis component could transform the gestures into a series of skeletal junction points, thereby producing a completely different but still very coarse representation. The coordinate series could then be mapped to a named discrete gesture such as "thumb up" or "thumb down."

At this stage, one might consider adding an interpretation such as "agreement" or "disagreement" to the input. A final interpretation should however be generally deferred to the fusion and discourse processing step, since it might involve context knowledge of the situation. Different cultures may have different understandings of gestures, which could not be resolved without resorting to user knowledge, and the same gesture may also have a different meaning in another discourse context. An exception to this general rule is the case of **early fusion**: Here, sensor information is fused with other sensor information on the input processing level. For example, the hand movement data could be combined with raw sensor data from a head tracking component to merge both into a single gesture, which is then easier to interpret.

### 4.3.2 Multimodal Fusion and Discourse Processing

A fusion and discourse processing module attempts to obtain the meaning (semantics) of an input and resolves ambiguities of interaction acts which are caused by different interpretations or the absence of context information when observing single interactions of a user, within a single modality or across modalities. For example, in case of speech, a Natural Language Understanding (NLU) component might perform a classification of a textual representation into intents, or do a complete semantic parsing of a sentence. It is also possible to generate multiple mutually exclusive hypotheses, for example if an utterance is not clearly understood.

The main source for this are the world and dialogue context. By combining the information contained in multiple inputs from different modalities arriving in close timely manner, new interpretations can be derived and references can be

resolved. While an NLU could consider the dialogue context to interpret the command “a little more please” differently depending on the context of the interaction, most current NLUs are separate from the dialogue and do not work that way. In fact, since there is only a limited number of possible inputs with a fixed meaning, simple dialogue systems often merely map inputs to pre-defined interpretations.

A further task of this step is to select between multiple inputs. If an utterance is not clearly understood, for example, a speech recognition may generate multiple mutually exclusive hypotheses. In the Multimodal Fusion and Discourse Processing step, the system could eliminate hypotheses not matching the discourse or re-rank them accordingly.

A more thorough look at the aspects of the fusion process is presented in Section 4.5.

### 4.3.3 Dialogue and Meta Dialogue Management

The dialogue manager is the component in a dialogue system, which controls the architecture and structure of the dialogue. Traum and Larsson [2003] define the following functions as the main tasks of a dialogue manager:

- updating the dialogue context on the basis of interpreted communication. The communication can originate from the human user, the system itself, or any other connected software agent;
- providing context dependent expectations for interpretation of observed signals as communicative behavior;
- interfacing with task/domain processing (e.g., database, planner, execution module, other back-end system), to coordinate dialogue and non-dialogue behavior and reasoning; and
- deciding (e.g., applying rules based on the current context) on any output that is to be generated or content to be expressed next and when to express it.

While the first two points concern the management of the dialogue context and the context-based interpretation of communication, the latter two rather handle the control of the conversation with the user. A distinction is made between user-initiative, system-initiative, and mixed-initiative systems, whereby the initiative belongs to the speaker in control of the conversation [Jurafsky and Martin 2009]. *User-initiative systems* are typical command and control systems. In a *system-initiative system*, the conversation is completely controlled by the system. Thus, the system asks a question to the user and solely reacts on inputs of the user that exactly answer the question. An improvement of this concept is universal commands that

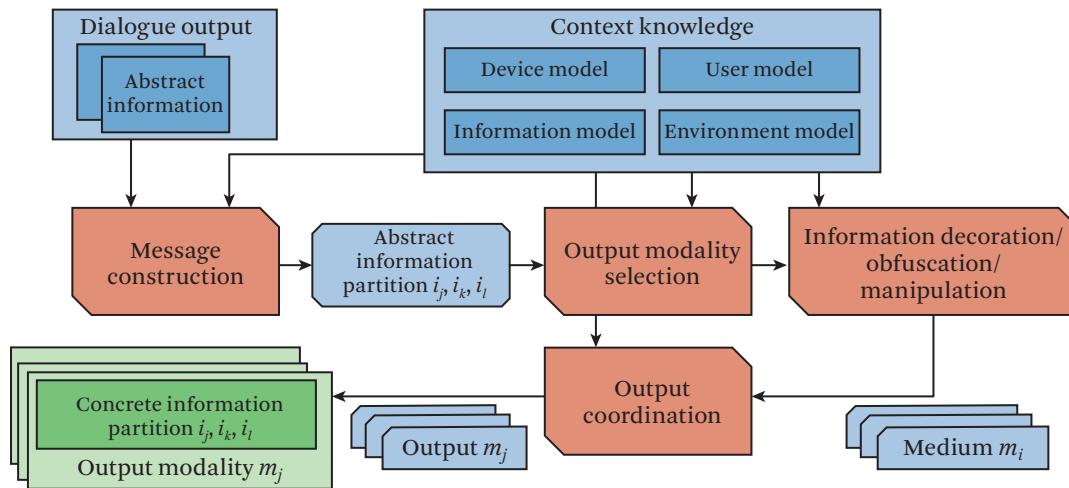
can be said anywhere in the dialogue and for example provide the user a shortcut for the navigation to a help or a main menu.

Often a user wants to communicate something that is not exactly the answer to a specific question, maybe in a sentence that provides more than one relevant piece of information. In a restaurant-reservation example, this would be the sentence “I want to reserve a table at five p.m. on Monday for three people.” Here, besides the day and the time, the number of people for the reservation is also given. This type of interaction is more natural for humans. Systems that support such a shift of conversational initiative are called *mixed-initiative systems*. Another example for mixed-initiative are systems where the user would normally initiate requests as needed, but the system can proactively render material from a different context between user turns.

We further distinguish between *dialogue management* and *meta-dialogue management* layers. While dialogue management usually refers to a specific and application-dependent flow of information, meta-dialogue management revolves around more general concepts that are repeated over the course of the dialogue in what is typical mixed-initiative. An example is back-channeling, which refers to feedback given by the user on somebody else’s turn, such as nodding or looking puzzled. Another example are turn management functions that mediate between speakers. Meta-dialogue logic should be implemented on top of other dialogue logic in order to avoid having to consider each and every case where they might occur. For example, instead of implementing dialogue logic for a user feedback such as “I did not understand this” after each turn in a dialogue model explaining a task to the user, the preferred way would be to globally listen for this type of question and handle the response more generally on the meta-dialogue level, e.g., by asking clarification questions (which can nevertheless be based on the dialogue context).

#### **4.3.4 Multimodal Fission and Presentation Planning**

The previous section discussed the fusion of modalities during input; in this section we look at the output side. Modality fission is the process of splitting semantic representation from an intended modality-free output into a multimodal presentation to be realized. Like the term fusion, the term fission has also been borrowed from physics and emphasizes that the output is split into several output modality presentations. Modality fission includes the decision of how the output is channeled and coordinated throughout the diverse available output modalities based on the user’s perceptual abilities and preferences [Costa and Duarte 2009]. Foster [2002] subdivides the tasks of a multimodal fission component into three parts.



**Figure 4.3** The multimodal fission process for the creation of context aware output from abstract information.

- Content selection and structuring: Often the content to present is already selected by the dialogue management component and provided on a semantic level. In the first step, a fission component must divide the overall meaning into elementary elements that can be presented to the user. Here the main approaches are schema-based [McKeown 1985] or plan-based [Moore 1995].
- Modality selection: In the next step the devices and modalities are selected that contribute to the multimodal output. The available devices are described by several features that include the type of information they can handle, the perceptual task they permit, the availability, the characteristic of information to present, resource limitations, and the user's profile, which includes his abilities, skills, and impairments.
- Output Coordination: After the output modalities are selected, the modality specific realizations must be created. For these modality attributes, spatial and temporal parameters, as well as user characteristics, etc., are considered. Especially for multimodal outputs with cross-modal references, the scheduling, synchronization, and coordination of presentations play a large role.

Honold et al. [2012] implement this concept in an adaptive probabilistic approach for multimodal fission. Their fission process is depicted in Figure 4.3. Outgoing from a modality-independent dialogue output, they first partition the data

items into elementary data. Then they use a probabilistic reasoning approach for the selection of the devices and modalities to involve. In the next step they post-process the output data and, e.g., obfuscate private messages if only public devices for presentation are available. In the last intermediate step, the final concrete output realizations are created for each information item. Finally, the output is distributed between the target output devices.

All the processing steps involve context knowledge about available devices, users, and the environment in their decisions. [Wasinger et al. \[2003\]](#) present several features from diverse context sources that affect the presentation output planning. [Endres \[2012a\]](#), [Endres \[2012b\]](#) presents *PrestTK*, a platform for the situation-aware presentation of messages and infotainment content for drivers. The main goal of the platform is to present a dramatically increasing amount of information from various in-car systems to the driver without increasing the risk of driver distraction. In the first step, the system applies techniques from scheduling and presentation planning in order to avoid conflicts when competing for scarce resources such as screen space. In the second step, the system considers the cognitive capacity of the driver.

[Wasinger \[2006\]](#) also sets a focus on symmetric multimodality that has been introduced by [Wahlster \[2003\]](#) within the SmartKom project. The main statement herein is that all input modes should also be available for output, and vice versa: “only true multimodal dialogue systems . . . create a natural experience for the user in the form of daily human-to-human communication, by allowing both the user and the system to combine the same spectrum of modalities” [[Wahlster 2003](#) (p.2)]. Thus, the presentation planner must take into account the modality a user applied for input, and adequately adapt the selection of output modalities. For example, speech input is responded to with speech output, interaction with a GUI is responded to with GUI updates and if a user performs a pointing gesture in order to identify an entity in the room, the reaction of the system could use light spots in order to highlight the referred entity or maybe even use a robot arm. One important design principle in the SmartKom project [[Wahlster 2006](#)] was “no presentation without representation”. This means that the generated multimodal presentations must be explicitly represented in order to ensure the dialogue coherence in multimodal communication. This plays a relevant role for the resolution of anaphoric, cross-modal, and gestural references of the user. In the SmartKom system, a text generator provided a list of referential items that were mentioned in the last turn of the system. A display management component permanently kept track of the currently presented screen content in an internally managed model for the display context.

### 4.3.5 Middleware

The aforementioned dialogue system components need to exchange information in order to reach their respective goals. User input needs to be forwarded from the input processing component to the fusion component, and its analyzed and annotated interpretation to dialogue management. Even within the fusion component, references may need to be resolved in multiple passes while further hypotheses are being generated, and presentation planning may involve a complex set of device-dependent incremental output generation rules. Depending on the desired degree of extensibility and performance, this communication mechanism needs to satisfy certain quality criteria. There are numerous architectures available in AI known to be applicable for this type of task. Examples are a simple bus system, a black-board communication architecture, an event-based messaging queue architecture, or a publisher-subscriber pattern. In Figure 4.2, this component is implicitly represented by the connecting arrows between the other components.

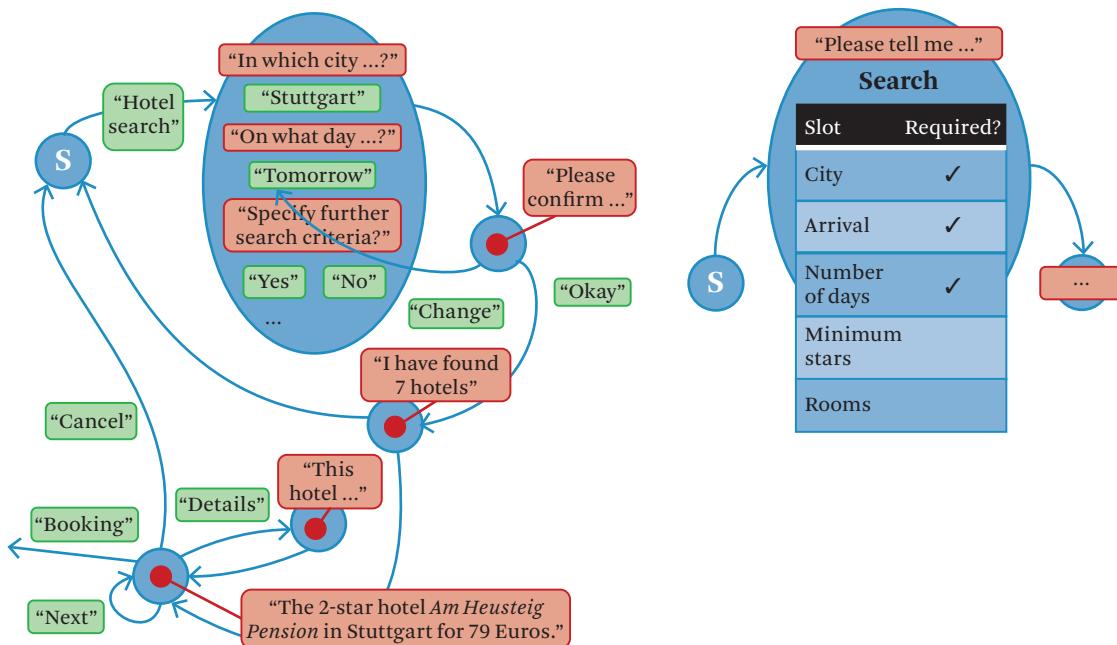
## 4.4

### Dialogue Management Architectures

The choice of the dialogue management architecture is crucial to a successful implementation, and depends on the type of application to be implemented (e.g., pro-active assistant, question answering system, troubleshooting chatbot, social bot etc.). Today, dialogue system designers can choose from a vast amount of architectures and combine them within a single dialogue application to suit their needs. While older architectures follow a relatively strict model-based approach with a strong impact of expert knowledge, newer variants are more dynamic and incorporate statistical, machine-learned aspects. Some of the presented list originates from [Bui 2006, Jurafsky and Martin 2009].

#### 4.4.1 Finite State-based

A dialogue manager based on this model uses a finite set of states (usually representing system turns) and transitions (usually representing user turns) connecting them. This approach is simple yet flexible. It is very well suited for less complex dialogue systems with well-structured tasks. However, for handling a mixed-initiative dialogue, a finite-state architecture is inappropriate due to the enormous number of states that would be required to handle each possible subset of questions and answers. There are numerous software libraries available for executing state machines with high efficiency. In many cases, more complex types of dialogue systems such as frame-based systems can be translated into finite state charts. While these machine-translated or “compiled” state charts are still fast to execute, they may no



**Figure 4.4** Left: Example dialogue state chart. Right: Example frame-based dialogue.  $S$  denotes the starting state. Green boxes represent user turns, while red boxes represent system turns.

longer be human-readable. An example dialogue modeled as a finite state chart is depicted in Figure 4.4 (left). In this user-initiative hotel booking example, a subset of possible states and transitions with alternating system and user turns is shown. This figure also presents the concept of *composite* or *compound* states, in which a single state can host an arbitrary number of other states or even a different dialogue management architecture.

#### 4.4.2 Frame-based

This approach is analogous to a form-filling task in which a predetermined set of information is collected. A frame-based dialogue manager collects information from the user by asking questions until enough information is available to perform a task. A method for dynamically generating the corresponding questions can be part of such a dialogue manager, or it could choose from a set of pre-formulated questions. If a user happens to answer more than one question within a single turn, the system has to fill the appropriate slots. This concept is illustrated in Figure 4.4

(right), where the hotel booking information from the previous example is modeled as a frame that contains *required* and *optional* information about the booking. Contrary to required information, the dialogue manager will not proactively ask for optional information once all required slots have been filled in. The first frame-based dialogue manager was presented by [Bobrow et al. \[1977\]](#).

#### 4.4.3 Information State-based

This architecture consists of five components: information state, dialogue act interpreter, dialogue act generator, a set of update rules, and a control structure that selects which update rules to apply. The term “information state” is quite abstract and might include things like the discourse context and the common ground of dialogue participants, their beliefs or intentions, user models, environment models, and so on. Thus, in contrast to a static state in the finite state-based approach, the information state is more complex and includes the values of many variables, the discourse context, and other elements. The update rules are responsible for modifying the information state based on the information of the dialogue acts. One subset of these rules is called selection rules and is used to generate dialogue acts in order to control the dialogue. For more information on the information state approach and the design of update functions, see, e.g., [Traum and Larsson \[2003\]](#).

#### 4.4.4 Plan-based Dialogue Agents

Plan-based approaches are based on the idea that people communicate in order to achieve goals, which includes the change of the mental state of the listener. Thus, plan-based models are often referred to as beliefs, desires, and intentions (BDI) models, which were first introduced by [Perrault and Allen \[1980\]](#) and [Cohen and Perrault \[1979\]](#). In the plan-based theory, the speaker’s speech act is part of a plan and it is the listener’s job to identify and respond to this plan. Communication and conversation are thus just special cases of rational actions in the world that can be planned as any other action by applying AI planning techniques like the TRIPS agent [\[Allen et al. 2001\]](#).

#### 4.4.5 Probabilistic

Recent approaches represent the underlying structure of a dialogue using probabilistic models [\[Lison 2012\]](#). For this, rules are specified using high-level conditions and effects and are defined as structured mappings over variables of the dialogue state. Nowadays, probabilistic models such as Bayesian Networks are in widespread use in spoken dialogue systems, but their scalability to complex interaction domains remains a challenge. Probabilistic models should help to make dialogue

system more robust against noise and uncertainty and to be capable of automatically learning and optimizing from data, making them more flexible and adaptive.

## 4.5

### Fusion and Communicative Functions

[Johnston et al. \[2009\]](#) define *multimodal integration* as the process of combining input from different modes to create an interpretation of composite input. A synonym is the term *multimodal fusion*, which is adopted from the terminology in physics. Multimodal fusion is a process that combines manifold types of input data, each associated with a particular modality. It is a fundamental task in the integration of various modalities. For an in-depth review of this task and the applied techniques, see Chapter 1.

#### 4.5.1 Classification of Multimodal Input

[Nigay and Coutaz \[1993\]](#) call the absence of fusion *independent modalities* and its presence *combined modalities*. [Serrano and Nigay \[2009\]](#) organize the combination space of interaction modalities into two dimensions, the type of the relationship between modalities and the temporal relationship. The type of relationship is explained with the CARE properties [[Coutaz et al. 1995](#)]. The CARE properties (Complementary, Assignment, Redundancy, and Equivalence) characterize multimodal interaction from the usability perspective on HCI. They are a set of properties that describe the relationship between modalities for reaching a goal or the next state in a multimodal system (see the [Glossary](#)).

Equivalence and Assignment are independent modalities and can be interpreted individually from one another. Redundancy and Complementary are combined modalities and require a multimodal fusion of the input. While redundant input must be compared and verified for an identical meaning, the complementary input must be combined in order to express the meaning.

#### 4.5.2 Temporal Relationship and Synchronization

The temporal relations can provide relevant indications of whether and how multimodal input should be combined. Early multimodal systems like the “Put-That-There” system by [Bolt \[1980\]](#) relied on the fact that multimodal constructions temporally co-occur. The meaning of the deictic term “that” in the spoken utterance “put that there” for example was resolved with the object at which the user was pointing when it was spoken.

This multimodal integration approach seems to be suitable for multimodal speak-and-point systems but has a restricted practical use in the design of future

### Glossary

**Assignment** expresses the absence of choice. Exactly one specific modality can be used in order to reach a goal. An example is the steering wheel of a car.

**Back-channeling** refers to feedback given by the user on somebody else's turn, such as nodding or looking puzzled.

**Complementary** modalities are used in complementary manner within a temporal window for reaching a goal, i.e., both modalities are needed to describe the desired meaning. A speak-and-point system is a classic example of this (“change the color of this (*pointing gesture*) item to blue”).

**Cyber-physical Environments** are characterized by a large number of individual systems and devices with their sensors and actuators, and shift the interaction paradigm from the user’s perspective toward system-environment interaction.

**Dialogue Applications** are dialogue systems that are written for specific application use-cases.

**Dialogue Management** controls the flow of the dialogue with a computer. It basically updates the context and determines how to react to dialogue acts.

**Dialogue Platforms** are underlying frameworks that are used to execute a dialogue system.

**Dialogue Systems** are software agents that allow users to converse with a machine in a coherent structure.

**Discourse** describes the sequence of dialogue acts, which in turn are a communicative function unit.

**Equivalence** expresses the concept of free choice of modality. Multiple modalities can reach the same goal and it is sufficient to use only one of them without any temporal constraint on them.

**Meta-Dialogue Management** deals with those aspects not related to a particular domain or dialogue application, typically particular types of communicative functions.

**Mixed-initiative system** allow either user or system to start a dialogue, and further allow for a shift of initiative within the discourse.

**Modality Fission** is a technique where a single semantic content is spread over multiple (complementary or alternative) output channels.

**Modality Fusion** describes the process of resolving the semantic intent of a dialogue act by combining different input modalities.

**Redundancy** is present when two modalities have the same expressive power, but are both required to be used within a temporal window in order to reach a goal. Redundancy can be important for safety relevant functions.

**Glossary** (*continued*)

**System-initiative Systems** typically represent dialogues in which the system asks a question to the user and solely reacts on inputs of the user that exactly answer the question.

**Turn management functions**, such as requesting or assigning a turn, are a family of communicative functions used in multi-party dialogue to mediate between agents.

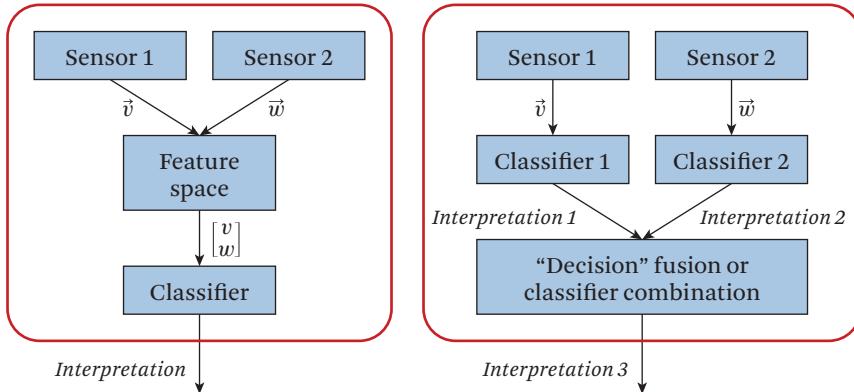
**User-initiative Systems** are typically command and control systems.

multimodal systems that involve other modes like gestures or body movements without deictic point relations [Oviatt 2012]. It turned out that the temporal overlap of signals not urgently determines which signals should be combined. A series of studies showed that there exist two distinct types of users with respect to integration patterns and that their integration patterns persist across the lifespan from children through the elderly [Xiao et al. 2002, Xiao et al. 2003]. An integration pattern here specifies the strategy of how users combine multimodal input with respect to the temporal relation. Simultaneous integrators overlap their input temporally, whereas sequential integrators begin with one mode after the other one has been finished [Oviatt 1999b, Oviatt et al. 2005]. Since a user's habitual integration pattern remains highly consistent during a session, this may allow systems to automatically detect and adapt to a user's dominant multimodal integration pattern. This may also include the temporal thresholds during the sequential use of modalities.

### 4.5.3 Fusion Level

An important aspect for multimodal fusion is the appropriate fusion technique that is applied to combine incoming unimodal events into a single representation of the user's intention. Literature often distinguishes between two stages where fusion occurs: *early fusion* and *late fusion* [Turk and Kölisch 2003, Jaimes and Sebe 2007, Nigay and Coutaz 1993]. There are also hybrid approaches that attempt to combine continuous and discrete signals (see [Cohen and Oviatt 2017]). The decisive factor here is the level of abstraction at which the fusion takes place.

*Early fusion* occurs at *feature* or *sub-symbolic* level. The input signals are concatenated and provided to a joint classifier that generates an interpretation (see



**Figure 4.5** Multimodal fusion on distinct levels. (Based on Oviatt and Cohen [2015b, p.70]). *Left:* Early fusion at feature level. *Right:* Late fusion at semantic level.

Figure 4.5 left). The interpretation (or classification) is mostly based on machine-learning techniques like neural networks or hidden Markov models. A classic example for early fusion is the audio-visual combination of speech and lip movements. Here, the motion data from the lips are concatenated with features from the recorded voice in order to recognize a spoken utterance [Tamura et al. 2004]. For example, Schuller et al. [2009] are using early fusion on a large set of visual and acoustic features (facial expression, eye activity, voice features, linguistic features etc.) to determine the user's interest in the current dialogue topic, which is valuable meta-dialogue information.

During the *late fusion* or *decision fusion*, the signals are first classified independently on the feature level. After that, the results are combined to a joint interpretation (see Figure 4.5 right). The late fusion is realized on the *semantic* or *symbolic* level and techniques like unification on graphs or Bayesian networks are employed in order to combine information. Late fusion is used, for example, in QuickSet [Cohen et al. 1997], Match [Johnston et al. 2002], SmartKom [Wahlster et al. 2001], MuDiS [Schwarzler et al. 2008], and the multimodal ShopAssist [Wasinger 2006] (e.g., speech plus gesture: “What is the price of this (*user points at product*) digital camera?”).

Atrey et al. [2010] mention several advantages of the late over the early fusion. One is that the interpretations at a semantic level have the same form making their fusion easier. A second one is that for each single modality, the most suitable methods for analyzing the input data can be applied, making the process more flexible than the early fusion. Wahlster [2003] explains that at a semantic level, the backtracking and reinterpretation of a result is easier. Furthermore, the development

process is less complex since multimodality with new modalities can be handled without specifying all varieties of cross-modal references in advance. [Oviatt and Cohen \[2015b\]](#), p. 71] point out that the development process is simplified because commercial “black box” recognizers can also be applied, which provide no access to their internal state or data. Late fusion is able to fuse modalities that are not time-synchronous. With early fusion, the feature vectors usually have a close temporal boundary. However, early fusion can gain potentially useful information that would already have been discarded when the late fusion is applied [[Wasinger 2006](#)].

#### **4.5.4 Dialogue Act Annotation**

A limited dialogue system may focus solely on the content of a dialogue act, which today is commonly represented in dialogue platforms by an intent identifier and optionally entities set to a value (a concept similar to variables in a programming language). This may be sufficient to look up a pre-entered answer from a database or to start a pre-programmed flow. However, in order to successfully build a conversational model, this simple description of conversational behavior is not sufficient. Even if the content of two conversations is equal, the actual set of exhibited behaviors differs from person to person and conversation to conversation [[Cassell 2000](#)]. This also means that one particular behavior can be employed in a variety of circumstances to produce distinct communicative effects and the same communicative function may be realized through various sets of behavior. The actually applied communicative behavior is dependent on the type of conversation, availability of modality, cultural patterns, and also the personal style.

Thus, in order to build a model for the description of how conversation works, it is necessary to identify high-level structural elements of interaction and their communicative function in the discourse. These elements describe a role or function in a contribution, e.g., conversation invitation, turn taking, providing feedback, or task requests [[Cassell 2000](#)]. Especially the examination and annotation of linguistic data at the semantic level in terms of their communicative function experience a growing interest resulting in various dialogue act annotation schemes. The maybe most widely used one is DAMSL (Dialogue Act Markup using Several Layers) by [Allen and Core \[1997\]](#).

The scheme DIT<sup>++</sup> [[Bunt 2009](#)] extends this scheme with the possibility to annotate multi-functional communicative acts, which is necessary because of the phenomenon that utterances in a dialogue often have more than one communicative function [[Bunt 2011b](#)]. For example, the request “Henry, could you take us through these slides?” contains, besides the function to assign the turn to Henry, also a specific request. It is argued that this phenomenon can be explained by analyzing the participation in a dialogue as involving the performance of several

types of activities in parallel, relating to the different dimensions of communication [Bunt 2011a]. Therefore, the annotation language of dialogue acts must be multi-dimensional, which means that multiple tags must be assignable to a single dialogue act [Petukhova 2011].

Since human communication is not only restricted to verbal communication, also nonverbal and multimodal dialogues can be annotated with these schemes. Petukhova and Bunt [2012] argue that nonverbal behavior may emphasize the intended meaning of synchronous verbal behavior. It may also express a separate dialogue act in parallel to a verbal speech act expressed by the same speaker and add multi-functionality to a communicative contribution. It can even be utilized to model parts of the dialogue in modality-agnostic way, by modeling based on functions rather than concrete interactions.

## 4.6

### Multimodal and Cross-Modal Reference Resolution

Often contributions in a dialogue have to be interpreted with respect to the actual context, which includes the context of the world, the actual discourse, but also coherent contributions of other input channels (e.g., Bunt [2000]). Hence, in multimodal dialogue systems it is inevitable to incorporate contextual information in order to resolve linguistic phenomena like referring expressions.

#### 4.6.1 Referring Expressions

Referring expressions are a key linguistic phenomenon in verbal utterances for the identification of specific entities in the real world. The referred entity is called the *referent* [Webber 1978]. Pfleger [2007] gives an extensive introduction of the role and the various types of referring expressions from a linguistic point of view. *Deixis* or *deictic expressions* is a group of referring expressions that refer to “some entity or concept of the physical, situational, or discourse context.” They are dependent on the contextual information and can only be interpreted if the context is considered. Since they play a relevant role in the representation of meaning in communicative acts, we give a short overview of some referring expressions Pfleger [2007] supported in his discourse resolution for multimodal dialogue.

- **Anaphora.** The term anaphora stems from the Greek word meaning “carrying back” and is a reference to an entity of the preceding discourse. In linguistics, an anaphoric expression is a pronoun or a nominal phrase that is linked to a noun which has been previously introduced in an utterance.
- **Place Deixis.** Describes a spatial reference either relative to the participants of a communicative act or to other entities in the context. The object with

respect to which the figure is located is called *relatum*. The important thing for the resolution of a spatial deixis is the *frame of reference*. This can be intrinsic if the speaker takes the viewpoint of the relatum; relative, if the object is located relative to another object; or absolute, if an unambiguous reference point is used.

- **Time Deixis.** Describes time points or time spans that are relative to the time point when a communicative act was produced. In natural language, adverbs like *then*, *now* or *tomorrow* express a temporal deixis. Temporal deixis also comprises complex compound temporal references like “next Monday” which consist of an adverb and a non-deictic name or unit of time that is modified.
- **Exophoric References.** These are references to the visual or situational context of the discourse. This can, e.g., be an object in the physical environment or objects that are presented on a graphical display.
- **References to Collections.** Sometimes an expression references an object in a collection of possible referents, e.g., in a list of entities. Here a differentiation criterion like “the third entry” describes the referred object.
- **Cross-modal References.** In a multimodal contribution, sometimes the content of one input modality refers to content that is provided by a second modality. For example, in the combination of a pointing gesture with the utterance “What is this building?”, the pronoun refers to the entity that is indicated by the pointing gesture. There exists plenty of research in this area [Oviatt et al. 1997, Kruijff et al. 2006, Prasov and Chai 2010, Schutte et al. 2010]. Recent work in this field includes Moniri et al. [2012], who used the speaker’s location in conjunction with the looking direction and the instruction given in the speech to resolve landmarks from within a driving vehicle.
- **Ellipsis.** In an elliptical construction, one or more words of an expression are omitted, e.g., if some of the constituents have already been mentioned in a previous turn. This especially may occur in information-seeking dialogues. The following example shows an elliptical construction: *User*: “What is the menu for today?” *System*: (Presents the actual menu) *User*: “And for tomorrow?”

### 4.6.2 Constraints on References

Communicative acts that contain referring expressions can additionally provide restrictions on the referred object which is valuable information for the resolution of matching referents. Pfleger [2007] introduced two types of constraints.

- **Syntactic Constraints.** A referring expression can contain linguistic information about number, person, and gender of the referent. Usually these features must match the result of the reference resolution.
- **Semantic Constraints.** A referring expression can also contain semantic information about the referent. For example in the utterance “turn on this lamp,” it is semantically clear that the user refers to a lamp and not to, e.g., a ventilator in a room. Semantic constraints can provide information about the type but also about features of an object like in the utterance “the green lamp.”

## 4.7 Review of Existing Dialogue Platforms

In this section, we will examine some of the established dialogue platforms.

### 4.7.1 AT&T Speech Mashup Architecture

The AT&T speech mashup architecture [Di Fabbrizio et al. 2009] aims to simplify the combination of web content with speech processing and makes speech recognition and text to speech (TTS) synthesis available by web services. The main idea is the support of speech practitioners and researchers in the easy and rapid development of speech and multimodal mobile services. The architecture consists of four components.

- Speech Mashup Server (SMS)—This server provides the services for speech recognition and speech synthesis. Furthermore, the server contains the AT&T speech mashup server that handles the direct connections between the connected client devices, including resolving device-dependency issues, performing authentication, and general accounting.
- Speech Mashup Client—A client application that runs on the mobile device which presents the UI (e.g., iPhone, Safari browser). The client sends and receives audio buffers for speech input and output and receives the recognition results from the server.
- Main Application Server—A web service (e.g., Apache or Tomcat). Depending on the application, the server provides access to a back-end database, collects and aggregates data from other application servers and could implement the application logic.

A mobile client application establishes connections to the application server and the SMS. Between the client and SMS, data and speech are exchanged. The

speech either consists of the voice of the user that is recorded for speech recognition and is sent to the server, or the synthesized speech output that is sent to the client for audio output. The speech processing resources are centrally managed in a web-based portal that supports rule-based and stochastic grammars. For rule-based grammar specification, the Speech Recognition Grammar Specification (SRGS) standard is used. One task of the web portal is the grammar management which allows one to extend the shared grammar rule-set with personalized user grammars. For the representation of recognition results, three different language formats are supported: XML, JavaScript Object Notation (JSON), and the EMMA markup language, a W3C standard for the inter-operable input representation in multimodal systems. TTS tasks are described with the W3C Speech Synthesis Markup Language (SSML).

The applications that have been implemented with this framework range from speech-based information retrieval from a business directory to the ordering of food from a pizza service. The applications combine graphical interfaces with speech interaction, the possible input is a mix of voice and touch inputs. Nevertheless, the support of cross-modal interaction is not mentioned.

#### **4.7.2 WAMI Toolkit**

The Web-Accessible Multimodal Interfaces (WAMI) toolkit [Gruenstein et al. 2008] provides a framework for the web-based development, deployment, and evaluation of multimodal user interfaces. The supported interaction channels are speech input, speech output, and graphical user interfaces that can be controlled by mouse, pen or touch, depending on the used device. The main goal of the framework is to be able to rapidly build applications that are accessible from outside a laboratory in order to more easily perform user evaluations. A secondary goal is that the toolkit supports a lightweight development model for non-expert dialogue application developers in order to build interactive multimodal applications. The framework supports desktop, laptop, and tablet PCs.

The framework uses a server-client architecture where the main application runs on the server. On the client, a web browser presents the GUI application in HTML and AJAX while an audio controller handles the audio stream between device and server. Speech recognition and synthesis are realized on the server. Thus, the client only has the task to establish the interface to the user; the program logic is located on the server. Grammars are specified using the Java Speech Grammar Format (JSGF), GUI information is either sent in final HTML that is rendered in the browser or in XML messages that update the content of an application specific GUI on the client. As example applications, some web-based showcases are presented

that support speech based interaction as well as interaction with the GUI. Cross-modal interactions are not included.

### 4.7.3 DIANE

The speech dialogue system DIANE (DIAlogmaschiNE) is a frame-based dialogue system for speech interaction developed by Siemens CT [[Song 2006](#), [Block et al. 2004](#)]. Applications are modeled with DIANE as a set of transactions, where a transaction is considered similar to a frame. Each function provided by the back-end application corresponds to one transaction in the dialogue model. Additional information that is required for the execution of a transaction is modeled as parameters of the transaction. At the example of a transaction for train ticket reservation, these parameters are information about departure, destination, date, and time of the itinerary. For each of these parameters an own grammar and prompts for query or confirmation are defined. Furthermore, trigger grammars for the direct access to the transactions can be defined. Siemens developed DIANEXML, an XML-based dialogue design language for supporting the development process of speech user interfaces. This design language is used to automatically generate the run-time resources for the DIANE dialogue system. A dialogue application is specified by three kinds of documents. With a transaction file the application is defined as a set of executable transactions. Additional necessary information is defined in a parameter file the transaction file can refer to. The grammar files provide important information for speech recognition and language understanding. Further functionalities like callback functions in order to specify inference rules, consistency conditions, repair mechanisms and to invoke back-end functions, are implemented by a Java interface.

The framework focuses on the frame-based dialogue management approach, which allows to build mixed-initiative dialogue applications. The support of the resolution of discourse phenomena and integration of additional modalities are not mentioned.

### 4.7.4 Dialog OS

The Dialog OS [[Bobbert and Wolska 2007](#)] is an extensible platform for the development of dialogue systems with the focus on spoken language. It is an educational tool that allows students with even non-technological background to develop relatively complex applications with flexible strategies for various domains. It is a commercial product that has been developed by the former CLT Sprachtechnologie GmbH.

The system mainly focuses on speech based dialogues but can be extended with new components over a communication API, for example actuators like a LEGO Mindstorm robot or an elevator. The dialogue modeling approach is based on a Finite state automaton (FSA) that can be built in a GUI workspace where dialogue nodes of the FSA are created and linked to each other. Nodes can be input nodes, output nodes or nodes that perform internal actions. The latter can be the execution of a JavaScript or the assignment of variables. Furthermore, sub-automata can be executed that describe recurring parts of the dialogue. Input nodes allow one to define a list of expected input values either as plain text or regular expression. If an incoming text or speech-based input message matches the value, the FSA follows the outgoing edge to the next node. Output nodes can be used to trigger TTS tasks.

#### **4.7.5 SmartKom**

The SmartKom project [[Wahlster 2006](#), [Wahlster 2003](#)] was one of the largest projects worldwide that examined multimodal interaction. The result of SmartKom was a multimodal dialogue system that combined speech, gesture, and facial expression for input and output. One focus was set on symmetric multimodality. For this, a virtual character was used to communicate with the user via speech, pointing-gestures, and facial expressions by imitating human characteristics. For input, the system supported gesture input recognized by an infrared camera, speech recognition, and a camera for facial analyses. Handwriting and hand contour recognition were used only during biometric signature identification. Furthermore, physical actions were recognized and used in order to validate whether the user executed a correct action in order to complete a collaborative task. SmartKom was built on an architecture for distributed components called *Multiplatform*, which is an open, flexible, and scalable software architecture that allows one to integrate heterogeneous software modules written in diverse programming languages and running on various platforms. In total, 40 asynchronously running modules were included in SmartKom. The data interface between the components is covered by the M3L (Multimodal Markup Language).

The integration and mutual disambiguation of multimodal input and output is solved by statistical and symbolic methods that are processed on both the semantic and pragmatic level. The uncertainty and ambiguity during the analysis of the various input modalities is corrected with the help of scored hypothesis graphs that stem from the user interactions. The speech recognizer provides word hypothesis graphs, the prosody component clause and sentence boundary hypothesis graphs; the gesture recognizer hypotheses about possible referenced objects, and the facial expression interpreter about the emotional state of the user. They are unified

by the fusion component using uni-cation and overlay and the resulting interaction hypotheses ranked by the intention recognizer. Here also, the particular context of the multimodal discourse model is considered making the final ranking highly context sensitive. The presentation planning in SmartKom is solved with a plan-based approach. The presentation goal for the planner is encoded in a modality-free representation. The goal is recursively decomposed into primitive presentation tasks. For this, the system contains 121 presentations strategies that are parameterized by the discourse context, the user model, and the environmental context. Finally, each presentation task is sent to the appropriate generator for concrete output realizations.

The reference resolution in SmartKom is based on a three-tiered multimodal discourse model. This consists of a discourse layer, a domain layer, and a modality layer. The discourse layer stores information about every discourse object mentioned. Since it is multimodal, this comprises verbal as well as visual and the conceptual context which includes all visible objects on the screen and the spatial relationships between them. Each of the stored objects can have multiple surface realizations on the modality layer. Each element in the discourse layer is also linked to an instance in the ontology-based domain model of SmartKom. The discourse model allows one to address the following multimodal dialogue discourse phenomena: multimodal deixis resolution and generation; multimodal anaphora resolution and generation, cross-modal reference-resolution and generation; and multimodal ellipsis resolution and generation. Furthermore, the dialogue management supports strategies for turn-taking and back-channeling.

#### 4.7.6 ODP

The Ontology-based Dialog Platform (ODP) framework [Porta et al. 2014, Neßelrath and Porta 2011] evolved from the SmartKom dialogue shell and was developed with the aim to enable developers to easily design and implement homogeneous ODP-based UIs for services. The complexity of the back-end architecture should be hidden from the user. The development mainly took place during the THESEUS research program that was funded by the Federal Ministry of Economics and Technology with the goal of developing new technologies and methodologies for the Internet of Services.

The ODP framework provided a platform and development methods for the rapid creation of multimodal dialogue applications with the main task to retrieve heterogeneous data from web services [Sonntag et al. 2009]. A declarative programming approach helped UI experts, who are not necessarily Java programming experts, to build applications. A main concern was to offer more abstract and comprehensible work levels instead of working on the most concrete level of abstrac-

tion. For example, the platform provides an advanced domain-independent grammar specification language. The language allows one to directly map utterances to a semantic interpretation of the input. Furthermore, algorithms are supported that handle the dynamic generation and management of named entity grammars. Rule-based speech synthesis templates allow a generic output generation strategy. A GUI model that can directly be connected to the semantic data model fulfills the paradigm “no presentation without representation” and is used to represent interaction-relevant graphical components. The platform out of the box supports the multimodal integration of speech recognition, speech synthesis and interaction with GUIs. Therefore, concepts for dialogue management, modality fusion, and discourse resolution have been adopted from the predecessor project SmartKom.

All content in ODP is semantically modeled. This is a necessary requirement for the robust multimodal processing and dialogue management. Information that is retrieved from integrated services must first be transformed into a semantic representation and mediated back into the internal domain model that is based on extended Typed Feature Structures (eTFSs).

Since the end of the project, ODP has been further developed by SemVox GmbH and is sold under the name ODP S3. Heavy improvements have been made regarding the software development kit with the aim that “developing dialog components becomes as easy as developing apps for a smartphone.”<sup>1</sup> The extended ODP S3 workbench provides an integrated tool chain based on Eclipse and supports all steps of the development process, from the specification of dialogues, over system integration, to an improved quality management for commercial and professional use. This, e.g., includes automatically generated test cases and system documentation. The platform itself demands low resources and is available for the target platforms Linux x86/ARM, QNX Neutrino, Android, Windows Embedded, and more. The supported features have also been extended with goal-oriented interaction, task-based dialogue models, hybrid speech recognition, and the consideration of personalization, user models, and cognitive load.

#### **4.7.7 CueMe**

The CueMe<sup>2</sup> software development platform is developed and licensed by Openstream and allows one to create multimodal systems based on the W3C international standard reference architecture (see also [Tumuluri et al. \[2019\]](#) in this book and [Dahl \[2013\]](#)). This standard embraces EMMA, a standard for content in messages for multimodal interaction. The framework is based on the OSGi standard.

- 
1. ODP S3 - The leading dialog and assistance technology, product sheet, SemVox GmbH (2015)
  2. Openstream CueMe, <http://www.openstream.com/cuemef>

The framework collects user input from diverse modalities like type, touch, talk, gestures, handwriting, or stylus and can also incorporate sensor information from on-device peripherals like camera, GPS, card reader, and bar-code scans. The information can be distributed across multiple devices, like smartphones, tablets, and PCs whereby all current operating systems are supported.

The system is context-aware and allows one to consider information from “hard sensors,” such as location awareness, and “soft sensors” such as user preferences. For example, if the user is moving (as in a car) the framework allows one to use speech interaction technology for navigation and control. Multimodality increases the robustness of applications by reducing data entry errors and common mistakes. One modality of input is hereby used to validate the content of other input modalities, which is called *mutual disambiguation*. An evaluation of a system employing complementary modalities has been investigated by [Oviatt \[1999a\]](#) in a user study with the QuickSet framework.

An interaction manager is responsible for the coordination of information. For this, the *Openstream MAM Server* holds application definitions that help to map collected input information to appropriate methods on connected application servers. A wide variety of enterprise SOA (Service Oriented Architecture) back-end systems are supported. The spectrum of applications covers form filling, question answering, personal information management, annotation of drawings and more and is settled in domains like health-care, financial services, media and entertainment, and utility and transportation sectors [[Oviatt and Cohen 2015a](#)] (see Chapter 15). Various corporations like Walmart, Merck, Roche, and the Bank of NY/Mellon are mentioned to have implemented multimodal applications based on the CueMe framework.

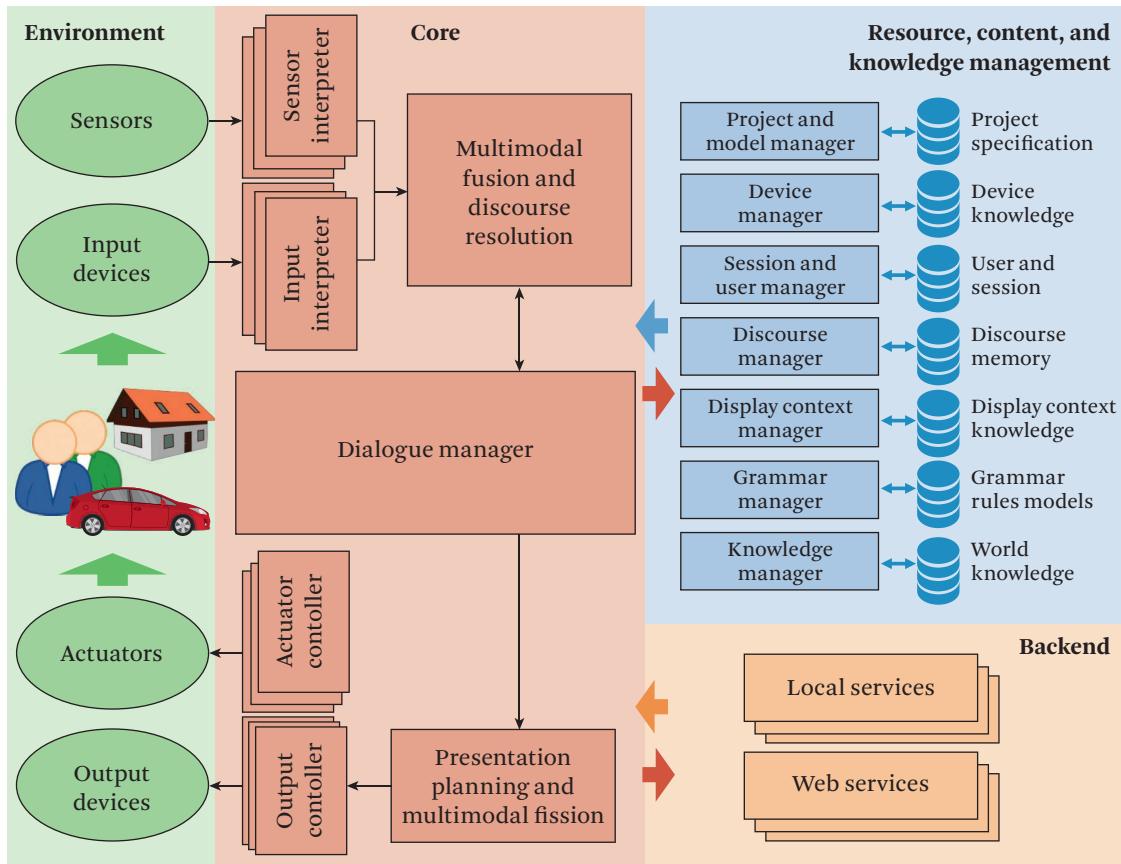
## 4.8

### **SiAM-dp—the Situation-Adaptive Multimodal Dialogue Platform**

The Situation-Adaptive Multimodal Dialogue Platform, in short *SiAM-dp*, was created to allow dialogue engineers to rapidly build dialogue applications that can make use of diverse modalities. Its development originates from a project with the goal to support user interaction between a car, its driver, and any other passengers, though it has now evolved beyond the automotive domain. It was first introduced by [Neßelrath and Feld \[2014\]](#).

#### **4.8.1 Basic Architecture**

One of the key goals of SiAM-dp is a modular architecture, which allows modules to be added, removed, and replaced as needed. This is particularly important for input and output components, since a modern multimodal dialogue platform



**Figure 4.6** The architecture of SiAM-dp.

has to allow the addition of new devices and modalities. Figure 4.6 shows an overview of the architecture, which largely follows the generic architecture outlined in Section 4.3. Each top-level box represents a separate system component, with the *main platform component* representing the largest unit implemented in Java/OSGi. Using OSGi allows modules to be started and stopped independently, and if needed, they can be replaced by custom implementations. SiAM-dp also utilizes OSGi's messaging architecture to enable communication between modules and to process dialogue acts.

The *fusion and discourse processing* module resolves ambiguities of interaction acts which are caused by different interpretations or the absence of context information when observing single interactions of a user. The main sources for this are the world and dialogue context. By combining the information contained in

multiple inputs from different modalities arriving in close timely manner, new interpretations can be derived and references can be resolved. Some of the dialogue phenomena that are recognized by SiAM-dp are deictic references, anaphoras, exophoric references, and spatial references. Closely related to fusion and discourse processing is the *dialogue management*. Its basic function is to determine the implications of user interaction and trigger system reactions, causing the dialogue to progress. In a situation-adaptive system, the context has a great impact on dialogue behavior. Currently, the dialogue manager executes as a finite state automaton, which is a very robust and transparent means for implementing dialogue systems. The automaton itself is also referred to as the *run-time dialogue model*, which can include executable code. A large number of dialogue models can be implemented this way; state charts are a natural way to model many dialogues, with flow charts as an intuitive extension. Due to the finite nature, however, some advanced NLU scenarios cannot be implemented directly and require a replacement of the dialogue management module. The last module in the dialogue platform core is the *presentation planning and fission*. Any output act (or more generally presentation task) passes through this module prior to being sent to devices. It determines where output should be displayed (media allocation), when (scheduling), and in what mode. Using fission, a single output can be split across devices if purposeful. Besides the dialogue management, this is the other module most involved in situation adaptivity. The strategies that describe the general adaptation effects to be performed are contained in *presentation* and *dialogue strategy modules*.

Further contained in the main platform component are *interpreters* and *generators*. They transform or enrich input or output acts, respectively. Interpreters could be used, e.g., for resolving references using context knowledge or for recognizing gestures. Generators might transform a generic modality-independent UI description into a modality-specific interface, or add further representations that are needed to support particular devices. There are some common interpreters and generators, but each dialogue application can add its own custom modules.

The dialogue platform can integrate with input and output devices via the i/o interface, which is a TCP-based message protocol. This allows a flexible and language-independent integration of new devices and modalities, including microphones, speakers, touchscreens, graphical UIs, tangible devices, eye-trackers, gesture recognition devices etc. A single device can cover multiple modalities and provide input, output, or both.

SiAM-dp uses a model based approach based on EMF (Eclipse Modeling Framework) for describing nearly everything: world knowledge, rules for grammars and mappings, input and output interfaces, users, communicative functions, etc. Some

concepts, such as that of states, devices, or communicative functions, are included in a so-called base model. In several areas, this model was designed to adopt existing standards, such as Grammar XML (GRXML) for describing grammars or State Chart XML (SCXML) for modeling dialogue states. Most applications will, however, introduce new semantic concepts that can be referred to in the dialogue, such as places, items, or states. We can extend the base model by creating a new EMF model in the SiAM-dp editor and define any number of concepts. The world state is stored in a separate knowledge base module, which enables easy sharing of information with other components in heterogeneous scenarios.

A key advantage of the unified data model is that dialogue systems can be created in a declarative fashion. A single dialogue system can be created as a *dialogue definition project*, which is essentially a file containing the dialogue state chart, plus some meta information such as the expected users and devices. While these project files could be created with any XML-capable editor, a *toolset* was created as part of the platform which provides comfortable access to a graphical environment in which the dialogue can be developed, including a graphical state designer. This toolset is also used for the dialogue evaluation function that we present as part of the resource awareness feature of the dialogue platform.

#### **4.8.2 Multimodal Dialogue in Cyber-physical Environments**

Supporting multimodal interaction with numerous devices used in intuitive fashion will be of increasing importance in the coming years. While the traditional input modalities to computers (keyboards and mice) will continue to be used in certain areas, most areas of daily life are dominated by other, more flexible modalities, including speech, eyegaze, gestures, etc. There are good reasons for this: For instance, in cyber-physical environments such as the smart home, users want the interaction possibilities to follow them into every room. In the car, speech is a prime modality since it is less distracting while driving than, e.g., a large number of knobs, and micro-gestures have been proposed for eyes-free operation too [Neßelrath et al. 2016]. In industrial environments, non-touch interaction can prevent screens and tangible controls from soil and dirt. Wearable devices such as smart watches with haptic feedback are becoming popular in mobile scenarios since they are more ubiquitously available.

Most of the aforementioned situations favor the use of multimodal combinations for interaction. For instance, in the car, many operations can be performed more easily by combining speech with buttons and switches [Castronovo et al. 2010]. Similarly, eyegaze can be used to quickly locate objects within range of sight in the physical world, but speech can be used to disambiguate when the precision

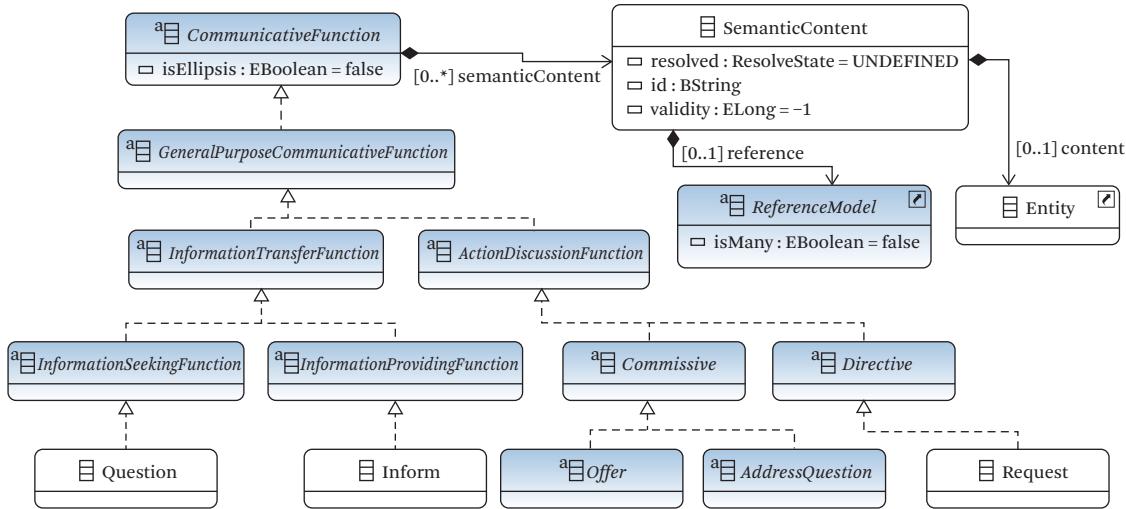
limitations of the eye-tracker are reached. Speech as humans' natural means of communication is probably the most universal modality for multimodal combinations. It therefore has increased priority in SiAM-dp and is supported out-of-the-box for services such as command-based and free text speech recognition, and text-to-speech synthesis.

A modern dialogue platform has to abstract from the actual modality being used, even from whether the input was unimodal or multimodal in nature. Recent studies and demonstrations performed at DFKI's *Advanced Driver Assistance Systems Lab* have utilized speech in combination with tangible devices [Castronovo et al. 2010], touch screens, eye-tracking [Moniri et al. 2012], and micro-gestures [Neßelrath et al. 2016] performed with the hands at the steering wheel of a car.

### 4.8.3 Semantic Dialogue Act Model

In Section 4.5.4 we discussed the fact that a disadvantage of representation on a purely syntactic level is that input or output messages from a device are very modality-specific. As a result, every description of a dialogue workflow must be adapted to each specific device and modality. In practice, when specifying the dialogue workflow, the application developer has to explicitly react to the input message of every input modality and send an output message for each output modality, respectively. In order to make SiAM-dp more flexible and adaptable to new modalities and devices, we introduced a model for dialogue acts that describes the user's or system's communicative intentions. Instances of this model contain the communicative intention of a dialogue act and eventually thereby provided semantic content. Thus, the model for controlling the dialogue workflow can be based on these semantic dialogue acts and be completely independent from the actually used modalities and devices.

The model for the definition of communicative functions is inspired by a standard for the semantic annotation of dialogue acts (ISO/DIS 24617). For the semantic description of dialogue acts, we adopted the type hierarchy of communicative acts that is specified by this standard and integrated it into our model for communicative functions (Figure 4.7 shows an excerpt of the upper level). This type hierarchy, inter alia, contains concepts for seeking and providing information, offering and demanding tasks, or controlling the dialogue, e.g., turn taking or giving feedback. The diagram shows that every communicative function can contain one or more elements of the type *SemanticContent*. This concept is a container for entities that are carried by the dialogue act. We can generally distinguish between resolved and unresolved semantic content elements and a *SemanticContent* instance can adopt one of the following two states.



**Figure 4.7** Upper level excerpt of the communicative function type hierarchy in SiAM-dp using Unified Modeling Language (UML) syntax.

- **Resolved.** Resolved semantic content is an entity that is introduced with the dialogue act into the discourse context. For example, the utterance “What is Harry Potter about?” introduces the book *Harry Potter* into the discourse.
- **Unresolved.** Unresolved semantic content contains referring expressions to entities that have already been introduced by a previous interaction turn (*anaphora*), will be subsequently introduced into the discourse context (*cataphora*), or are part of the environment context. The dialogue act in the utterance “Who is starring in this movie?,” for example, refers to an entity of type Movie that is not implicitly given but is already part of the discourse context.

Entities in the platform must derive from the base concept Entity. This concept also serves as the anchor point for new domain-specific concepts, thus the concepts in new domain specifications should be derived from this basic concept. Already resolved content is added to the content slot of the SemanticContent concept. If the content is unresolved, the information of the referring expression is represented by an instance of the ReferenceModel that uses concepts similar to the approach of Peger (2007). Instances of this model are added to the slot reference. During run-time it is the responsibility of the fusion and discourse resolution

engine to resolve the referring expressions and fill the content slot with entities from the context.

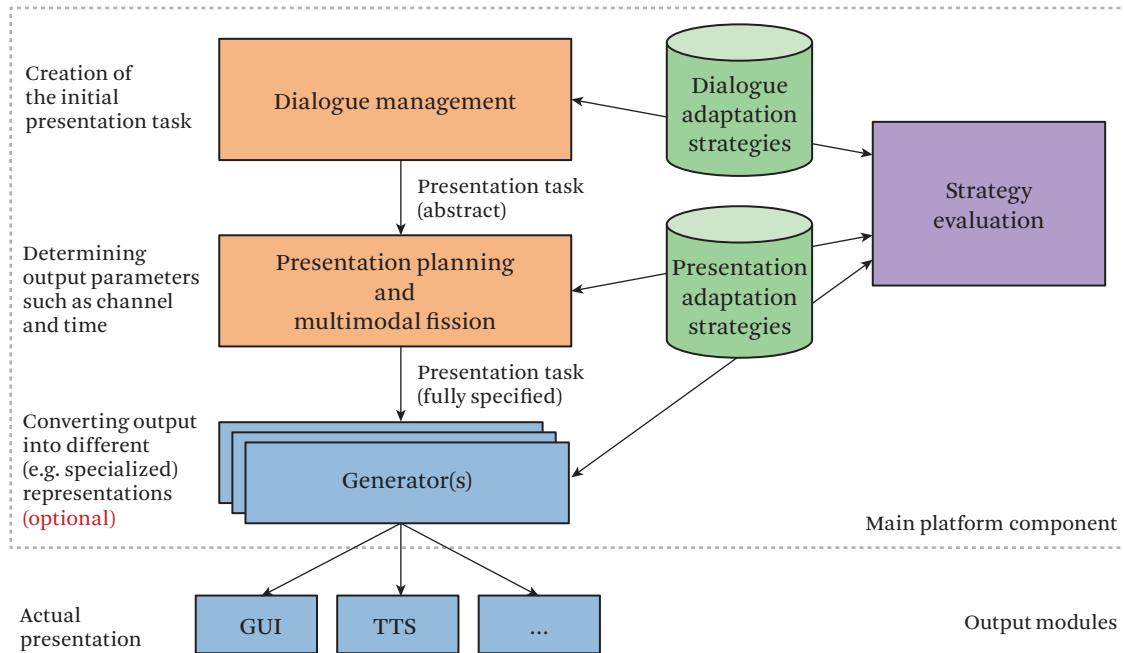
#### 4.8.4 Situation-Adaptive Behavior

In order to adapt to the user and situation, providing the right output at the right time, requires three main aspects to be covered: (1) acquiring knowledge about the user and context on which dynamic decisions can be based; (2) deciding how the knowledge should be structured, stored, and be made available to the platform and other applications; (3) and performing adaptation based on this knowledge by applying certain criteria.

(1) is also known as *knowledge acquisition*. Generally, we assume that for a new user, there is no prior knowledge available. If our dialogue depends on knowing his or her name, age, gender, preferences, needs, etc., we first have to acquire this knowledge. This can happen explicitly by asking the user for this information. However, in many scenarios, this is not appropriate: When people are talking to strangers, there are several general aspects which they learn or at least assume without asking, such as gender and approximate age. With sensor technology, it is possible for the system to also obtain some information in a non-intrusive manner. This is even more critical for dynamic information (e.g., stress, emotional state, fatigue, etc.), since it would be neither accurate nor practicable to repeatedly ask the user for this information. More obvious even, many context parameters can also be determined by the system much more accurately. These may include the time of day, calendar events, brightness, temperature, other people and devices nearby, etc.

(2) involves user modeling, context modeling, and knowledge management. As mentioned earlier, EMF models can be created or extended to cover the application domain in question. The actual knowledge can then be stored in a knowledge base that is part of SiAM-dp. Alternatively, a custom knowledge base can be used.

The aspects which for the most part affect the core of the dialogue platform are contained in (3). This again consists of two sub-tasks: (a) determining which adaptation should occur and (b) performing the actual adaptation. Without step (a), the system could still be adaptable, but not adaptive (see [Endres et al. \[2010\]](#)). As shown in Figure 4.8, these points are not implemented into a single module, but rather affect multiple modules across the main platform component. However, there is a recurring scheme: The adaptation is described in terms of declarative adaptation strategies, which are assigned to one of several pools. Each module which is capable of adapting its behavior has such an associated adaptation strategy pool or database. In the current release of SiAM-dp, the dialogue management



**Figure 4.8** Illustration of the components affected by adaptivity.

is adaptive (e.g., it can change the user's path through the dialogue), as is the presentation component (e.g., adapting the timing or output modalities), plus several of the generators, which are also responsible e.g., for rendering an adapted GUI with different colors or layout. In each of these cases, the item being adapted is a presentation task. A presentation task is a concept used in SiAM-dp to describe a request to send some output to the user. Such a task passes through several components before it is being rendered, and it can be created in abstract shape with some parameters missing or containing alternative presentations from which one has to be chosen. As part of the adaptation, each adaptive component compares the presentation task with its pool of adaptation strategies to determine if any apply. The preconditions of each strategy are compared to the world state. For those resulting in a match, atomic effects are used to simulate the result of the strategy in the current situation, which is again matches with a goal function. In general, any strategy improving the evaluation of the goal function is applied. For this purpose, the strategy contains an implementation section, which is specific to the module which contains the strategy in its database. For example, strategies for the GUI renderer will contain a description of the GUI changes to be applied. Therefore,

the module triggering the adaptation is also responsible for implementing and performing the effect for the applicable strategies.

#### 4.8.5 Cognitive Resource Awareness

Today, dialogue systems are fully ubiquitous: They are used in all kinds of situations, including those where the dialogue is only a secondary task with a more critical primary task being performed at the same time. Examples are the use of built-in navigation systems while driving a car, checking messages on the phone while walking, or verifying the state of the production system while operating a machine in a factory environment. In each of these cases, the different tasks compete for the user's attention or cognitive resources. This condition is amplified if multiple functions are introduced as part of the dialogue, such as notifications or background information. Since our resources are obviously limited, and it is not always under the user's control to arbitrarily shift resources, it is vital in many situations to prioritize the resource demand according to the user's capacities in order to ensure safety and efficiency. Essentially, the dialogue system must be cognitive resource aware.

Cognitive resource awareness in SiAM-dp builds on the situation adaptivity features introduced earlier. Conceptually, it describes an adaptation that is performed when certain conditions occur in the user model. The condition is typically connected to the variable known as cognitive load, which describes the sum of all currently utilized (and therefore unavailable) cognitive processing resources. The adaptation could be performed on the dialogue level, e.g., by omitting some dialogue acts or by reducing overall user interruptions, or on the presentation level, e.g., by showing a visually less complex UI or by highlighting critical elements. A remaining question at this point is how to obtain the state of the user's cognitive resources.

Usually, the most relevant number for cognitive resource awareness in this process is the cognitive load. The more traditional way of obtaining this number is by trying to measure the effects on the user, such as stress or distraction. This can be done e.g., through sensors (heart rate/galvanic skin response sensors, brain interfaces, eye-trackers, etc.) or secondary task based metrics (reaction time, accuracy, etc.). While there are certainly several proven methods, which can also be integrated with SiAM-dp, the disadvantage of many of these metrics is that they are either impractical and expensive, or require a ground truth that is only available in simulated environments. For further discussion of indicators of cognitive load, see, e.g., Zhou et al. [2018].

Hence, as an alternative or complementary way of obtaining the cognitive load, Neßelrath and Feld [2013] presented the idea to base cognitive load computation mainly on a prediction of the effects of the known stimuli affecting the user. The most obvious stimuli are represented by the interaction of the user with the dialogue system. The suggested method essentially consists of concluding—from cognitive costs associated with a set of “primitive tasks” that are performed by the user or the system—the amount of cognitive load produced by interaction in the current situation. Examples for such primitive tasks are entering text, scrolling a list, giving a yes-no answer, or reading a message. It is known that cognitive demands and processing resources are modality-dependent, therefore the costs have to be quantified for tasks in each modality. For instance, a message has different resource impact when seen in text form on the screen than when read out to the user. This also means that for multimodal interaction, a single dialogue act may have costs associated with multiple modalities.

Cognitive load assessment in SiAM-dp is based on the Wickens model of attention management [Wickens et al. 1983], a cognitive cost matrix, and a set of filter functions for incorporating external stimuli and dealing with effects such as cross-modal influences, long-term memory, etc. The current initial implementation is based on the relationships that were experimentally obtained for the impact of visual complexity in a driving study.

#### **4.8.6 Dialogue System Offline Evaluation**

Cognitive load is one type of a user’s resource, but there are others: time can also be critical for implementing efficient dialogues, some systems might have monetary resources linked, and others could involve safety as a resource. These resources have in common that they are updated dynamically when the dialogue system is running. Measuring the consumption of these resources is an important step in the evaluation of the system, and this typically involves performing user studies under varied conditions.

With SiAM-dp, we propose a feature that can help simplify these situations. The presented architecture includes a toolset where dialogue systems can be modeled in a declarative way. Since the default dialogue model is based on state charts, this enables us to simulate the execution of the dialogue without an actual user (“offline”) or even without using the event infrastructure. In other words, the system is evaluated computationally, which is much faster and exhaustive than other simulations and considerably less cost-intense than a user study. Being able to compute resource consumption at any point in the dialogue is however just one aspect; it is at least as important to have a proper visualization for the result that

points the designer to the relevant numbers and issues. The visualization of SiAM-dp will be integrated with the toolset's graphical editor. States that match certain criteria (e.g., producing a cognitive load that is beyond a predefined threshold) can be highlighted, and each transition is annotated with the corresponding cost. This is especially helpful to locate resource bottlenecks. Finally, graph searches will allow the designer to quickly locate shortest/longest paths, or paths matching other criteria.

Offline Evaluation is not intended to replace user studies—numerous categories of effects related to usability or user experience cannot be measured based only on the dialogue model. Yet, it is a straightforward means of reducing the cost of user studies, by enabling the dialogue author to make strategic decisions or find and eliminate certain flaws earlier and without requiring additional conditions in the user study. With respect to cognitive resources, we believe that it will be easier to see whether some parts of the dialogue may be responsible for an increased and possibly excessive cognitive load.

## 4.9

### Current Trends in Dialogue Architectures

While the landscape of established dialogue platforms focusing on native support for multimodal interaction is manageable, the development in other areas of dialogue systems has made strong progress and is influencing next-generation multimodal dialogue platforms.

#### 4.9.1 Personal Digital Assistants

Over the last decade, many big companies in the consumer electronics market have started creating a personalized Virtual Digital Assistant (VDA), with *Apple Siri* introduced in 2011 being the first prominent example. Others have followed, including *Microsoft Cortana*, *Amazon Alexa*, or *Google Assistant*. All of these assistants have in common that they are speech-based. Enabled through advances in NLU technology, these new natural language interfaces have replaced the command-and-control interfaces that were available (although not widely used) until then. The strength of these platforms lies in the input processing step, which uses statistical language models for intent recognition and entity extraction. In addition to the aforementioned VDAs, the technology has matured and is now available separately from various vendors for creating custom NLU applications, such as *IBM Watson Conversations*, *Nuance Nina IQ Studio*, *Nuance Mix*, *Google Dialogflow* (originally created by *Speaktoit* but now owned by *Google*), *Microsoft LUIS*, *wit.ai*, and others. The core of each of these platforms is a module for the training of a language model from

a large number of sample sentences for each intent, with the optional annotation of entities (*IBM Watson* also allows the annotation of relations). Some of the frameworks do not work directly on speech—they rather convert the vocal input to text using traditional ASR and then apply the model to the text.

A feature that is shared by many of the VDA platforms is their extensibility: developers can extend the dialogue engine with custom commands (sometimes called *skills*) that enable new functionality. However, there is a strict separation between the “native” dialogue and the extension, and only a limited set of framework features is exposed by the dialogue engine, hence the integration of these extensions into the dialogue is usually less smooth and resembles command-and-control (e.g., “Alexa, start home delivery”).

By observing current developments, it becomes obvious that newer generations of devices will include further options for input and output aside from speech. For instance, the Amazon Show table is a logical upgrade of the Echo speaker with a touch screen for added output display. It supports a limited set of fission options for crafting output already today, although a much greater advancement will be made if cameras and other sensors will be able to recognize gestures and other types of input. The same is true for VDAs available on desktop computers and mobile phones.

### **4.9.2 Chatbots**

Chatbots are another variant of emerging dialogue applications. They are often encountered on web sites (e.g., as a buying advisor) or in messenger applications (e.g., Facebook Messenger, Skype), and they use typed text as their primary input modality. Nevertheless, they share the goal of having a natural conversation with the user.

There are three broad categories of bots: social bots, question answering (QA) bots, and task-oriented bots. Social (or “chit-chat”) bots are merely intended to entertain the user and usually have no well-defined goal. QA bots do not need to understand the task and often consist only of single-turn dialogue. Some bots attempt to accurately parse the user’s question into a semantic query, while others will just look up a “closest match” from a set of possible answer documents (e.g., an FAQ). Task-oriented bots, however, require a much more elaborate dialogue flow.

End-to-end learning techniques have been successfully applied for both types of dialogues. Several approaches have been proposed, such as using supervised learning [Wen et al. 2017], reinforcement learning using RNNs [Williams et al. 2017], deep reinforcement learning [Zhao and Eskenazi 2016], and interactive reinforcement learning.

Due to the advancements in dealing with image data using deep neural networks, it is to be expected that multimodal dialogues can greatly benefit from end-to-end learning. For example, interacting with an autonomous vehicle using a combination of speech, gestures, and eyegaze to the outside environment, an early fusion incorporating the raw image data could achieve more dynamic dialogue and respect subtle aspects of the interaction.

### 4.9.3 Probabilistic Architectures Based on POMDPs

A major drawback of finite-state charts and other traditional architectures is that they consider and follow only a single hypothesis. If that hypothesis is wrong, e.g., during a speech recognition error, it can lead to incorrect beliefs about the user's intention. As of today, dialogue systems based on partially observable Markov decision processes (POMDPs) have matured as an effective means to counter these shortcomings for speech dialogue systems [Williams and Young 2007]. By modeling the user's goals as belief states, the POMDP keeps track of the complete dialogue state history with parallel hypotheses, updating the belief scores in every step. Furthermore, all belief state updates are scaled by a priori likelihoods that can be configured in a user profile.

As Young et al. [2013] argue, the concept is fully applicable to multimodal dialogue, though it would imply a larger state space (see, e.g., Lucignano et al. [2013] for an application). When used in conjunction with end-to-end learning, a possible downside of POMDPs in this context is their reliance on large amount of handcrafted features [Bordes et al. 2017].

### Focus questions

**4.1.** What is a dialogue? Which of the following examples can be called a dialog?

- A: "What's the time?" B: "It's a quarter past ten"
- A: "Switch on the light, please." B: switches on the light
- A: "I'm feeling cold"
- A: "I'm feeling cold" B: closes the window

**4.2.** What is the difference between a dialogue system and a dialogue platform?

**4.3.** What are the three sequential processing phases of a typical dialogue system architecture?

**4.4.** What is the main task of an input processing stage? Which conceptual extension has to be made in the input processing stage to enable proactive dialogue?

- 4.5.** What is the difference between “user initiative,” “system initiative,” and “mixed initiative” systems?
- 4.6.** Can you come up with more examples for typical meta dialogue functions besides “I did not understand this?”
- 4.7.** What is the difference between multimodal fusion and multimodal fission?
- 4.8.** Describe the concept of early fusion. What would be an example for early fusion?
- 4.9.** Is fusion on a semantic level early or late fusion? What are the advantages of late fusion?
- 4.10.** What does the term “symmetric multimodality” mean?
- 4.11.** Which dialogue management architecture would you choose for a simple dialogue system with well structured tasks? Would you use the same architecture for a mixed initiative system?
- 4.12.** Describe the difference between “simultaneous integrators” and “sequential integrators.”
- 4.13.** What is a “relatum”? What are intrinsic, relative, and absolute frames of references?
- 4.14.** Explain the difference between anaphoric and exophoric references.
- 4.15.** Try an ellipsis on your favorite commercial VDA (e.g. Amazon Alexa, Apple Siri, Microsoft Cortana, Google Assistant . . . ). How many subsequent ellipses does the system understand (e.g. “How high is Mount Everest?,” “and Mount Fuji?,” “and the Eiffel tower”)?

## References

- J. Allen and M. Core, 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript. <https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/>. 161
- J. Allen, G. Ferguson, and A. Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, IUI '01, pp. 1–8. ACM. DOI: [10.1145/359784.359822](https://doi.org/10.1145/359784.359822). 156
- P. K. Atrey, M. Hossain, A. El Saddik, and M. S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6): 345–379. DOI: [10.1007/s00530-010-0182-0](https://doi.org/10.1007/s00530-010-0182-0). 160

- H. U. Block, R. Caspari, and S. Schachtl. 2004. Callable manuals - access to product documentation via voice (Anrufbare Bedienungsanleitungen - Zugang zu Produktdokumentation über Sprache). *Information Technology*, 46(6): 299–305. DOI: [10.1524/itit.46.6.299.54679](https://doi.org/10.1524/itit.46.6.299.54679). 166
- D. Bobbert and M. Wolska. 2007. Dialog OS: An extensible platform for teaching spoken dialogue systems. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, Trento*, pp. 159–160. 166
- D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. 1977. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2): 155–173. DOI: [10.1016/0004-3702\(77\)90018-2](https://doi.org/10.1016/0004-3702(77)90018-2). 156
- R. A. Bolt. 1980. “Put-that-there”: Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '80*, pp. 262–270. ACM, New York. DOI: [10.1145/800250.807503](https://doi.org/10.1145/800250.807503).. 157
- A. Bordes, Y.-L. Boureau, and J. Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the 5th International Conference on Learning Representations*. 182
- T. H. Bui. 2006. Multimodal dialogue management - state of the art. Technical Report TR-CTIT-06-01, Centre for Telematics and Information Technology, University of Twente, Enschede, The Netherlands. 154
- H. Bunt. 2000. Dialogue pragmatics and context specification. In *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, pp. 81–150. John Benjamins. 162
- H. Bunt. 2009. The DIT<sup>++</sup> taxonomy for functional dialogue markup. In D. Heylen, C. Pelachaud, R. Catizone, and D. Traum, editors, *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 13–24. 161
- H. Bunt. 2011a. Multifunctionality in dialogue. *Journal Computer Speech and Language*, 25(2): 222–245. DOI: [10.1016/j.csl.2010.04.006](https://doi.org/10.1016/j.csl.2010.04.006). 162
- H. Bunt. 2011b. The semantics of dialogue acts. In *Proceedings of the 9th International Conference on Computational Semantics, IWCS '11*, pp. 1–13. Association for Computational Linguistics, Stroudsburg, PA. <http://portal.acm.org/citation.cfm?id=2002670>. 161
- H. Bunt, M. Kipp, M. T. Maybury, and W. Wahlster. 2005. Fusion and coordination for multimodal interactive information presentation. In O. Stock and M. Zancanaro, editors, *Multimodal Intelligent Information Presentation*, vol. 27 of *Text, Speech and Language Technology*, pp. 325–339. Springer, Dordrecht, The Netherlands. 146
- J. Cassell. 2000. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 43(4): 70–78. DOI: [10.1145/33550.33562](https://doi.org/10.1145/33550.33562). 161
- S. Castronovo, A. Mahr, M. Pentcheva, and C. Müller. September 2010. Multimodal dialog in the car: Combining speech and turn-and-push dial to control comfort functions. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pp. 510–513. ISCA, Makuhari, Japan. 173, 174

- P. R. Cohen and S. Oviatt. 2017. Multimodal speech and pen interfaces. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3015783.3015795](https://doi.org/10.1145/3015783.3015795). 159
- P. R. Cohen and C. R. Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3): 177–212. DOI: [10.1016/S0364-0213\(79\)80006-3](https://doi.org/10.1016/S0364-0213(79)80006-3). 156
- P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In *Proceedings of the 5th ACM International Conference on Multimedia*, pp. 31–40. ACM. DOI: [10.1145/266180.266328](https://doi.org/10.1145/266180.266328). 160
- D. Costa and C. Duarte. 2009. Improving interaction with tv-based applications through adaptive multimodal fission. In K. Blashki and P. Isaias, editors, *Emerging Research and Trends in Interactivity and the Human-Computer Interface*, Ch. 3, pp. 54–73. IGI Global, Hershey, PA. DOI: [10.4018/978-1-4666-4623-0.ch003](https://doi.org/10.4018/978-1-4666-4623-0.ch003). 151
- J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. M. Young. 1995. Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The Care Properties. In *Proceedings of the INTER-ACT 95-IFIP TC13 Fifth International Conference on Human-Computer Interaction*, vol. 95, pp. 115–120. Springer US, Boston, MA. DOI: [10.1007/978-1-5041-2896-4\\_19](https://doi.org/10.1007/978-1-5041-2896-4_19). 157
- D. A. Dahl. November 2013. The W3C multimodal architecture and interfaces standard. *Journal on Multimodal User Interfaces*, 7(3): 171–182. DOI: [10.1007/s12193-013-0120-5](https://doi.org/10.1007/s12193-013-0120-5). 169
- G. Di Fabbrizio, J. Wilpon, and T. Okken. 2009. A speech mashup framework for multimodal mobile services. In *Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interfaces ICMI-MLMI*, pp. 71–78. Cambridge, MA. DOI: [10.1145/1647314.1647329](https://doi.org/10.1145/1647314.1647329). 164
- C. Endres. 2012a. PresTK: Situation-aware presentation of messages and infotainment content for drivers. Ph.D. thesis, Saarland University, Saarbrücken, Saarland, Germany. 153
- C. Endres. 2012b. Real-time assessment of driver cognitive load as a prerequisite for the situation-aware presentation toolkit PresTK. In *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2012*, pp. 76–79. 153
- C. Endres, T. Schwartz, M. Feld, and C. Müller. February 2010. Cinematic analysis of automotive personalization. In *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010*, pp. 1–6. ACM, Hong Kong, China. DOI: [10.1145/2002368.2002372](https://doi.org/10.1145/2002368.2002372). 176
- M. E. Foster. 2002. COMIC project deliverable: State of the art review: Multimodal fission. Available at <http://groups.inf.ed.ac.uk/comic/documents/deliverables/Del6-1.pdf>. Last accessed January 2019. 151

- A. Gruenstein, I. McGraw, and I. Badr. 2008. The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pp. 141–148. ACM, New York, NY. DOI: [10.1145/1452392.1452420](https://doi.org/10.1145/1452392.1452420). 165
- F. Honold, F. Schüssel, and M. Weber. 2012. Adaptive probabilistic fission for multimodal systems. In *Proceedings 2012 Conference of the Computer-Human Interaction Special Interest Group (CHSIG) of Australia on Computer-Human Interaction*, pp. 222–231. ACM. DOI: [10.1145/2414536.2414575](https://doi.org/10.1145/2414536.2414575). 152
- ISO 24617-2:2012. 2012. *Language Resource Management—Semantic Annotation Framework (SemAF)—Part 2: Dialogue Acts*. ISO, Geneva, Switzerland.
- A. Jaimes and N. Sebe. 2007. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2): 116–134. DOI: [10.1016/j.cviu.2006.10.019](https://doi.org/10.1016/j.cviu.2006.10.019). 159
- M. Johnston. 2019. Multimodal integration for Interactive Conversational systems. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA.
- M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. 2002. MATCH: An Architecture for Multimodal Dialogue Systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 376–383. Association for Computational Linguistics. 160
- M. Johnston, P. Baggio, D. C. Burnett, J. Carter, D. A. Dahl, G. McCobb, and D. Raggett, February 2009. EMMA: Extensible MultiModal Annotation Markup Language - W3C Recommendation. <http://www.w3.org/TR/emma/>. 157
- D. Jurafsky and J. H. Martin. 2009. *Speech and Language Processing, 2nd Edition*, Dialogue and Conversational Agents, pp. 863–891. Pearson, Upper Saddle River, New Jersey. 150, 154
- G.-J. M. Kruijff, J. D. Kelleher, and N. Hawes. 2006. Information fusion for visual reference resolution in dynamic situated dialogue. In E. André, L. Dybkjær, W. Minker, H. Neumann, and M. Weber, editors, *Perception and Interactive Technologies*, pp. 117–128. Springer, Berlin, Heidelberg. DOI: [10.1007/11768029\\_12](https://doi.org/10.1007/11768029_12). 163
- P. Lison. 2012. Probabilistic dialogue models with prior domain knowledge. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pp. 179–188. Association for Computational Linguistics. 156
- L. Lucignano, F. Cutugno, S. Rossi, and A. Finzi. 2013. A dialogue system for multimodal human-robot interaction. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 197–204. ACM. DOI: [10.1145/2522848.2522873](https://doi.org/10.1145/2522848.2522873). 182
- M. T. Maybury and W. Wahlster. 1998. Intelligent user interfaces: An introduction. In M. T. Maybury and W. Wahlster, editors, *Readings in Intelligent User Interfaces*, pp. 1–13. Morgan-Kaufmann, San Francisco, CA. DOI: [10.1145/291080.291081](https://doi.org/10.1145/291080.291081). 146, 147

- K. McKeown. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, New York, NY. 152
- M. M. Moniri, M. Feld, and C. Müller. June 2012. Personalized in-vehicle information systems: Building an application infrastructure for smart cars in smart spaces. In *Proceedings of the 8th International Conference on Intelligent Environments*, IE'12, pp. 379–382. IEEE, Guanajuato, Mexico. DOI: [10.1109/IE.2012.40](https://doi.org/10.1109/IE.2012.40). 163, 174
- J. D. Moore. 1995. *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*. MIT Press, Cambridge, MA. 152
- R. Neßelrath and M. Feld. 2013. Towards a cognitive load ready multimodal dialogue system for in-vehicle human-machine interaction. In *Adjunct Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2013*, pp. 49–52. Eindhoven. 179
- R. Neßelrath and M. Feld. July 2014. SiAM-dp: A platform for the model-based development of context-aware multimodal dialogue applications. In *Proceedings of the 10th International Conference on Intelligent Environments*. IEEE. DOI: [10.1109/IE.2014.31](https://doi.org/10.1109/IE.2014.31). 170
- R. Neßelrath and D. Porta. July 2011. Rapid development of multimodal dialogue applications with semantic models. In *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, KRPD-11. Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI-11. Barcelona, Spain. 168
- R. Neßelrath, M. M. Moniri, and M. Feld. 2016. Combining speech, gaze, and micro gestures for the multimodal control of in-car functions. In *Proceedings of the International Conference on Intelligent Environments*, IE-16. IEEE, London. 173, 174
- L. Nigay and J. Coutaz. 1993. A design space for multimodal systems: Concurrent processing and data fusion. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pp. 172–178. ACM, New York. DOI: [10.1145/169059.169143](https://doi.org/10.1145/169059.169143). 157, 159
- S. Oviatt. 1999a. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 576–583. ACM. DOI: [10.1145/302979.303163](https://doi.org/10.1145/302979.303163). 170
- S. Oviatt. 1999b. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11): 74–81. DOI: [10.1145/319382.319398](https://doi.org/10.1145/319382.319398). 159
- S. Oviatt. 2012. Multimodal interfaces. In J. A. Jacko and A. Sears, editors, *The Human Computer Interaction Handbook*, ch. 18, pp. 405–430. CRC Press, Boca Raton, FL. DOI: [10.1201/b11963-22](https://doi.org/10.1201/b11963-22). 159
- S. Oviatt, A. DeAngeli, and K. Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Referring Phenomena in a Multimedia Context and Their Computational Treatment*, ReferringPhenomena '97, pp. 1–13. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: [10.1145/258549.258821](https://doi.org/10.1145/258549.258821). 163

- S. L. Oviatt and P. R. Cohen. 2015a. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*, Ch. 9. Morgan & Claypool Publishers, San Rafael, CA. [170](#)
- S. L. Oviatt and P. R. Cohen. 2015b. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*, Ch. 7. Morgan & Claypool Publishers, San Rafael, CA. [160](#), [161](#)
- S. L. Oviatt, R. Lunsford, and R. Coulston. 2005. Individual differences in multimodal integration patterns: What are they and why do they exist? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pp. 241–249. ACM, New York. DOI: [10.1145/1054972.1055006](#). [159](#)
- C. R. Perrault and J. F. Allen. 1980. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3–4): 167–182. [156](#)
- V. Petukhova. 2011. Multidimensional dialogue modelling. Ph.D. thesis, Tilburg University. [162](#)
- V. Petukhova and H. Bunt. May 2012. The coding and annotation of multimodal dialogue acts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12. European Language Resources Association (ELRA), Istanbul. DOI: [10.1.1.466.2190](#). [162](#)
- N. Pfleger. 2007. Context-based multimodal interpretation: An integrated approach to multimodal fusion and discourse processing. Ph.D. thesis, Universität des Saarlandes. [162](#), [163](#)
- D. Porta, M. Deru, S. Bergweiler, G. Herzog, and P. Poller. 10 2014. *Building Multimodal Dialog User Interfaces in the Context of the Internet of Services*, pp. 145–162. Cognitive Technologies. Springer, Cham. [168](#)
- Z. Prasov and J. Y. Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 471–481. Association for Computational Linguistics. [163](#)
- B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörmller, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. 2009. Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12): 1760–1774. DOI: [10.1016/j.imavis.2009.02.013](#). [160](#)
- N. Schutte, J. Kelleher, and B. Mac Namee. 2010. Visual salience and reference resolution in situated dialogues: A corpus-based evaluation. In *Proceedings of the AAAI Symposium on Dialog with Robots*, pp. 109–114. Dublin Institute of Technology. [163](#)
- S. Schwärzler, A. Bannat, J. Gast, F. Wallhoff, M. Giuliani, M. Kasseecker, C. Mayer, M. Wimmer, C. Wendt, and S. Schmidt. 2008. MuDiS – a multimodal dialogue system for human-robot interaction. Technical report, Cotesys. [160](#)
- M. Serrano and L. Nigay. 2009. Temporal aspects of CARE-based multimodal fusion: From a fusion mechanism to composition components and WoZ components. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pp. 177–184. ACM, New York, NY. DOI: [10.1145/1647314.1647346](#). [157](#)

- D. Song. December 2006. Combining speech user interfaces of different applications. Ph.D. thesis, Ludwig-Maximilians-Universität München. [http://nbn-resolving.de/urn:nbn:de:bvb:19-62088\\_166](http://nbn-resolving.de/urn:nbn:de:bvb:19-62088_166)
- D. Sonntag, R. Neßelrath, G. Sonnenberg, and G. Herzog, December 2009. Supporting a rapid dialogue engineering process. Paper presented at the First International Workshop on Spoken Dialogue Systems Technology, IWSDS-2009, Kloster Irsee, Germany. DOI: [10.1007/978-3-642-32790-2\\_76](https://doi.org/10.1007/978-3-642-32790-2_76). **168**
- S. Tamura, K. Iwano, and S. Furui. 2004. Multimodal speech recognition using optical-flow analysis for lip images. In J.-F. Wang, S. Furui, and B.-H. Juang, editors, *Real World Speech Processing*, pp. 43–50. Springer US. DOI: [10.1007/978-1-4757-6363-8\\_4](https://doi.org/10.1007/978-1-4757-6363-8_4). **160**
- D. Traum and S. Larsson. 2003. The information state approach to dialogue management. In J. van Kuppevelt and R. W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, vol. 22 of *Text, Speech and Language Technology*, pp. 325–353. Springer, Dordrecht, The Netherlands. DOI: [10.1007/978-94-010-0019-2\\_15](https://doi.org/10.1007/978-94-010-0019-2_15). **150, 156**
- R. Tumuluri and R. Cohen, P. 2019. Commercialization of multimodal systems. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA.
- R. Tumuluri, D. Dahl, F. Paterno, and M. Zancanaro. 2019. Standardized representations and markup languages for MMI. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. **169**
- M. Turk and M. Kölsch. 2003. Perceptual interfaces. Technical report, University of California, Santa Barbara. [https://www.cs.ucsb.edu/research/tech\\_reports/reports/2003-33.pdf](https://www.cs.ucsb.edu/research/tech_reports/reports/2003-33.pdf). **159**
- W. Wahlster. 2003. SmartKom: Symmetric multimodality in an adaptive and reusable dialogue shell. In R. Krahl and D. Günther, editors, *Proceedings of the Human Computer Interaction Status Conference 2003*, pp. 47–62. DLR. **153, 160, 167**
- W. Wahlster, editor. 2006. *SmartKom: Foundations of Multimodal Dialogue Systems*. Cognitive Technologies. Springer, Berlin Heidelberg, Germany. **153, 167**
- W. Wahlster. 2014. Multiadaptive interfaces to cyber-physical environments. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pp. 1–2. ACM. Keynote. DOI: [10.1145/2557500.2568055](https://doi.org/10.1145/2557500.2568055). **148**
- W. Wahlster, N. Reithinger, and A. Blocher. 2001. SmartKom: Multimodal communication with a life-like character. In *Proceedings of the 7th European Conference on Speech Communication and Technology Eurospeech 2001*, vol. 3, pp. 1547–1550. **160**
- R. Wasinger. 2006. *Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations*. Aka Verlag, Heidelberg, Germany. **153, 160, 161**

- R. Wasinger, C. Kray, and C. Endres. 2003. Controlling multiple devices. In *Physical Interaction (PI03) – Workshop on Real World User Interfaces in Conjunction with MobileHCI '03*, pp. 60–63. [153](#)
- B. L. Webber. 1978. Description formation and discourse model synthesis. In *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*, pp. 42–50. Association for Computational Linguistics. DOI: [10.3115/980262.980270](https://doi.org/10.3115/980262.980270). [162](#)
- T. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. Rojas-Barahona, P. Su, S. Ultes, and S. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Proceedings of Conference*, vol. 1, pp. 438–449. DOI: [10.18653/v1/E17-1042](https://doi.org/10.18653/v1/E17-1042). [181](#)
- C. D. Wickens, D. L. Sandry, and M. Vidulich. 1983. Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 25(2): 227–248. DOI: [10.1177/001872088302500209](https://doi.org/10.1177/001872088302500209). [179](#)
- J. D. Williams and S. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2): 393–422. DOI: [10.1016/j.csl.2006.06.008](https://doi.org/10.1016/j.csl.2006.06.008). [182](#)
- J. D. Williams, K. Asadi, and G. Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 665–677. [181](#)
- B. Xiao, C. Girand, and S. L. Oviatt. 2002. Multimodal integration patterns in children. In J. H. L. Hansen and B. L. Pellom, editors, *Proceedings of the 7th International Conference on Spoken Language Processing*, ICSLP 2002, pp. 629–632. ISCA. [159](#)
- B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. Oviatt. 2003. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, ICMI '03, pp. 265–272. ACM, New York, NY. DOI: [10.1145/958432.958480](https://doi.org/10.1145/958432.958480). [159](#)
- S. Young, M. Gašić, B. Thomson, and J. D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, 101(5): 1160–1179. [182](#)
- T. Zhao and M. Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. DOI: [10.18653/v1/W16-3601](https://doi.org/10.18653/v1/W16-3601). [181](#)
- J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad. 2018. Multimodal Behavioural and Physiological Signals as Indicators of Cognitive Load. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108002](https://doi.org/10.1145/3107990.3108002). [178](#)

# Challenge Discussion: Advancing Multimodal Dialogue

James Allen, Elisabeth André, Philip R. Cohen,  
Dilek Hakkani-Tür, Ronald Kaplan, Oliver Lemon,  
David Traum

## 5.1

### Introduction

Arguably, dialogue management is the most underdeveloped technology in the field of multimodal interaction. The purpose of these challenge questions is to start to set directions for the next generation of conversational systems that can participate in a variety of dialogues with humans. The topics to be discussed are not merely of academic interest, but are crucial to commercial applications.

There are many dimensions along which human-computer conversations differ, including: Modalities of interaction, initiative, dialogue purpose, collaboration and dyadic vs. multi-party interaction. Please consult the [Glossary](#) for definitions of these terms. Given the huge variation in this multidimensional space, it is no surprise that the most frequently deployed commercial dialogue systems are very simple, including system-initiated Interactive Voice Response systems that take over a conversation and provide a limited set of options to the user, such as “Say ‘yes’, ‘no’, or ‘operator’”), and finite-state dialogue controllers that typically offer rigid dialogue flows. Other commercial systems are based on machine learning of dialogue responses from corpora of human-human dialogues. Typically, such conversations are voice-only interactions, but occasionally have graphical user interface output, such as the “chatbots” built on the Facebook Messenger™ platform. Overall, these commercial systems are fine as far as they go, but they don’t yet go far enough. Most

### Glossary

**Collaboration.** Dialogues often are expected to further participants' plans and goals, rather than just respond literally to the overt utterances. On the other hand, dialogues such as legal proceedings or negotiations can be adversarial or non-collaborative dialogues, in that some of the participants' goals are not shared, but the parties still obey collaborative conversational rules.

**Dialogue purpose.** Dialogue purposes can range from helping the user or system to perform a task, provide information, to social dialogues, in which the participants are making friendships, building rapport and trust, exploring their likes and dislikes, flirting, arguing, telling jokes, passing the time, etc.

**Dyadic vs. Multi-party dialogues.** In addition to the dyadic (two-party) case, a system may act as a full-fledged dialogue participant in a multi-way conversation, requiring it to process and respond to human-human dialogues in addition to responding to utterances addressed to it.

**Initiative.** Dialogue participants may: Take the initiative and prompt others for information, be totally reactive and respond only when requested, or have a mixed initiative strategy in which one party starts a conversation, and provides information about their intent after which a respondent may perform other dialogue acts, such as to request information from the first party before furthering the conversation or performing desired actions.

**Modalities of interaction.** Dialogues can take place in a richly coordinated multimodal context involving voice, gesture, vision, touch, etc., which is typical of human-robot or human-avatar interaction. At the other end of the spectrum, conversations can take place in unimodal settings, including so-called typed "chat" interactions, or can be supplemented with graphical user interface elements (e.g., buttons for selecting options).

of these commercial systems are unable to conduct dialogues that: accurately resolve pronouns and definite descriptions, depend on linguistic and/or situational context, process and respond to multimodal information, handle changes of mind, handle clarifications, express attitudes (irony, sarcasm, emotion, etc.), extend beyond a single question/answer pair, respond collaboratively to users' underlying plans and intentions, or provide their own responses rather than selecting among a prescribed set of responses, etc. These failings are widely recognized, and research is underway both to broaden systems' dialogue abilities and to deepen their understanding.

In the dialogue research community, the currently most popular mixed initiative dialogue systems are so-called "frame-based" or "slot-filling" systems that attempt

to acquire required information from the user to establish parameters for an action schema. The dialogue management strategies are either hand-crafted or machine learned from data provided by Wizard-of-Oz experiments and/or user simulations. Such systems engage in question–answer interactions, along with providing confirmations and performing a small number of other types of speech acts. Future systems will need a larger set of communicative and task-related capabilities and to be capable of dialogue in a broader range of settings and modalities. A complementary discussion of dialogue systems can be found in [Ward and DeVault 2016].

## 5.2

### Discussion Questions

The questions below are designed to stimulate discussion of how we can extend or develop new system capabilities to improve the processing of dialogue.

#### 5.2.1 Collaborative Dialogues

In collaborative dialogue, people are expected to respond in ways that further each other's goals/intentions/plans.

Imagine a Traveler landing in a different country (say, South Korea) for the first time who needs to connect by bus to her destination. The Traveler has approached an information provider (person or system) at 10:45 pm.

**Traveler:** *Do you know when is the next bus to Suwon?*

**Provider:** *Sorry, the last bus has left for the evening. You will have to take a bus to Seongnam and then transfer to the bus to Suwon. The bus leaves here from bay number 6 at 11:00 pm.*

Notice that the Provider does not answer the literal question. Rather, it has inferred the Traveler's higher level goal of being in Suwon that evening, and discovered an obstacle to the plan to achieve it. The Provider then informs the Traveler of the obstacle and volunteers a different option that would achieve her higher level goal. Thus, the Provider's goal may not necessarily be to search for the *optimal* response, but for an adequate one. However, there are truthful responses that would be *inadequate*. For instance, a collaborative information provider should *not* merely answer literally:

**Traveler:** *Do you know when is the next bus to Suwon?*

**Provider:** *Yes, or*

**Provider:** *The next bus leaves tomorrow at 5am.*

Such collaborative dialogues that address the conversants' plans are very common. Indeed, it has been argued that collaboration is the key to communication. Clearly, such collaboration capabilities are domain-independent, because inferring someone's plan, finding obstacles, and planning a response can apply to many task-related domains, and are independent of language.

- How can current approaches to building human-computer dialogue systems, including but not limited to, approaches based on machine learning (e.g., partially observable Markov decision procedures (POMDPs), reinforcement learning, neural networks), planning/plan-recognition, information state approaches, etc., be extended to handle these more collaborative exchanges in a domain-independent manner?
- How can dialogue systems be built to be robust to dialogue failures, such as failures of understanding, or even simply to changes in the world (e.g., the bus broke down)? Do we need different capabilities to handle dialogue vs. world failures?

### 5.2.2 Chatbot Dialogues

Many groups have been building so-called "chatbots" that mimic conversational engagement. Such systems learn their conversational responses from a vast collection of human-human dialogues, often using recurrent neural networks or other deep learning technologies. Corporate chatbots have been built, for example to route calls in a call-center to the right corporate department, or route users to the correct question in a "frequently asked question" database. Other chatbots have been built to enable a messaging user to engage a business for a service. Some simple examples exist, but as yet this class of dialogue systems requires a deeper understanding of conversation. They tend to imitate the superficial levels of dialogue without any understanding of dialogue semantics and pragmatics, or the user's emotional state.

- Do you believe these machine learning techniques will enable systems eventually to participate in dialogues for extended periods? If so, why; if not, why not?
- Dialogues can be both transactional and social. However, the dialogue field so far tends to use different technologies to deal with these dialogue purposes. How can one develop systems that seamlessly handle both transactional and social purposes?

### 5.2.3 Multimodal Dialogues

Future dialogue systems will need to process multimodal input and produce multimodal output. In the example above, the information Provider could have responded to the Traveler's multimodal cues:

**Provider:** *The next direct bus leaves at 5am. You can find a hotel 200 yards west of here with economy rooms.*

**Traveler:** [Frowns, shakes head and appears distressed]

**Provider:** *If you prefer to get there this evening, you will have to take a bus to Seongnam and then transfer to the bus to Suwon. The Seongnam bus leaves here from bay number 6 at 11:00 pm.*

Multimodal dialogue systems will need to process voice and other modalities, including recognizing meaningful information from users' faces and bodies, as in the above example. Likewise, robots and avatars will need to produce coordinated multimodal output.

- What will need to change in our approaches to handling dialogue in order to accommodate multimodal input and output? For example, how can we handle coordinated speech, gesture, eye gaze, head movement, and facial expressions? And how would the approach handle the temporal coordination between modalities?
- Where will the data for machine learning of the interpretation of coordinated modalities come from? What examples exist now? What type of data will be needed next, and how can we collect it?

People process and respond to utterances incrementally, as seen in their properly timed gestures (e.g., head nods or verbal backchannel utterances). If those expected backchannels signaling understanding are delayed, conversations break down. On the other hand, most current conversational systems are staccato turn-by-turn interactions.

- How can your favored approach to dialogue processing handle *incremental multimodal processing*?
- How can it handle spontaneous speech effects (dysfluencies, etc.)?

### 5.2.4 Hybrid Dialogue Reasoning

Many researchers have called for hybrid symbolic-statistical methods to be used to build dialogue engines.

- What hybrid methods would you advocate?
- What would they be used for?
- Where will we get the data?

### **5.2.5 Transfer Learning**

- What can a system carry over to a new domain from having learned to converse in another?
- Why does the system need to learn to converse all over again?

### **5.2.6 New Opportunities**

There are many directions that academic and commercial research into dialogue might explore, ranging from improved machine learning algorithms to new kinds of multimodal and/or multi-user interaction.

- What directions do you think are likely to be the most profitable for dialogue research in the near and medium term, both for academic research and commercial application?
- What research topics would you recommend for young researchers wanting to break into the field?

## **5.3 Conversation**

### **Participants**

**Prof. James Allen.** John H. Dessauer Professor of Computer Science, University of Rochester, and Sr. Researcher, Institute for Human and Machine Cognition.

**Prof. Elisabeth André.** Full Professor of Computer Science (W3), Chair of Human-Centered Multimedia, Faculty of Applied Informatics, Augsburg University, Germany.

**Prof. Phil Cohen.** (moderator) Professor of Artificial Intelligence and Director Laboratory for Dialogue Research, Faculty of Information Technology, Monash University.

**Dr. Dilek Hakkani-Tür.** Researcher, Conversation Modeling Technical Lead and Research Manager, Google Deep Dialogue Research Group.

**Dr. Ronald Kaplan.** Adjunct Professor of Linguistics, Stanford University.

**Prof. Oliver Lemon.** Director of the Interaction Laboratory, Heriot-Watt University, UK.

**Dr. David Traum.** Head of Dialogue Research Group, Institute for Creative Technologies, Univ. of Southern California. Dr. Traum provided comments after the group discussion.

---

**Question: How can current approaches to building human-computer dialogue systems be extended to handle collaborative dialogues in a domain-independent manner?**

**Phil Cohen:** The first topic that I want to raise has to do with collaborative dialogue. Generally, for the kind of systems we are trying to build, we want the system to be collaborating with us. The example I gave you is one in which a person asks a literal do-you-know question: “Do you know when is the next bus to Suwon?” The information provider doesn’t actually answer the literal question, but says “Sorry the last bus has already left for the evening you’ll have to take a bus to Seongnam, and then transfer to the bus to Suwon. And that bus leaves from bay number 6 at 11.” In other words, the person has addressed the traveler’s goals, but has not actually answered the literal question. In fact, answers given only to the literal questions would be uncooperative: “Do you know?,” “Yes I do know,” or “The next bus leaves at 5:00 in the morning.” But that’s not what they wanted to know. They wanted to know how to get there this evening.

So given all the techniques that you all have worked with over the years, how do we go about getting systems to handle dialogues like this? The machine learning approaches may have trouble finding enough examples, although I submit the examples are everywhere. Regarding the planning and plan recognition approaches, well this sort of dialogue came from that general mode of thinking, but I would like to know how the other approaches that we might be thinking about can address these kinds of topics.

As the moderator I’ll always take the opportunity to call on somebody. Oliver?

**Oliver Lemon:** I don’t mind saying a few words about this because I spoke about it last year at SigDial, though for me it’s like going back in a time warp to when I worked on a robot helicopter project and we were doing plan-based dialogues [Lemon et al. 2002]. First, you will have two problems to solve—you have to infer what the user’s plan is, and then you have to have something like a recipe for that plan and activate that recipe and see if preconditions are met or whether there’s anything making the plan fail. So I know at least several people have worked on

this in the past, but then the field kind of got dominated by slot-filling dialogues for a while where everyone got obsessed with just getting slots filled correctly when speech recognition is a problem. And I think we actually need to go back to working on these much more plan-based systems. It's a very interesting question how you could possibly do that and use statistical models. So what we're doing right now is trying to come up with some sort of hybrid systems where you do use some statistical components, but underneath you need to have plan representations. I'm not sure you escape from having to actually represent things like plans, recipes, constraints, and these kind of things. So I'm a bit skeptical that a fully statistical approach can actually deal with these things. I'll be interested to hear what other people think.

**Phil Cohen:** Dilek, what do you think?

**Dilek Hakkani-Tür:** I think that to be able to really understand and respond to these types of questions, the system needs more information than just the word sequence because the user could as well mean the next bus, they may just want to know when the next bus is. But you prepared us with some context and some of this context is actually available and is already in use by the smart phones for example. So I think we do know what the context is, and if we have ways of integrating it, for example via learning if we have data, then we could probably give a better answer. Unfortunately with spoken dialogue, it may be hard to present multiple available buses. With visual interfaces, something like this is really easy.

**Phil Cohen:** That's interesting, why do you think it's easy with a visual interface?

**Dilek Hakkani-Tür:** For example, every time I search for a flight from the Bay Area to Seattle, I am actually given a list of options that would ask me do you want non-stop flights, what time frame do you want, etc. on the web form. There's just so much information in the visual interface, even though I give them the minimum amount of information: just the origin and destination.

**Phil Cohen:** Yes, it's surely the case that when we are having a spoken dialogue, we don't have access to all that kind of information.

**Dilek Hakkani-Tür:** Exactly.

**Phil Cohen:** But I guess the question is where will we get the data from that will enable us to train a system to do this kind of reasoning. There's a whole collection of people in the dialogue world right now who think they can avoid explicit representations, and we can simply train end-to-end, and we'll get to that in the next set of questions. But in this case there are a potentially significant number of plans that

people are judging and figuring out which is the most likely, and then using that to guide their answers.

**Oliver Lemon:** To do this in an end-to-end style you'd have to condition the output sequence on all the relevant context features that you can get hold of and I'm not sure you can actually find big enough data sets that allow you to train that. Then this question of even if you did that for one domain, you manage to collect some huge very rich data set, it wouldn't necessarily transfer very well to another domain. So it's for those reasons I think it's actually still sensible to look at explicit representations of plans and activities and goals and things like that.

**Ron Kaplan:** I agree.

**Dilek Hakkani-Tür:** I agree with the multi-domain but I think in certain domains, and especially given the richness of visual interfaces, getting some data could be possible. I mean I'm not saying that it's a solved problem, but I think that there could be some creative solutions especially around visual interfaces. That could help us get some idea of what this user may be doing.

**Phil Cohen:** Ron, you had a comment?

**Ron Kaplan:** I wanted to say that I agree with Oliver that the key to this is to have some representation of the plan. And that's an abstraction away from the domain. The operations on plans include revision, replanning, backup, and so forth. It's not just the selection among a set of available plans you might think of, but it's actually operations on the abstraction itself. And the thing that's really important is that the abstraction is abstracted away from the particular properties or the particular elements of a particular domain; those are more general than that. What happens when someone's frustrated or something can't be achieved, or in the middle of a collaborative dialogue information that the system provides back to the user causes the user to change their intention?

**Phil Cohen:** Right.

**Ron Kaplan:** Which changes what they're trying to do. So that's where this kind of shared plan and shared collaborate dialogue stuff comes in. So I think that to the extent that the data-driven approaches are driven by properties or features of a particular domain and don't have the ability to abstract away from that, they're not going to be very effective in the long run. I also think, by the way, that a lot of attention goes into predicting the next best move by a lot of these statistical systems. I actually don't think that's a particularly interesting question. What's more interesting is to predict what's the really worst thing that you could do next; what's the next bad move? I think there are a lot of good next moves. If you're

doing a movie application, you know I want to watch a movie. There are lots of next questions that the assistant could ask, like “what actor,” “what director,” or “what location.” But if the next question was “What do you think the running length should be?,” well that’s bizarre. A lot of the focus goes into optimizing using the data for the next best move, as opposed to making sure that you filter out the bad ones.

**James Allen:** I obviously agree with a lot of what people have been saying and just wanted to throw out that I think it is absolutely essential to have an abstract level of collaborative problem-solving that is independent of the actual plans or tasks performed in the application domain. I believe it is quite possible. In fact, we’re building a number of systems this way now, where you have a domain-independent collaborative problem-solving shell interacting with a domain dependent reasoning system that might do arbitrary types of computation [Galescu et al. 2018]. It might use statistical intention recognition or plan recognition, or it might just be an agent-based system that reacts in various ways. The critical thing is that there is a well-defined interface between the domain-independent part and the domain-specific part. And that is, I think, what people have been talking about: An abstract level of operations on plans where you can do things like propose goals, you can refine a plan, you can change something, or you can modify goals. You can do all those kind of operations and those can all be expressed independent of the actual task, and independently of how the task reasoning is actually implemented.

---

**Question: How can the technology underlying chatbot dialogues be extended to handle transactional dialogues?**

**Phil Cohen:** OK, I think we have a lot of questions to go through. We could talk about all of these questions for a long time, so let’s move on to the next topic, which is Chatbots. Now of course many people mean very many different things by the term “chatbot,” but essentially what I am referring to here are end-to-end trained systems that essentially simulate dialogues but don’t actually participate in them.

It’s too bad David Traum isn’t here because he’s built one of these and it’s been extraordinarily effective. I don’t know if you all have seen his Holocaust survivors’ stories in which he trained a chatbot, but the chatbot was actually represented by a person who was interviewed [Artstein et al. 2016]. Essentially they were classifying the users’ queries into remarks that he [the survivor] had already made. And it had a beautiful graphical user interface, namely a 3D rendition of the person, and it was extraordinarily convincing. People even were crying and apologizing to the Holo-

caust survivor about what happened to him. But it goes back to Joe Weizenbaum's Eliza, and people were believing that too. So what will these techniques ultimately be useful for? I think we agree on their limitations, but you know there's something very clever going on here, and it would be interesting to speculate what we could actually do with this kind of training in a hybrid type of scenario in which you may be combining it with other types of information.

**Elisabeth André:** Chatbot systems are quite useful for role play. We developed virtual agents with chatbot functionality for job interview training. The basic idea is to employ virtual agents as job interviewers that interact with users while their social cues are recorded. The virtual characters do not understand much about the content of the dialogue. But this does not matter. If the users are actually willing to train their social skills, the approach usually works. But of course if the users try to break the system—this can be done very easily—they would not benefit from the training. However, if the users are cooperative and willing to engage in role play with the virtual characters, then the approach is very effective, more effective than, for example, learning how to behave in a job interview from a book [Damian et al. 2015] I believe that role play is quite a good application for chatbots.

**Phil Cohen:** That's very interesting Elisabeth. It sounds as though you need a willing suspension of disbelief and you realize you're dealing with a fictitious entity, and then if you're willing to cooperate with it, then the system can actually be fairly useful.

**Oliver Lemon:** We've been competing in this Amazon Alexa challenge last year, and we're also in it this year, and it's not the Turing test; we're not trying to fool people into thinking they're talking with a human. So they know they're talking with an audio chatbot, but conversations can go really well and can actually go on for quite a long time, and be very engaging. And people actually can really enjoy it and get into it and that's something that surprised me because we build a lot of task-based systems for which the task is supposed to be done as quickly as possible and then you're done. This is the opposite kind of thing—you're actually trying to be engaging and entertaining. And people go along with it to a large extent. We have 10% of our conversations being over 10 min long. We get a lot of people returning to call the system and talk to it again and again. It really surprised me that even though the system doesn't understand what it's talking about on some level, or has a very superficial understanding of what the conversation is actually about, there's a level at which that willing suspension of disbelief is enjoyable for people, and they can still get a lot out of it, and they can even learn a lot. For example, our chatbot [Papaioannou et al. 2017] has indexed all of Wikipedia and it updates news items

every night. You can actually learn a lot from chatting with it, and it's enjoyable even though it doesn't have a level of deep understanding.

**Ron Kaplan:** Is this like a cocktail party scenario? It's the experience itself, and there's no particular topic? Find out about somebody and talk to them?

**Oliver Lemon:** It's open domain conversation and we have a persona that tries to find out about you and what kind of movies you like. We do this kind of thing where we try and find out, yeah the idea is that you're meeting someone for the first time. You try and find out what they're interested in and this kind of stuff.

**Ron Kaplan:** Or what do we have in common?

**Oliver Lemon:** Yeah.

**Ron Kaplan:** It's a different purpose even in the role-playing training example that Elisabeth mentioned,

**Oliver Lemon:** Yes, so this is just completely open-domain dialogue that is supposed to be engaging and entertaining.

**Dilek Hakkani-Tür:** What type of persona does it have, and does the persona have any impact on how people would perceive any errors that it is making? One example I want to give you is Microsoft's Xioice has so many active users but the persona is such that if it makes a mistake or if it says something silly, my understanding is people are more willing to be OK about it. Did you see anything like that?

**Oliver Lemon:** Yeah; we tend to apologize if the user says "oh, that's dumb" or "you've gone off topic" or something. We have a persona that apologizes for doing those things, and then proposes to do something else. That works relatively well. We get all kinds of very interesting and strange kinds of abusive behavior as well. People kind of indulge in it. It's very interesting because you're doing something which is just for the purpose of engaging people but I kind of believe that it's also important for task-based dialogue systems because what I think we need to do is have task-based systems [which] are embedded within this wider kind of social chat, so that people will actually begin to enjoy using these systems. I do feel it's a problem that you know you can make a pretty good task-based system but it's not really that much fun to use it. So I think there's a very interesting future area and trying to combine these.

**Elisabeth André:** I completely agree. In this context, I would like to mention Tim Bickmore's and Justine Cassel's early work on relational agents [[Bickmore and Cassell 2001](#)]. They investigated the use of small talk in a real-estate agent. Such an agent should not only show a great deal of knowledge about housing, but at

the same time needs to build up trust with the user. To this end, it uses a mixture of task-based and social dialogue. Often it is more important to create a trustful relationship with users than to fulfill a particular task. If the user regards the dialogue agent as a companion, simple chatbots might be more effective than natural language systems that go into depth.

**Phil Cohen:** But if you're going to build trust, you have to say true things and meet your commitments, and the chat systems don't know if they're saying something that's true or if they're meeting their commitments.

**Elisabeth André:** I am not sure whether chat systems always need to say things that are 100% correct. Social lies are an essential component of everyday conversations. For example, people often make deceptive remarks to flatter their friends. Such lies help build up trust and are socially accepted. It is, of course, another thing when a chat system is supposed to provide health recommendations to the user. In this case, it really matters that a chatbot is telling the truth in order not to harm the user.

**Ron Kaplan:** Well, even in the Wikipedia example that Oliver mentioned, let's assume that the Wikipedia is true, but the responses have to be relevant to whatever the topic is. So the trust depends also on the view that the system properly interprets what it is that the user is talking about and wants to know about rather than say something that's true but completely off the wall.

**Phil Cohen:** Yes, it's not just any true thing.

**Ron Kaplan:** Yes, you're getting back almost to this task-based kind of thing where the task is to be having a conversation about something that is meaningful with respect to the [topic of] conversation.

**Phil Cohen:** One aspect of trust is explanation. In principle, I can ask a plan-based system: "Why did you say that?"

**Ron Kaplan:** Right.

**Phil Cohen:** Because it has as a plan standing behind the utterances, but for most of these systems, be they chatbot systems or just slot-filling systems, if you ask "why did you say that?", you're gonna get catatonic responses.

**Oliver Lemon:** Yeah that's true, and that's one of the big places where these things fall over. So I mean we retrieve relevant information. If someone asks for a named entity we extract that and we go and look for it on Wikipedia or news stories and we retrieve something that's relevant to that named entity that is coherent and relevant

and all these conversational things. But then as soon as somebody asks a follow-up question that digs into that content that you've just delivered, these systems immediately fall over. There's a really big challenge there in maintaining coherence over multiple turns, especially with these kind of follow-up questions; things like trust start to fall away.

**Ron Kaplan:** So the system we built at Nuance a couple of years ago [Breen et al. 2014] did have an explanation capability and you could say well "why do you say that?" It's particularly important when the system does something that is implausible, or may seem to be bizarre. If the user can say "well why did you say that?" and there's a plausible explanation, that's OK. In this case, it was about being able to articulate the path of links through Freebase that led from one part to another of the connected things. So if the system can explain bizarre behavior and the user can say "it's not the way I would have thought about it but I could see how you would have thought about it," that's again part of building trust. You can be bizarre as long as there's kind of an interpretable rationale that you can explain, that you can present.

**Oliver Lemon:** That's another reason why you need some of these explicit representations underneath like ontologies or something you can navigate through.

**Ron Kaplan:** Yes, so it was based on Freebase [Ramachandran et al. 2015]. So you could say "I want to watch a movie about lawyers." Who would ever say that? But it said "Legally Blonde" and that's obvious. It also would bring up "Batman Returns" or the "Dark Knight." "Why did you say that one?" It would say well because one of the main characters is the district attorney. Well, OK, you're not a total idiot.

**Oliver Lemon:** We're trying to add that sort of thing into our Alexa system that's pretty powerful I think.

**David Traum:** (after the discussion) I think the panel did a good job of discussing the contrast between "chatbots" that are just looking to predict the next utterance (and use that to try to continue coherently), and dialogue systems that need to develop some shared understanding of what is under discussion. The former can make do, to some degree, with statistical methods based on large sets of in-domain training data, but the latter really require some kinds of representations (plans, ontologies, etc.) and reasoning to be able to function acceptably.

I want to point out that there is a wide range of systems that fall in the middle—and in fact, most of the dozens of systems we've built in my almost 20 years at ICT have fallen in this middle area. For these kinds of applications we need both human-like dialogue and some amount of common ground about the topic under

discussion, but generally not as much as is needed for assistant-systems where the goal is only to efficiently perform a service for the user. Elisabeth mentioned role-play dialogue, often used for education or practicing social and communication skills. For these kinds of systems, what's most important is not that the system correctly and concisely recognizes and implements the user's intent, but that the user spends a significant amount of time and cognitive effort on the task, and is engaged in constructive thinking about the task—some misunderstanding can actually be beneficial here, as it leads the user to try to think about what may be going wrong in their communication and whether the issue might be a different model of the domain. The systems need to be human-like enough to cause this engagement with the task, and not have conversational breakdown when the system character is oblivious rather than responsive to what the person is saying (including meta-dialogue about communication issues).

Some of our systems, like MRE [Traum et al. 2003] and SASO [Traum et al. 2008] have a lot of domain-independent knowledge and fairly developed models of dialogue information state, plan representations, emotion models, negotiation models, etc. and some of them, like the New Dimensions in Testimony Holocaust survivor systems are more focused on the surface interaction, but all of them are “hybrid” in that they use some data-driven aspects of understanding and content selection/generation but also some representations of the dialogue state and when to just use the top “predictive” response vs. something else.

This includes some notion of social state and relationships between interlocutors, as Elisabeth mentions, and how to talk to individuals in different relationships, even when the system character's overall goals are the same.

---

**Question: Multimodal dialogues: what will need to change in the approaches to handling dialogue in order to accommodate multimodal input and output? For machine learning approaches, where will we get the data?**

**Phil Cohen:** So I want to move to another topic: multimodal dialogues. Now multimodality will appear in many different places particularly if you're dealing with robots or avatars or even GUIs. But if we're dealing with continuous processing of speech and other continuous modalities, we have an issue with current ways that we're doing speech recognition and natural language processing. It's a very staccato process—one utterance, one response; one utterance, one response. There's no co-timed modality understanding of something as seemingly simple as pointing, but there are obviously gestures and eye-gaze and prosody, head movement, and facial expressions. All these modalities combine to change the meaning, or fuse

their meanings with that of the utterance. So given the approaches we see today, for instance for natural language processing, maybe it's bi-directional recurrent neural networks for linguistic processing, where you have to get to the end of the utterance before you're going to do very much, how is the current state of statistical language processing, for instance, going to start handling the multimodality and where would we get the data to train the [recognizer(s)] for all of this multimodal input that otherwise would be difficult to capture?

**Elisabeth André:** There is an excellent survey on multimodal machine learning by Baltrušaitis, Ahuja and Morency in this handbook [[Baltrušaitis et al. 2018](#)] that discusses how to capture the dependencies between multiple modalities and how they are temporally aligned to each other. However, multimodal machine learning approaches usually do not aim at a deeper semantic analysis of the single components. There is a limited understanding of the meaning of a multimodal utterance. Earlier, you gave an example of a traveler that frowns, shakes her head, and appears distressed. The analysis of nonverbal cues is no longer a big deal for statistical machine learning approaches. However, these approaches alone are not able to figure out why somebody is distressed.

**Phil Cohen:** Well it ties back into the plan recognition in this particular example. But, yes, I haven't seen that in any of the multimodal scenarios that we've come across in the research literature.

**Elisabeth André:** Yeah.

**Phil Cohen:** Typically the work in the multimodal literature has used fairly standard grammatical processing, and has not been using the most recent recurrent neural nets or what-have-you, but once you start adding in those, and you feel like you have to train everything, you won't have very much in the way of data to do this with. It would be hard to crowdsource this [multimodal input].

**Oliver Lemon:** I agree in [the difficulty of] multimodal data collection in order to do this at the semantic level because from what I've seen, that's what's missing. It's tying in at the level of meaning in context. For example, facial expressions or even if somebody pauses and they don't say anything—silence can be incredibly meaningful in a particular context. And what you need in order to capture that in an RNN or whatever is actually lots and lots of data where silence in different contexts has different meanings. And that's just for silence, right? So again it's not that these statistical models can't in principle capture this kind of thing, it's just that the amount of data I think you would need to do it just seems impractical. That's a worry that I have anyway.

**Elisabeth André:** I would like to give another example from the EmpaT project [Damian et al. 2015]. In the job interviews with the virtual agent, people were smiling a lot. But they were also smiling when the system was making a negative comment on their performance. But in this case they were of course not happy. Rather, they were embarrassed, or they were just polite. A smile can be very ambiguous. But the systems nowadays would always map a smile onto a positive emotional state. Thus, the connection between nonverbal cues and their causes is completely missing. To draw the right conclusions from a smile, an immense amount of data would be required.

**Ron Kaplan:** These examples are about the interpretation of multimodal communication signals. I think that's one really important aspect of this multiple interaction. But there's also the notion of a shared environment the system and the user are operating in, and it may be an environment that's partly created by the system like I think Dilek was mentioning—visual output, like putting up a display if you had a display, of the bus schedule. And now you have an artifact that both the system and the user are aware of. Now the conversation can be about that. You can have demonstratives, you know the third line, the fourth line, and so forth, and that's another kind of aspect of multimodal, which is talking about things in the common environment.

**Elisabeth André:** To establish common ground in a dialogue, people make use of multiple modalities that have to be tightly coordinated. Language plays an important role, but also pointing and gaze behaviors. It is quite a challenge to implement dialogue systems that simulate the complex interplay of the highly interwoven interpersonal coordination and grounding processes.

**Dilek Hakkani-Tür:** We did some collections related to pointing and gaze, mostly using Wizard of Oz, but that said, you would need to collect these signals synchronously to speech and it's not easy, I mean just the engineering of such infrastructures is not easy, but once you get over that the data is very useful. Unfortunately our approach is not scalable to crowds currently. With such data, I think it's quite interesting to understand what is really happening in there.

**Phil Cohen:** Would some of those techniques work if, for instance, you had sequential performance of the various modes? So we've often seen people point and then stop and then speak about what they pointed at.

**Dilek Hakkani-Tür:** Sure, I mean if you can align the input language and gaze, then you should be able to process it. But the collection of the amount of data that you need and the way to get there I think is quite challenging.

**Phil Cohen:** James do you have any comments?

**James Allen:** I don't think I have much to add given what people have been saying. It's exceptionally challenging to figure out how you might collect the data for such approaches.

**Phil Cohen:** So the multimodal shades into the incremental. If you're going to be processing, for instance backchannels, or performing actions before the person has even finished uttering a noun phrase, such as "there is a little yellow" and "I picked up the yellow thing" and "I didn't have to even hear the rest of the utterance." So incremental processing of utterances in the currently fashionable way of doing things may be very challenging. It's not obvious how you do this in the various RNN flavors, for example.

**James Allen:** That is a very good point, Phil, in the work we've done we often have a GUI display that represent the state of the world, the state of the task, or other things like that. This is absolutely essential for maintaining an effective common ground between the user and the system, which is very hard to accomplish by just using language. So certainly in these more complex domains or in physical domains when you're actually interacting with the world, having that ability to use that as for the grounding purposes is critical. I think a lot of the systems we've built would not work without having a good graphical user interface that is summarizing what's going on right now in the task.

**Phil Cohen:** But I may want to give my robot instructions on what to do at some future time. It can't do it now; it's going to only do it when it gets to the bus station, etc., and so it's hard to ground it physically, but you're still going to get back channels "uh huh," or you should, "yeah I understand, I turn left and I turn right." So I'm going to ground my understanding. I may not ground it in the world though because the noun phrases don't refer right now, they'll refer later.

**Ron Kaplan:** That's sort of setting up a monitor on the future of the system. Another example is transfer of \$1000 from my checking account to my savings account when my salary comes in. Now you have the system having to be able to set up a predicate on reality that gets evaluated continuously or in a polling sort of way so that it can fire in the future. Again, there's a sparseness of data if you believe that the only way you're going to construct these systems is by training, which is kind of an academic perspective on this kind of thing. We're going to try to prove how much we can do with as many hands tied behind our back as possible. But you know, why bother learning things that you already know or that somebody can tell you? So back to the I guess the next question, which is the hybrid [architectures].

---

**Question:** Transfer learning: what can a system carry over from having learned to converse in one domain to another? can dialogue management be domain-independent?

**Phil Cohen:** Right. Also transfer learning. Why do I have to learn to converse all over again in another domain?

**Ron Kaplan:** Exactly, so the machine learning can be very effective but also very silly if what you're going to do is need 100,000 utterances labeled to know that English has subject verb agreement. You know you can ask any taxi driver they'll tell you. It's not a mystery. So if you look at the way the field is evolving right now, it is a very purist kind of attitude towards machine learning, and not recognition, I think James pointed out before, about the important thing of merging the statistical components that represents things that nobody can articulate, nobody can actually tell you about, with things where generalizations have already been discovered and represented that are known.

**Phil Cohen:** So that brings up the question of hybrid reasoning methods. What [hybrid] methods have you come across and are investigating that you think would be fruitful for students or future researchers to investigate? What would you use them for, and of course the next question is, where do we get the data? But let's ask what methods are being explored right now?

**Oliver Lemon:** So I would like to follow up on what Ron said. I think it's interesting too, and it's a hybrid method, to do things like machine learning, but some of the features that you were computing over are actually provided by a symbolic system. One example of that is an incremental semantic parser, that incrementally produces logical forms representing meanings. Then we do reinforcement learning over states that are created from those symbolic representations. So what I totally agree you should be using, linguists have come up with great grammars and semantic parsing systems, for example, that encode a lot of prior knowledge [so] that we don't actually need to spend a lot of time training to learn those things because we already know them. But then you can use machine learning models where appropriate, for example sometimes in some kinds of optimization. So I guess it's a bit of a dark art at the moment, of trying to figure out where that boundary lies and how to cleverly combine these kind of systems.

**Phil Cohen:** Dilek, what do you think?

**Dilek Hakkani-Tür:** I agree with the dark art part; beyond that I don't really have much to add.

**Ron Kaplan:** One dimension, I don't know, if you think of the symbolic system producing a framework or a skeleton of alternatives derived by abstractions, derived by generalizations from a knowledge base, and that's kind of an input to the system where the machine learning is now responsible for selection or disambiguation. So all these things are possible, but which one is optimal? It may well be that nobody can articulate exactly how you make that selection among the candidates, and that's something that the machine learning system with maybe relatively little data can deal with, much less than the data that it takes to learn how to generate the candidates.

**Oliver Lemon:** Yes, sometimes it's just because the feature space for consideration is just so large that no human can get their head around it, at least it's not in an explicit way that they understand [or] can articulate to anybody else.

**Ron Kaplan:** Right.

**Oliver Lemon:** Machine learning's really good if you use these kind of word embeddings for similarity judgments, and that kind of stuff. Those can capture, at least to some extent, human intuitions, that are difficult to explain in detail. That's just one example of where those kinds of models can be really useful.

**Ron Kaplan:** Even there, the interaction of that kind of embedding space with certain logical and known things like ontologies or negation or whatever such as similarity or other relationships. I mean an example from the Amazon context, you're looking for "snacks without nuts."

[laughter]

**Ron Kaplan:** First of all, I always use the with/without example to point out that normal search techniques think of 'with' and 'without' as basically being the equivalent of stop words. So you throw them out. But you also have to know what a nut is, and it's known what a nut is. Peanuts are not nuts, and almonds are, and people that are allergic to tree nuts are not allergic to peanuts and vice versa. So now you have to know something about the statistical space. We have to combine that with these ontological relations, these category relations that are known. It would be silly, and probably inaccurate, to try to learn them because peanuts and walnuts are probably used in a lot of the same contexts. A lot of people don't care about the difference. But if your life depended on it, you would care.

**Oliver Lemon:** Yeah, the notion of similarity is incredibly context dependent as well, and I think it's very interesting [that people] can on-the-fly create the right kind of similarity space for this particular topic that we're talking about. Nobody I know has done know that. That seems a great problem to work on in the future.

**Ron Kaplan:** So more generally I've always talked about two kinds of learning. I learned long ago that you don't say that you're doing learning. So there's learning by observation which is all this data-driven work. Then there's learning by instruction, somebody who knows tells you, and the problem there is they have to tell you in a notation that you can compute with and can combine with the other stuff.

**Phil Cohen:** Sounds like McCarthy's Advice Taker.

**Ron Kaplan:** Which is an important notion.

**Phil Cohen:** Well most of the systems that we see, you can't tell them anything. They have to learn it.

**Ron Kaplan:** Well, that's right.

**Phil Cohen:** They don't take input lightly, or at all. So let's shade this conversation then into what can a system carry forward having learned how to engage in a slot-filling dialogue in one domain. Is there not a level of generality that all you have to do is parameterize the next type of action or frame, so to speak, and it will have those same capabilities but in a different topic? Do we need to learn it or do we just need to parameterize?

**Oliver Lemon:** We've done a bit of work, for example, learning repair and hesitation and [phenomena] like that. You know those seem to be general across all different types of dialogues. So I think there's a very interesting level of generality that you can capture also with statistical models that removing hesitations and computing repairs properly. Those kind of processes I think can be domain-general but a lot of it really is domain-specific.

**Elisabeth André:** There are many context factors to consider. Look at social status as one example. Depending on whether you talk to a friend or your boss, you would use a different style of speaking and gesturing. A dialogue system could learn the style of conversing from data. On the one hand, we need of course a tremendous amount of data to take the large variety of context factors into account. On the other hand, it is difficult to define rules that encode all the implicit knowledge that is involved in conversation. Overall, I believe that statistical approaches are more suitable to simulate specific styles of behavior than manually coded rules—in particular when multiple modalities are involved.

**Dilek Hakkani-Tür:** We are trying to do more than just these tools going at the level of slots and the level of state tracking. If I'm saying "not this but that," when "that"

thing is the value of a slot, then it is possible to transfer what you learn from one domain to another. It is more complex when you go beyond a single slot.

**Phil Cohen:** OK now, James is talking more about a stratified approach in which the life cycle of goals basically is domain independent. James, would you say a little bit more about how to transfer from one domain to another?

**James Allen:** The way we transfer from one domain to the other is as you just described. We have a dialogue system front end that goes all the way from speech recognition and parsing to managing the life cycle of goals and plans. And those components will run in most any domain. We have it running in about six or seven domains right now. But each one of them has a very different backend for task-specific/domain-specific task reasoning. I don't have much to say about a hybrid using statistical learning here because in fact what people typically do is engineer the backend. And, for most of the tasks they're doing, which might involve complex planning or simulation or other things like that, it's not clear how you would ever generate data to be able to build a statistical model for that.

**Phil Cohen:** But you do have a level that's domain general, and then as you said the domain dependent will be more at the level of the action with its arguments and the other types of knowledge that you might have. I think you've worked in the blocks world and you worked in bio-curation.

**James Allen:** Right, we have a dialogue system that interacts with a person manipulating objects in a blocks world, and another with a biologist to build and evaluate causal models of biological processes. Each has very different actions in the backend. In order to have a uniform framework across the domains, even though those actions have domain-specific implementations in the backend, they all have natural language descriptions, which is what people use to refer to them. So we build a representation of those domain-specific actions in terms of how they are described in language. Then as the descriptions move to the backend, they are translated into whatever the representation would be for the backend. That's basically the way we're approaching it. So there are two parts of the interface between the generic dialog management and the backend. One is this level of collaborative problem-solving acts, i.e., introducing goals, refining plans, editing plans, explaining things or whatever, and the other one is this ontology mapping where we're converting from a generic semantic representation computed by the parser to a domain-specific representation that's used by the backend reasoners. This is all hand-engineered. If we could use some machine learning we would love to try and improve the models that way but so far we haven't figured out how we might do that.

**Phil Cohen:** Ron, you had a point?

**Ron Kaplan:** I just want to say that there's one issue, which is having a system or set of tools and capabilities that can transfer from one domain to another, maybe by abstraction, maybe by having domain-independent features that are there, but there's another challenge, which is behaviors that depend on crosstalk between domains. An automotive example: You're on a trip from Austin to Toronto and you want a recommendation for a good place to have lunch for kids. So now you have to have the mapping and the route planning and all that kind of stuff together with the time for lunch and also properties of restaurants in order to satisfy or make plausible recommendations.

**James Allen:** For simplicity's sake so far, I have been talking about our systems like there's a single backend, but in fact you can load as many backends as you want, and when they load, they declare their capabilities. Specifically, they declare what kind of actions they can do, and so then when you get to the point in the dialogue where something needs to be done, the generic system sends out a message asking who can handle this. It's basically an agent-based system. Now that is fairly straightforward to do. What can get complicated, and I'm not sure it rises in your situation but it does arise in some cases, is when some task might and requires reasoning that combines both of them together at the same time. We do not have a good solution for that.

**Ron Kaplan:** Sure, that is a complicated issue that we have mostly ignored so far.

**Phil Cohen:** Planning and logical forms are one step to getting there.

---

**Question: What directions do you think are likely to be the most profitable for dialogue research in the near and medium term?**

**Phil Cohen:** OK. I think we would be remiss in not talking about where we think the future will lie particularly for next generations of students. The commercial world is doing certain things, academics are pushing the envelope in a slightly different direction, where do we think the field is going at this point and what would be the most profitable things [to work on] for near term, medium term, and what students should be thinking about for Ph.D.s?

**James Allen:** I think one of the most exciting possibilities right now is the work going on building "chat bot" conversational agents that perform much better than we had imagined was possible. A very interesting challenge is how do you combine

such systems with more symbolic task-based systems that can do complex tasks, because clearly both approaches have real strengths. Right now there's some attempts to merge the approaches here and there by, say, adding some task knowledge into a neural network or whatever that are driving the dialogue, but there's not been a good integration so far.

**Phil Cohen:** Well, there are people who are trying to do that. [For instance,] Chris Manning is working on that.

**James Allen:** I think that's going to be a really promising area.

**Phil Cohen:** But then we have to push it beyond slot filling given all the other topics we have talked about.

**James Allen:** Exactly what I am saying. Right, slot-filling is not what you are thinking of, yes.

**Phil Cohen:** But that's one step. I mean you'll do slot filling as part of doing something else.

**James Allen:** Yes, I think slot-filling would be low-hanging fruit at this stage.

**Phil Cohen:** But how do we move the field forward that combines task-based reasoning and end-to-end training?

**James Allen:** If I may make one suggestion. It seems that to move the field forward these days, you have to create challenges. And you need to create something that clearly goes beyond the boundaries of what people can do right now, but is not overwhelmingly complicated. So if we had a dialogue system challenge that involved some complex backend reasoning systems, or even making some baby steps towards that, that would be a way to do it. If that doesn't happen, the field is going to move very incrementally because people are very comfortable working on problems where you can easily get data.

**Phil Cohen:** Well having a common backend might be very useful and with enough complexity to it, maybe it's a simulator, maybe it's a planner, maybe it's a mapping system, but it has to have enough complexity that the tasks people can perform are interesting, and potentially difficult.

**Oliver Lemon:** Yes, a shared challenge where statistical approaches could be attempted on plan-based systems, and would be enough data collected and a nice kind of practical problem that people could work on. It would be a very promising avenue for different types of Ph.D. that could be done. But I absolutely agree that we have to look at trying to combine statistical and plan-based systems. That's a big, big topic.

**Ron Kaplan:** I think in creating a challenge like that, I would want to set it up so that there's barely enough data in training something you provide. Because you want it to be enough so that it looks like it has a machine learning set of issues in there to attract people into it, but not so much that they think that, well, if they keep mining the data this way and that way, that will solve the problem.

**Phil Cohen:** The goal is always to do it with the least amount of data.

**Ron Kaplan:** You don't want too much data. You want people to really operate in the sparse data world for all this kind of stuff.

**Oliver Lemon:** Yes, you want a very data-efficient way to do that so as to kind of weave that into the way the challenge is defined.

**Phil Cohen:** But now let's take an orthogonal axis there. How do you throw in incrementality and multimodality into that kind of a scenario?

**Elisabeth André:** Yes, multimodality and incrementality are tricky issues. The research community has organized a number of challenges in the area of social signal processing. However, these challenges focus on offline analysis often using pre-processed and cleaned data to make the analysis tasks manageable. The problem with offline analysis is that the results cannot be transferred to online settings without further redo. As a consequence, classifiers that have been trained for offline analysis often show a poor performance in online analysis. Multimodality has been addressed to some extent in challenges, but incrementality has been very much neglected. So far, there is no challenge on dialogue agents that have to deal with unexpected events, such as misconceptions and interruptions, and exploit the full-range human social cues while interacting with users in naturalistic settings.

**Phil Cohen:** Wow, that's hard. We need to leave it there.

Thank you all for a wide-ranging and informative conversation.

**David Traum (after the discussion)**

Some top topics for students (some of this is already in the discussion to some degree).

1. Hybrid techniques to capture the best of both worlds for knowledge-based reasoning where people can program or discover principles with data-driven, statistical processing. Some such systems exist, but there is not a lot of theory about how and where to put the divide exactly or how these kinds of processing should work together.
2. Along the same lines, but focusing specifically on multimodal expression and understanding of meaning, how do continuous expression modalities

like gesture, prosody, facial expression, etc. interact with more symbolic information like phonemes, words, and syntactic argument structures to get a fuller and nuanced understanding (or expression)?

3. How do systems participate fully in multi-party conversation and multi-conversation interaction, in the sense that multiple conversations or floors are going on at the same time, sometimes with overlapping topics or sets of participants.

## References

- R. Artstein, A. Gainer, K. Georgila, A. Leuski, A. Shapiro, D. Traum. 2016. New Dimensions in Testimony Demonstration. In *Proceedings of North American Chapter of the Association for Computational Linguistics—Human Language Technologies, Association for Computational Linguistics*, pp. 32–36. San Diego, CA. [200](#)
- T. Baltrušaitis, C. Ahuja, L.P. Morency. 2018. Challenges and applications in multimodal machine learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. [206](#)
- T. W. Bickmore and J. Cassell. 2001. Relational agents: a model and implementation of building user trust. CHI 2001: 396–403. DOI: [10.1145/365024.365304](https://doi.org/10.1145/365024.365304). [202](#)
- A. Breen, H. Bui, R. Crouch, K. Farrell, F. Faubel, R. Gemello, W. Ganong III, T. Haulick, R. Kaplan, C. Ortiz, P. Patel-Schneider, H. Quast, A. Ratnaparkhi, V. Sejnoha, J. Shen, P. Stuble, and P. van Mulbregt. 2014. Voice in the user interface. In A. Bhowmik, editor, *Interactive Displays*. Hoboken, New Jersey: John Wiley & Sons, 107–163. [204](#)
- I. Damian, T. Baur, B. Lugrin, P. Gebhard, G. Mehlmann, E. André. 2015. Games are Better than Books: In-Situ Comparison of an Interactive Job Interview Game with Conventional Training. AIED, 84–94. DOI: [10.1007/978-3-319-19773-9\\_9](https://doi.org/10.1007/978-3-319-19773-9_9). [201](#), [207](#)
- L. Galescu, C. M. Teng, J. Allen, and I. Perera. 2018. A generic dialogue system shell based on a collaborative problem solving model. In *Proceedings of the SIGDIAL 2018 Conference*. Melbourne, Australia: Association for Computational Linguistics. [200](#)
- O. Lemon, A. Gruenstein, A. Battle, and S. Peters. 2002. “Multi-tasking and collaborative activities in dialogue systems.” In *Proceedings of the Third SIGdial Workshop on Discourse and Dialog*, Philadelphia. [197](#)
- I. Papaioannou, A. C. Curry, J. L. Part, I. Shalyminov, X. Xu, Y. Yu, O. Dušek, V. Rieser, and O. Lemon. 2017. “An ensemble model with ranking for social dialogue”. *NIPS Workshop on Conversational AI*, Long Beach, CA. [201](#)
- D. Ramachandran, M. Fanty, R. Provine, P. Yeh, W. Jarrold, A. Ratnaparkhi, and B. Douglas. 2015. A TV program discovery dialog system using recommendations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 435–437. DOI: . [204](#)

- D. Traum, J. Rickel, J. Gratch, and S. Marsella. 2003. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 441–448. Melbourne, Australia. DOI: [10.1145/860575.860646](https://doi.org/10.1145/860575.860646). 205
- D. R. Traum, W. Swartout, J. Gratch, and S. Marsella. 2008. A virtual human dialogue model for non-team interaction. Laila Dybkjaer and Wolfgang Minker, editors. In *Recent Trends in Discourse and Dialogue*, pp. 45–67. Springer. DOI: [10.1007/978-1-4020-6821-8\\_3](https://doi.org/10.1007/978-1-4020-6821-8_3). 205
- N. G. Ward, and D. DeVault. 2016. Challenges in building highly interactive dialogue systems. *AI Magazine, AAAI*, 7-18. DOI: [10.1609/aimag.v37i4.2687](https://doi.org/10.1609/aimag.v37i4.2687). 193





# **Nonverbal Behavior in Multimodal Performances**

**Angelo Cafaro, Catherine Pelachaud, Stacy C. Marsella**

## **6.1**

### **Introduction**

The physical, nonverbal behaviors that accompany face-to-face interaction convey a wide variety of information that powerfully influences the interaction. A nod can convey agreement, a gesture can emphasize a point, and facial expressions can convey emotions. A speaker's aversion of gaze reflects they are thinking, in essence regulating cognitive load as they consider what to say next while also signaling they want to hold onto the dialogue turn [Argyle and Cook 1976b, Bavelas 1994]. Nonverbal behaviors are so pervasive in every moment of the dialogue that their absence also signals information—that something is wrong, for example, about the physical health or mental state of the person.

Our interest here in such behaviors lies in efforts to automate the selection and generation of nonverbal behavior for convincing, life-like virtual character performances. Specifically, in this chapter we discuss efforts to generate a character's nonverbal behavior and the many challenges such efforts face. However, most fundamental challenges can be distilled down to determining what behaviors to exhibit and when to exhibit them. The relation between nonverbal behavior and speech is complex. Nonverbals can stand in different, critical relations to the verbal content, providing information that embellishes, substitutes for, contradicts or is even independent of the information provided verbally (e.g., Ekman and Friesen [1969], Kendon [2000]).

In this chapter, we begin by briefly discussing the role that nonverbals play in face-to-face interaction from the perspective of the relation between a participant's

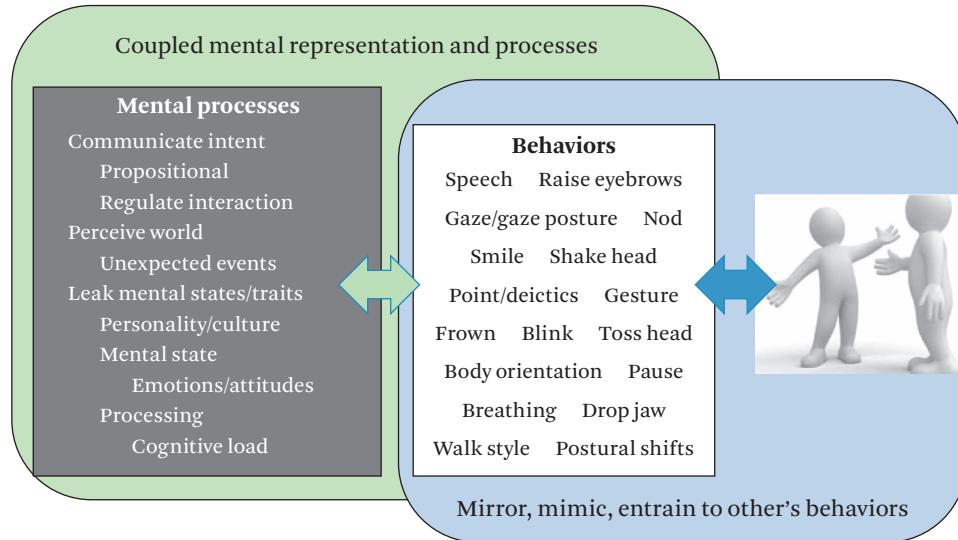
mental states and their nonverbal behavior within the context of an unfolding interaction. The discussion helps identify some of the key complexities of nonverbal behavior. This leads discussion, in the following sections, of the difficult challenges that must be addressed in realizing multimodal behavior. We then explore the various technologies researchers have used to realize these behaviors in virtual characters as well as international standards that have been used to guide the development of these technologies.

## 6.2

### Embodiment: The Mind, Bodies, and Nonverbal Behavior

Figure 6.1 depicts the overall complex relation between internal mental states, nonverbal behavior and the face-to-face interaction between conversants in the real world. In this section we tease apart this diagram to illustrate the challenges it presents for research on creating multimodal behavior in virtual agents.

As Figure 6.1 posits, face-to-face interaction is through the body. The interaction is over a wide channel of multimodal information that flows between participants including vibration of vocal cords, facial expressions, head movements, postures, hand gestures, and gaze. This wide channel exists in large measure because there is also a wide channel of internal mental states and processes that map to or drive these behaviors. The channels are bidirectional. A person's mental states are



**Figure 6.1** Mind and body in face-to-face interactions.

influenced by the other participant's behavior, of course. In addition, a person's own physical behavior can influence their mental state. In this section, we lay out some of the complexity of this mapping between internal states and behavior as it relates to face-to-face interaction.

### **6.2.1 Mapping of Mental States to Nonverbal Behavior**

The left side of Figure 6.1 roughly classifies the types of mental states that inform nonverbal behavior into states associated with the communicative intention, perceptual states associated with a participant being embedded in a larger physical world and leakage of information that we can characterize as unintentional. As we will see, these distinctions tend to break down but nevertheless they provide a useful starting point.

**Communicative Intentions.** There is a wide array of behavior that we use to intentionally convey referential content. For example, nods and shakes can signal agreement or disagreement. One can make reference to a physical object in the space shared by conversants by using a pointing (or deictic) gesture, for example pointing while saying "That is the door to my office"). One can depict an object's physical properties using what are called iconic gestures. For example, a person may show the size of an object with a two arms gesture that spreads the hands apart. In the case of metaphoric gestures, the form of the nonverbal behavior is tied to physical metaphors that treat abstract ideas as if they were physical objects and properties and action taken on the ideas become properties of, and action taken on, the physical object. One can, for example, illustrate the rejection of an idea using a sideways flip of the hand that suggests discarding an object [Calbris 2011].

Nonverbal behaviors also serve a variety of discourse and rhetorical functions. Shifts in topic can be marked by shifts in posture or shifts in head pose. Comparison and contrasts between abstract ideas can be emphasized by abstract deictic (pointing) gestures that point at the opposing ideas as if they each had a distinct physical locus in space [McNeill 1992].

Nonverbal behavior also helps regulate the conversation, for example by signaling the desire to hold on to, get, or hand over the dialogue turn [Bavelas 1994].

**Leakage.** A wide range of mental states and character traits are essentially leaked in the sense that they may not be part of the explicit intended communication. For example, gaze [Argyle and Cook 1976a], specifically what one is gazing at, reveals thought processes, gaze aversion signals cognitive load, blushing suggests

### Glossary

**Alignment** is the coordination along one modality (e.g., verbal modality) that occurs when interlocutors reach a common understanding [Pickering and Garrod 2004].

**Beat gesture** is “defined as movements that do not present a discernible meaning, [ . . . ]. They are typically biphasic (two movement components ), small, low energy, rapid flicks of the fingers or hand” [McNeill 1992].

**Behavior Markup Language bml** is an XML-like mark up language specially suited for representing communicative behavior.

**Chameleon effect** “refers to non-conscious mimicry of the postures, mannerisms, facial expressions, and other behaviors of one’s interaction partners, such that one’s behavior passively and unintentionally changes to match that of others in one’s current social environment” [Chartrand and Bargh 1999].

**Coping** is the process of dealing with emotion, either by acting externally on the world (problem-focused coping), or by acting internally to change beliefs or attention (emotion-focused coping).

**Embodied Conversational Agent** is a virtual or robotic human-like character that demonstrates many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication.

**Empathy** is the capacity to put oneself in the shoes of another one. Emotional empathy refers to feel what the other is feeling while cognitive empathy to understand what the other feels by taking his perspective [Paiva et al. 2004].

**Function Markup Language** is an XML-like mark up language specially suited for representing communicative intentions.

**Gesture phases** is a type of communicative gesture [McNeill 1992]. The optional phase preparation brings the hand into the gesture space. This may be followed by a *pre-stroke hold* where the hand hold its position until the stroke. The *stroke* corresponds to the forceful part of the gesture. It carries the meaning of the gesture and is synchronized with the linguistic segments it “coexpressed.” A *post-stroke hold* phase may follow where the hand remains in its position. Finally, the hand may come to a rest position within the *relaxation* phase.

**Ideational unit** a unit that makes up discourse and that may span over several gestures that convey related information.

**Imitation** is the production of a behavior, be it verbal or motor, with a learning or a communicative goal that was perceived earlier [Nadel and Butterworth 1999].

**Mimicry** “is behavior displayed by an interaction participant who does what another person does and refers to an automatic tendency to imitate others” (Van Baaren et al. [2003] in Glas and Pelachaud [2015].

**Glossary** (*continued*)

**Mirroring** is a similar definition as mimicry.

**Rapport** is a feeling of mutual attentiveness, positivity and connection with another [Zhao et al. 2014, Huang et al. 2011].

**Synchronization** can be defined as “the dynamic and reciprocal adaptation of the temporal structure of behaviours between interactive partners” [Delaherche et al. 2012].

shyness, and facial expressions intentionally or unintentionally can convey emotions and attitudes. This latter point is critical. In practice, any observable behavior communicates information, even though we may characterize it as leakage of an internal mental state such as a facial expression of anger. This in turn means it can also be used intentionally by a person to convey anger even if the person is not angry. In fact, the relation between emotion and facial expression is a central debate in research on emotion in psychology. See [Russell and Fernadez-Dols \[1997\]](#) for a discussion.

### 6.2.2 Overall Characteristics of the Mapping

The mapping between these communicative functions and the behaviors that realize them is many-to-many. Parts of the utterance can be emphasized using a hand gesture, a nod, or eyebrow raise. On the other hand, a nod can be used for affirmation, emphasis or to hand over the dialogue turn [[Kendon 2002](#), [McClave 2000](#)]. The context in which the behavior occurs can transform the interpretation, as can change in the dynamics of the behavior: head nods signaling affirmation vs. emphasis typically have different dynamics. Moreover, behaviors can be composed with each other, further transforming their interpretation.

A range of cultural, individual, and situational factors influence this mapping as well as modulate the form and manner of any behavior used.

**State, Trait Factors.** In particular, individuals differ considerably in amount and types of behaviors they exhibit. For example, even a casual observation of people will reveal differences in the frequency of gesturing, with some people rarely gesturing while others gesture frequently. They will also differ in the types of gestures. Some people largely rely on **beat gestures**, simple rhythmic motions of the hands tied to prosody and specific points the speaker is emphasizing while other will more

extensively use iconic and metaphoric gestures. Even when using similar gestures, there will still be differences between people in form and manner of the gesture. Further, mental and physiological state, such as high arousal, can have a powerful influence on nonverbal by modulating the types, frequency, and dynamics of behaviors.

**Cultural Factors.** There are also considerable cultural differences in nonverbal behavior and how it relates to underlying mental states [Matsumoto 2006]. A classic example of this is the thumbs-up gesture. In some western cultures, this signals agreement or compliment of a job well done while in some cultures it is an obscene, derogatory gesture. Cultural differences may be found not only for emblematic gestures are those just mentioned but for communicative gestures. For example, references to time happen on an axis perpendicular to the speaker; but the front means future events in some cultures or past events in some others [Calbris 2011]. Similarly, gaze patterns and proxemics differ across cultures. Even smiles convey different information and more generally there are different norms about what can be expressed.

**Physical and Social Situational Factors.** Consider the following situations: talking to a friend, talking to a boss, talking across a busy intersection, talking up close in a quiet room, or lecturing. Each of these will tend to elicit different behaviors. Talking up close one can use subtle behaviors, for an eyebrow lift to convey emphasis or surprise. Across an intersection or during a lecture such subtle behaviors will be missed and therefore one may rely instead on beat gestures. Talking to a peer or friend may elicit gestures and gaze patterns that would be avoided when in a more regulated context such as talking to a more dominant individual such as a boss.

### 6.2.3 Shared Representations, Metaphor, and Metaphoric Gesture

One of the most intriguing aspects of nonverbal behaviors is that they transform mental states into physical behaviors. This is especially intriguing when the mental state is comprised of abstract concepts. For example, ideas are immaterial. To reject them by a sideways flip of the hand suggests that they are conceptualized as concrete objects with physical features such as form and location and that we can act on them. This view is in line with the Embodied Cognition paradigm that argues that the same set of sensory and motor representations we use to make sense of and act in our world are also used to make sense of, reason and communicate about abstract concepts [Kendon 2000, Tversky and Hart 2009]. Thought, and the message

to convey, is therefore construed in terms of concrete elements and actions. For example, an idea conceptualized as a concrete object potentially inherits physical properties such as a size, location or weight. A metaphorical mapping process can provide an important idea with a big size, an uncertain idea with a floppy shape, and locate an old idea on the left. Beyond offering a physical representation to abstract elements, embodied cognition considers that reasoning and thought processing are actions taken on these conceptual representations, and that gestures, in particular metaphoric gestures, are physical representations of these actions realized at the conceptual level [Hostetter and Alibali 2008]. In other words, holding an idea in our hand of rejecting it by a sideway flip of the hand is a mirroring of actions taken at the conceptual level, i.e., considering an idea to examine or dismissing it.

#### **6.2.4 Nonverbal Behavior Feedback on Mental States**

**Nonverbals Impact Mental States.** Embodied cognition goes beyond arguing that mental states and processes employ physical and perceptual representations, which in turn influences nonverbal behavior. There is a body of work, going back to Darwin [1872], that argues that physical behavior, such as head movements [Wells and Petty 1980] and facial expressions [Edelmann and Zajonc 1989] influence internal mental states. For example, smiling induces happiness. Similar work has that nodding while making a statement leads one to find arguments more persuasive [Wells and Petty 1980].

**Nonverbals Impact on Nonverbals: Mirror, Mimic.** During an interaction, people may mirror each other behavior [Giles et al. 1991]. Conversational partners may show the same posture, pick up on the same vocabulary [Garrod and Anderson 1987], and use similar prosody features, or facial expressions [Dimberg 1997]. This coordination among the conversational partners is not only seen in the similarity of the signals they display but also how they temporally arrange their signals. Signals are tightly connected in time [Condon and Osgton 1967]. A partner may respond to the smile of another partner or to another emotional expression [Bourgeois and Hess 2008]. Such phenomenon has been referred to in many terms such as *alignment* [Pickering and Garrod 2004], *imitation* [Nadel and Tremblay-Leveau 1999], synchronization, *mimicry*, *mirroring*, and the *chameleon* [Lakin et al. 2003]. The act of imitation may signal engagement [Poggi 2007], common understanding, *empathy* [Nadel and Tremblay-Leveau 1999], and *rappor* [Tickle-Degnen and Rosenthal 1990].

## 6.3

### Toward Building Multimodal Behaviors Control Models

Creating computational models of communicative gestures require an *ECA* be able to display these gestures as well as be able to know when to do a gesture and which one. To build behavior control models, researchers have relied on data described in the psycho-linguistic literature (cf. works by [Kendon \[2004\]](#), [McNeill \[1992\]](#), [Calbris \[2011\]](#)) but also on the manual or automatic analysis of human data. In this section we report on efforts carried over to gather data and annotate them as well as we present main approaches to animate virtual agents.

#### 6.3.1 Data Collection and Annotation

Collecting data requires one to address many issues such as deciding on contextual factors (monolog, dyadic, group discussion, etc.); the number of participants to be recorded; which technology to use; which camera angle should be selected for video data; and so on.

**Naturalistic vs. Acted.** A particular issue regards the type of participants that are recorded. Should they be actors acting out a scripted dialogue [[Douglas-Cowie et al. 2007](#)]? Should they be naive participants unaware of being filmed [[Scherer and Ceschi 1997](#)]? Should they be naive participants in a controlled environment [[Douglas-Cowie et al. 2007](#)]? Each of these choices has pros and cons. Recording data in a controlled environment ensures defining precisely the scenario, controlling all the variables, to set carefully the recording material. However, recorded data often lacks of naturalness and spontaneity. Recording data in a natural environment overcomes this problem but it is highly controlled by ethical committees making it very difficult to perform. So most of the current databases are recorded in a well designed setting. To limit the lack of naturalness, researchers have made use of professional actors [[Bänziger and Scherer 2010](#)] or have developed induction techniques to frame participants in a specific state of mind [[Douglas-Cowie et al. 2007](#)].

Annotating data may be done at different levels, from signals to dialogue acts, mental states, emotions, and attitudes. FACS [[Ekman et al. 2002](#)] is a well-known schema to describe muscular activity involved in facial expressions. MUMIN [[Allwood et al. 2007](#)] covers the multimodal behaviors within an interaction. Several standards have been developed: The W3C coding schema EmotionML [[Schröder et al. 2011](#)] allows representing emotional states following the main theoretical approaches of emotion (namely discrete, dimensional, and cognitive). Dialogue acts can be annotated with ISO DIS 24617-2 [[Bunt 2014](#)]. Data may be annotated discretely by segmenting the data and adding a label for each segment or continuously

(e.g., using GTrace [Cowie et al. 2011]). The level of granularity of the annotation is another variable; this is particularly pertinent when annotating high level variables such as emotional states or social attitudes. Interested readers on the issue of data collection and annotation may be interested in Cowie et al. [2011], Douglas-Cowie et al. [2007], Douglas-Cowie et al. [2011] and Jokinen and Pelachaud [2013].

### 6.3.2 What to Communicate

Behaviors display an extremely large palette of information from own mental and emotional states, indication of the world, attitude on own and other's actions, etc. An upward movement of the eyebrow may indicate surprise, signal new information, mark an emphasis [Ekman 1979] or even be used to greet someone [Walker and Trimboli 1983]. Head nods may mean "yes" in some countries, coincides with new information, indicates approval, etc. Several taxonomies have been elaborated to encompass this variety. They differ on the types of information and on the modalities they consider. Gestures [Kendon 2004, McNeill 1992] can be classified as deictic to indicate a point in space as reference to physical or abstract entity, metaphoric when it represents an abstract idea, iconic to mimic a physical property of an entity, emblem when it replaces a word. Ekman [1979] differentiates expressions depending on whether they are linked to emotions, prosody, punctuation, or turn-taking management. Poggi [2007] classifies nonverbal behaviors depending on their communicative meanings. They can provide information on the speaker identity (age, gender, culture), the word (point in space), and mental state. For this later category, a behavior may communicate on his/her beliefs, intentions, meta-cognition and emotions. Taxonomies on gaze [Argyle and Cook 1976a], emotions [Ortony et al. 1988, Scherer 2000], and turn-taking [Duncan and Fiske 1985, Clark 1996] have also been proposed.

The display of behaviors depends not only on the message to be communicated but also on the social and cultural settings of the interaction, the role of the interactants, and their interpersonal relationship. Expressions of emotions may be masked, played down, or exaggerated depending on who is our interlocutor. More gestures may be done when explaining a task to beginners or giving indication to foreigners. One may use more facial expressions in a noisy environment. Ekman [2003] talks about display rules that state which information (in particular emotional states) can be shown how, when and to whom. Thus, there is not a simple mapping between the type of information to be conveyed and the behaviors used to elicit it.

Computational models of nonverbal behaviors for ECAs ought to encompass the variability and complexity of nonverbal behaviors. They often rely on existing

taxonomies. Some make use of a lexicon; that is a dictionary mapping communicative meanings and multimodal behaviors. Several of the ECA models do take into consideration the socio-cultural factors to determine which behaviors to display.

### 6.3.3 Behavior Synchronization

The generation of the nonverbal behaviors must additionally take into account that they are synchronized, often tightly, with the dialogue and changes in this synchronization can lead to significant changes in what is conveyed to a listener. For instance, the stroke of a hand gesture, a nod, or eyebrow raise performed individually or together are often used to emphasize the significance of a word or phrase in the speech. To achieve that emphasis the behavior must be closely synchronized with the utterance of the associated words being emphasized. Alteration of the timing will change what words are being emphasized and consequently change how the utterance is understood.

## 6.4 Approaches

In this section we describe several approaches that have been adopted for multimodal behavior generation. We first describe the SAIBA (Situation, Agent, Intention, Behavior, Animation)<sup>1</sup> framework, a reference architecture for multimodal behavior generation that makes clear the distinction between communicative functions (i.e. intents) and behavior. Then, we discuss the broad differences that characterize existing techniques, we provide a review of specific complete approaches and we conclude with a comparative summary.

### 6.4.1 The SAIBA Framework

#### 6.4.1.1 Communicative Functions vs. Behavior

Many ECA systems [Cassell et al. 1999, Niewiadomski et al. 2009, Cassell et al. 2001, Vilhjálmsdóttir 2005] adopted the strategy of using separate interfaces to specify an agent's communicative function and its communicative behavior at two levels of abstraction, where the functional level determines the intent of the agent, that is what it wants to communicate and the behavioral level determines how the agent will communicate by instantiating the intent as a particular multimodal realization.

This separation can be seen as two independent components where one component represents the *mind* of an agent and the other component represents the

---

1. <http://www.mindmakers.org/projects/saiba>.

*body*. During a user-agent dyadic interaction, for example, the agent's mind decides what function to accomplish (e.g., greeting), while the body receives what the mind decides to communicate and renders it at the surface level, according to available communication channels and capabilities of the agent.

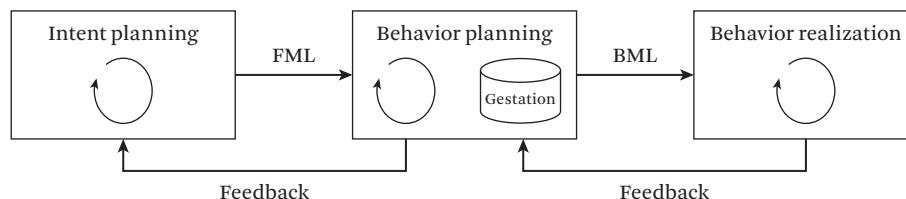
This design strategy has several advantages. First of all, the agent's mind can produce decisions and intents independently of the body, so for example the same mind can be used for different agent's embodiment (e.g., virtual vs. robotic) or shared across systems. Second, the same communicative function can be delivered with different surface forms (i.e., verbal or nonverbal behavior in case of ECAs) depending on the mental state of the agent or intended attitude that the agent aims to show off toward the user. Thus, an agent that wants to express a friendly attitude toward the user might accomplish the same function (e.g., greeting) by using different nonverbal behaviors compared to another agent that aims to express hostility.

This strategy and the need for sharing and reusing existing working components to speed up the process of getting full conversational systems up and running, led research groups in the ECA community to propose the SAIBA framework.

#### 6.4.1.2 SAIBA

The SAIBA framework is the result of an international effort to unify a multimodal behavior generation framework for Embodied Conversational Agents [Kopp et al. 2006].

This framework divides the overall behavior generation process into three subprocesses, as depicted in Figure 6.2, each bringing the level of communicative intent closer to actual realization through the agent's embodiment [Vilhjálmsson 2009]. The interfaces connecting the components are one at the high level, between intent planning and behavior planning, and another interface at the lower



**Figure 6.2** The SAIBA framework for multimodal behavior generation, showing how the overall process consists of three sub-processes at different levels of abstraction.

level, between behavior planning and behavior realization. They are called **FML** [Heylen et al. 2008, Cafaro et al. 2014] and **BML** [Kopp et al. 2006, Vilhjálmsdóttir et al. 2007], respectively. As opposed to the W3C standards for multimodal behavior interaction described in Chapter 9 they are designed to be independent of (1) a particular application or domain, (2) independent of the employed graphics and sound player model, and (3) to represent a clear-cut separation between information types (function-related vs. process-related specification of behavior) [Kopp et al. 2006].

**FML** describes communicative and expressive functions without any reference to physical behavior, representing in essence what the agent's mind decides. It is meant to provide a semantic description that accounts for the aspects that are relevant and influential in the planning of verbal and nonverbal behavior. An FML description must thus fulfill two tasks. First, it must define the basic semantic units associated with a communicative event. Secondly, it should allow the annotation of these units with properties that further describe communicative function such as expressive, affective, discursive, epistemic, or pragmatic functions [Kopp et al. 2006, Cafaro et al. 2014].

**BML** describes the behaviors to express given a function, therefore multimodal behavior should be described so that it can be used to control an agent [Vilhjálmsdóttir et al. 2007]. The last stage handles the realization of the behavior by interpreting the incoming BML and making sure the virtual character behaves accordingly. The behavior realization depends on the particular realization model and can be very diverse. Animations for example can be procedural or fixed and chosen from a repository. Sounds can be generated by a text-to-speech engine or played from file. Therefore, what is specified by BML is independent of any specific realization method [Kopp et al. 2006, Vilhjálmsdóttir et al. 2007].

The framework also presents a *Gesticon*, that can be used by the *Behavior Planner*. This Gesticon is a dictionary which could contain predefined BML behavior definitions. Currently BML has been standardized to a first official version adopted by international researchers, however a unified language specification for FML is still work in progress (c.f. Cafaro et al. [2014]).

#### 6.4.2 Broad Differences and Techniques

A number of approaches for generating ECAs' behavior has emerged in the past years that can be broadly categorized in **data-driven**, **rule-based**, or **combined**.

Data-driven approaches are often based on *annotated corpora* and the direct simulation of human behaviors [Rehm and André 2008]. For instance, Chollet and colleagues first automatically extracted sequences of nonverbal signals character-

izing different interpersonal attitudes in a dyadic interaction from a corpus of job interviews [Chollet et al. 2014]. Then the generation model uses a probabilistic framework to compute a set of candidate sequences and then selects the best sequence for expressing the given attitude using a classification method based on the frequent sequences previously extracted from the corpus.<sup>2</sup>

An interesting data-driven approach based on Motion Capture (Mo-cap) data has been proposed in [Ennis et al. 2010, Ennis and O’Sullivan 2012]. This approach aims at minimizing the amount of audio and Mo-cap data (and possibly reusing it on different autonomous agents) while still producing plausible and varying group conversing behavior in simulated crowds. To this end, they investigated users’ sensitivity to visual desynchronization (i.e., when ECAs body motions in the group are misaligned in time) and mismatched audio (i.e., when ECAs’ speech content is not matched to their gestures). They found that humans are more sensitive to visual desynchronization of body motions, than to mismatches between the characters’ gestures and their voices.

Crowdsourcing has been recently used to obtain annotated datasets [Rossen and Lok 2012, Ravenet et al. 2013, Breazeal et al. 2013]. Ravenet et al. [2013], for example, implemented a user-friendly interface for a tool that allows users to directly configure an ECA’s nonverbal behaviors for conveying particular socio-emotional states (e.g., friendliness).

Conversely, a rule-based approach often consists of performing statistical analysis of human-data and deriving rules for guiding the generation process [Cassell et al. 2001, Bickmore et al. 2009, Cafaro et al. 2009, Lee and Marsella 2006]. This approach can also be combined with manually crafted rules grounded on social psychology theories [Pedica and Vilhjálmsson 2012, Cafaro et al. 2016b]. We illustrate this approach with an example. Consider the goal of providing an autonomous ECA (A) with gaze behavior when another agent (B) is perceived (i.e., B is in A’s field of view) depending on its gender in a social simulation environment. A naive rule, statistically derived from human interaction data, could simply be a frequency analysis of the number of gazes happening for each person featured in the analyzed video and categorizing those gaze instances by target’s gender (i.e., gender of the person looked at). We can then make a statistical inference based on the number of gazes at a specific gender compared to another given the total number of persons currently in the field of view of the examined person. The usage of existing social

---

2. <https://www.youtube.com/watch?v=8fmJMzC18C4>

psychology theories can be illustrated, following our example, by establishing that a friendly agent is more likely to look at others compared to an unfriendly one. We then manually craft this rule by giving more probability to look at others when an agent aims at expressing friendliness as indicated in human-human interaction literature.

Combined approaches generating behavior in a continuous manner have been discussed in [Pedica and Vilhjálmsdóttir \[2012\]](#) and [van Welbergen et al. \[2014\]](#). [Pedica and Vilhjálmsdóttir \[2012\]](#), for instance, argue that human behavior is continuous and is modulated by a constantly changing environment, thus the body is being pushed or pulled by an ever present “social force.” They combine the behavior tree-based social steering mechanism with social theories of territorially and face-to-face interaction. At any given moment one or more of these steering forces would motivate agent’s motion in any of the degrees of freedom while implementing the social norms associated with particular situations. For example, only agents in a conversation would need to worry about maintaining their F-formation.<sup>3</sup>

Another combined approach adopts the over-generate-and-filter technique as shown in [Cassell et al. 2001](#). The basic idea is to generate spoken lines of text augmented with plausible gestures’ annotations based on rules that are derived from studies of human-human dialogue. There might be conflicts in the co-occurrence of proposed behaviors such as gestures. Therefore modifiable filters are applied to trim the gestures down to a set appropriate for a particular character [[Heloir and Kipp 2009](#)].

Finally, there is a growing interest in machine learning techniques based on deep learning algorithms [[Chiu et al. 2015](#)] and probabilistic models [[Bergmann and Kopp 2009a](#), [Bergmann and Kopp 2009b](#), [GroBekathöfer et al. 2012](#), [Pecune et al. 2015](#), [Prepin et al. 2013](#)]. [Chiu et al. \[2015\]](#) proposed a gestural sign scheme to facilitate supervised learning and presented a deep learning approach that realizes both the mapping relation between speech and gestures while taking account temporal relations among gestures. Further work seeks to address correlation between gesture timing and audio changes taking into account deep semantics and prosody features [[Marsella et al. 2013](#)].

Focusing on audio instead of text, [Čerešović et al. \[2010\]](#) solved the problem of synchronizing non-verbal behaviors with synthesized speech by using a BNN [[Rojas 1996](#)]. BNNs are used to determine timing data and extract prosody features from speech. More specifically, they estimate word duration and align them with animation in real time.

---

3. [https://www.youtube.com/watch?v=\\_EMLvGnYLww](https://www.youtube.com/watch?v=_EMLvGnYLww)

The following section analyzes in more detail specific systems for behavior generation that use one or a combination of the techniques mentioned in this section.

### 6.4.3 Specific Complete Systems

#### 6.4.3.1 The Virtual Interactive Behavior Platform

The Virtual Interactive Behavior platform (VIB, formerly Greta) [Pecune et al. 2014] is a fully SAIBA compliant system for the real-time generation and animation of ECA's verbal and nonverbal behavior. The modular architecture of this platform supports the interconnection with external tools (e.g., SSI social signal interpretation framework [Wagner et al. 2013], Cereproc text-to-speech engine [Aylett and Pidcock 2007b]) enhancing an agent's detection and synthesis capabilities.

VIB uses a specific implementation of FML named FML-Affective Presentation Markup Language (FML-APML) [Mancini and Pelachaud 2008] which enables the expression of the degree of certainty, meta-cognitive source of information (thinking, remembering, planning), the speech act (called performative), rhetorical relations such as contradiction or cause-effect (named belief-relations), turn allocation, affect, and emphasis. An FML-APML example is shown in Figure 6.3. The internal BML is used as a “shortcut” for including speech content which is normally an instance of verbal behavior and thus should not be within an FML request and synchronize it with intentions (in the FML part for example) by using temporal markers (i.e., mark).

Furthermore, the FML-APML set of tags has some interesting features regarding the timing and importance of communicative intents, the emotional state of the agent, and information on the world. The timing is specified with attributes inspired from the *BML* recommendations [Kopp et al. 2006, Vilhjálmsdóttir et al. 2007] and makes possible absolute or relative timings of intents with symbolic labels for referencing. A *BML* excerpt is shown in Figure 6.4. The emotional state tags gave the possibility to specify an intensity (as a numeric parameter from 0–1) and a regulation type, that was controlling for felt, faked (emotion aimed at simulating), and inhibited emotions (felt but aimed at being inhibited by the agent). The world tag made possible reference to entities in the world and their properties (physical or abstract) [Mancini and Pelachaud 2008].

A Behavior Lexicon contains pairs of mappings from communicative intentions to multimodal signals. The internal behavior realizer instantiates the corresponding multimodal behaviors, it handles the synchronization with speech, and procedurally generates animations for the ECA.

```

<fml-apml>

    <bml id="bml-1" xmlns="http://www.mindmakers.org/projects/BML">

        <speech id="s1" language="en-GB ssml:xmlns="http://www.w3.org/2001/10/synthesis">
            <ssml:mark name="s1:tm1"/> Synchronization time marker
            Hello
            <ssml:mark name="s1:tm2"/>
            my name is
            <ssml:mark name="s1:tm2"/>
            Greta.
            <ssml:mark name="s1:tm4"/>
            <pitchaccent id="pa1" start="s1:tm1" end="s1:tm2"/>
            <pitchaccent id="pa2" start="s1:tm3" end="s1:tm4"/>
            <boundary id="b1" time="s1:tm4"/>
        </speech>

    </bml>

    <fml>
        <performance type="greet" id="p1" start=s1:tm1" end=s1:tm2"/>
        <emotion type="joy" id="e1" start=s1:tm1" end=s1:tm4"/>
    </fml>

```

**Speech (SSML)** ←

Synchronization time marker

Speech specific intentions →

→ **Intentions**

fml-apml>

The diagram illustrates the FML-APML code structure. A large blue box encloses the entire FML-APML request. Inside, a blue bracket on the left labeled 'Speech (SSML)' points to the `<speech>` block. A red box highlights the `<ssml:mark name="s1:tm1"/>` element, with a red arrow pointing to it labeled 'Synchronization time marker'. Another red box highlights the `<pitchaccent>` and `<boundary>` elements, with a red arrow pointing to them labeled 'Speech specific intentions'. A red box also highlights the `<performance>` and `<emotion>` elements, with a red arrow pointing to them labeled 'Intentions'.

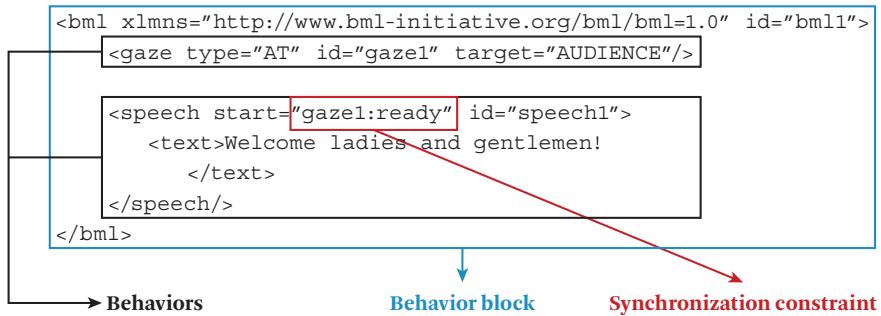
**Figure 6.3** Example of a FML-APML request indicating the speech block specified using the SSML standard and communicative intentions.

#### 6.4.3.2 The Virtual Human Toolkit

The Virtual Human Toolkit<sup>4</sup> (VHT) [Hartholt et al. 2013] offers a collection of modules, tools, and libraries, as well as a framework and open architecture that integrates these components for the creation of ECAs. It offers coverage of subareas including speech recognition, audio-visual sensing, natural language processing, dialogue management, nonverbal behavior generation and realization, text-to-speech and rendering.

In particular, it uses a *Non-verbal Behavior Generator* (NVBG) [Lee and Marsella 2006] module to plan an ECA's nonverbal behavior by using several FML-inspired concepts. The NVBG adopts a rule-based approach that analyzes the agent's spoken text augmented with functional markup to propose nonverbal behaviors in BML. The functional markup, similar to FML-APML, specifies the communicative functions (e.g., speech acts and turn management), cognitive operators that drive the

4. <https://vhtoolkit.ict.usc.edu>



**Figure 6.4** Example of a BML request indicating a speech element starting when the specified gaze behavior is ready to be displayed.

gaze state of an agent and elements that relate to emotional states (e.g., affect states such as joy, distress, fear, etc.), and *coping* strategies (cf. [Glossary](#) or [[Swartout et al. 2006](#)] for a definition of coping). The gaze model associates behaviors with what are called cognitive operators by providing a specification of the form and function of gaze patterns. These functions specify detailed reasons behind a particular gaze behavior related to four determinants: conversation regulation, updating of an internal cognitive state (desire, intention . . . ), monitoring for events and goal status and *coping* strategy. An internal set of rules (written in the XSLT mark-up language) within the NVBG determines which nonverbal behaviors should be generated in a given context and those behaviors are than realized by using SmartBody [[Thiebaut et al. 2008](#)]. SmartBody supports facial expressions, speech, body postures, body gestures, gaze, head motions, feedback notifications, and interruptions of running behaviors.

#### 6.4.3.3 The Articulated Social Agents Platform

The Articulated Social Agents Platform (ASAP) [[Kopp et al. 2014](#)] provides a collection of software modules for building social robots and virtual humans.

The platform embeds a SAIBA compliant architecture that include Flipper [[ter Maat and Heylen 2011](#)] and AsapRealizer 2.0 [[van Welbergen et al. 2014](#)] for behavior planning and generation. Flipper is a library for specifying dialogue rules for dialogue systems, that uses XML-templates to describe the preconditions, effects and BML behaviors of these rules. A simple example of a greeting template is shown in Figure 6.6.



**Figure 6.5** Rachel (left) and Brad (right) are the two representative ECAs of the Virtual Human Toolkit. Permission to use this photo granted by the Institute for Creative Technologies (<http://ict.usc.edu>).

AsapRealizer 2.0 builds on two existing tools that have focused on either incremental multimodal utterance construction [Kopp and Wachsmuth 2004] or interactional coordination [van Welbergen et al. 2010] as isolated problems.

The ASAP platform unifies these fluent behavior realization capabilities and it supports a continuous perception, interpretation, reasoning and behavior generation in order to keep the interaction between the user and the agent natural and fluent. In particular, for behavior generation, it supports incremental behavior plan construction and graceful interruption and adaptation to ongoing behavior. Incremental behavior planning and realization is obtained by constructing a behavior plan out of small increments that allow for the realization behavior early while still planning its construction and executing previous increments. Adaptation can be steered (1) by the behavior planner (top-down), for example when requesting the ECA to speak louder, (2) by the behavior realizer (bottom-up), for instance to achieve co-articulation between gestures on-the-fly, and (3) by external constraints from the environment, for example to support alignment and a tight synchronization between the agent and the user's behavior.

```

<behaviortemplates>
    <template id="1" name="greeting">

        <preconditions>
            <compare value1="$userstates.intention" value2="greeting"/>
            <compare value1="$dialoguestates.topic" value2="greeting"/>
        </preconditions>

        <behavior class="BehaviorToDisplay">
            <argument name="response" value = "Hello, my name is Alice."/>
        </behavior>

    </template>
</behaviortemplates>

```

**Preconditions** ←

**Behavior (speech)** ←

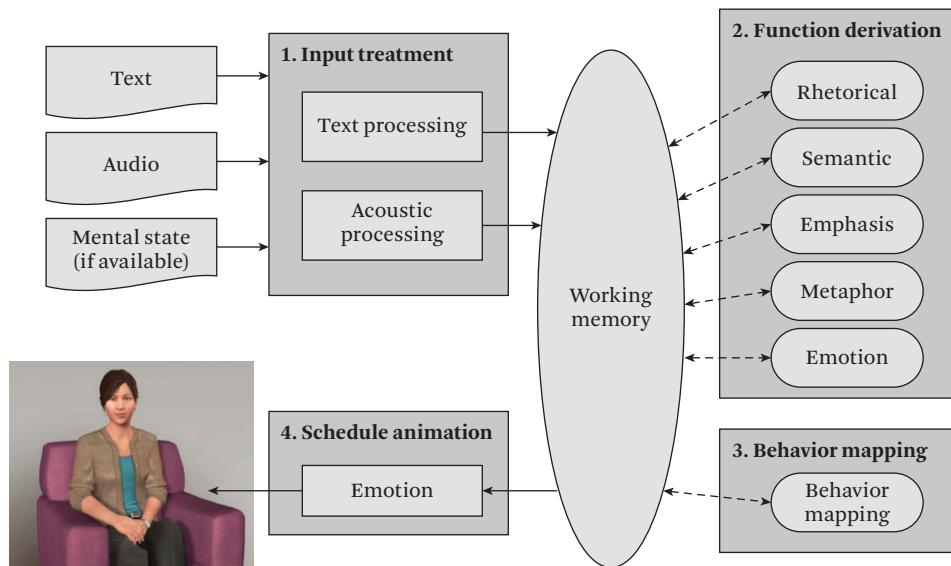
**Figure 6.6** Example of a Flipper template showing a greeting intention.

#### 6.4.3.4 Cerebella

The Cerebella system realizes a flexible technique that leverages information about the character's mental state and communicative intent to generate nonverbal behavior when that information is available [Marsella et al. 2013, Lhommet et al. 2015]. The architecture of Cerebella is depicted in Figure 6.7. It combines prosody and semantic features of an agent's utterance for generating appropriate conversational behavior.

In the absence of explicit input on the character's mental and communicative intent, Cerebella uses acoustic, syntactic, lexical, pragmatic, metaphoric, and rhetorical analyses of the agent's utterance text and audio to infer the associated Communicative Functions (CFs). This includes deriving both the communicative intent of the utterance as well as the underlying emotional and mental state of the speaker.

In either case, regardless of whether the mental states are provided or inferred, the CFs are then mapped to nonverbal behaviors, including head movements, facial expressions, gaze, and gestures, that are composed and co-articulated into a final performance by a character animation system. In addition, working closely with the SmartBody animation system, Cerebella can structure the overall nonverbal performance into ideational units [Xu et al. 2014]. This mapping can use character's specific rules designed to support individual differences including personality, culture, gender, and body types. The final result is a schedule of behaviors described in BML.



**Figure 6.7** Cerebella is a rule-based system that in real time automatically derives the nonverbal behavior of an agent. Starting from an analysis of the lexical and acoustic properties of the utterance, increasingly richer representations are built up spanning rhetorical structure, metaphor use, and emotional content. The communicative function analysis is then mapped to behaviors tailored to the character as well as the overall structure of the nonverbal performance.

#### 6.4.3.5 LiteBody Tool and DTask Dialogue Planner—Relational Agents

*LiteBody* and *DTask* are open source, standards-based tools for creating flash web-based ECAs [Bickmore et al. 2009].

*DTask* is a dialogue planner designed to model and execute system-directed dialogue, with multiple-choice user input. Dialogue is specified in a declarative way, as a hierarchical task decomposition. The accompanying nonverbal behavior of an ECA working with *DTask* is not specified in the dialogue model but it is automatically added at run-time by *BEAT* [Cassell et al. 2001]. *DTask* readily interoperates with *LiteBody*.

*LiteBody* is a web-enabled user interface which renders an ECA given BML commands from a dialogue engine. An http or https based client-server protocol is used for delivering audio and animation scripts from a server to a client (flash enabled browser), and for delivering user input from client to server.

Together, these two tools have been actively used for building ECAs that are particularly effective in the health domain thanks to their relational capabilities. Those agents are defined as “Relational Agents” because they are designed for establishing and maintaining long-term socio-emotional relationship with their users.

#### **6.4.3.6 Behavior Expression Animation Toolkit**

The Behavior Expression Animation Toolkit (BEAT) [Cassell et al. 2001] is a tool to generate, in an XML-based pipeline, multimodal co-verbal behavior based on linguistic and contextual analysis of the text to be spoken [Cassell et al. 2001]. The division between communicative function and behavior is made very clear with the definition of two separate XML tag sets. Therefore, we can consider this system as a precursor of the SAIBA framework architecture. In the XML pipeline an input text message was first annotated with XML tags in terms of various discourse functions related to content and information structure (e.g., theme/rheme, emphasis, contrast, topic-shifts) and interaction processes (turn-taking and grounding). These functions were then mapped into supporting nonverbal behavior for a full multimodal delivery by using a separate set of tags as depicted in Figure 6.8.

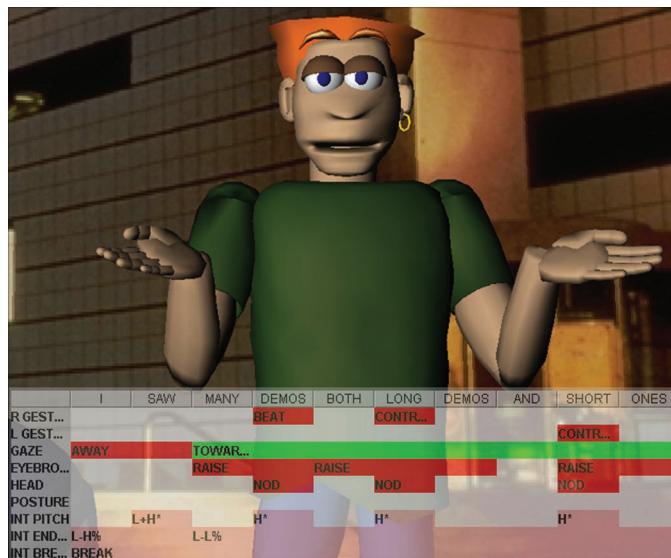
The XML annotation was all done inline with the spoken text and while that made temporal co-occurrence easy to process, it did not allow partially overlapping temporal spans. The term FML was first introduced into this toolkit to describe these tags system to contrast them with the set of tags used to describe the supporting behavior (BML).

#### **6.4.3.7 Other Systems**

Several other specific approaches are briefly reviewed in this section, however they focus on specific application domains, such as healthcare [Rossen and Lok 2012] and interactive narrative [Shoulson et al. 2013], they focus on specific communicative functions (e.g., active listening [Schröder 2010]) or they only generate specific behaviors (e.g., facial and postural expressions [Courgeon and Clavel 2013]).

The Virtual People Factory [Rossen and Lok 2012] is a web-based authoring tool that uses a crowd sourcing approach focused on natural language interaction modeling that can be used by medical and pharmacy educators for creating natural language virtual patient interactions on their own.

ADAPT is a platform that incorporates character animation, navigation, and behavior with modular interchangeable components to produce narrative scenes in interactive multi-agent environments [Shoulson et al. 2013]. A navigation component, based on a centralized event-driven model implemented with parametrized



**Figure 6.8** A screenshot of BEAT in action. After the automatic annotation of communicative functions in the input text, as the time-line below displays, these functions are mapped into supporting nonverbal behavior that can be exhibited in real time by the ECA. Permission to use this photo courtesy of Dr. Hannes H. Vilhjálmsson (<http://www.ru.is/faculty/hannes/>)

behavior trees (PBTs), allows the characters to generate complex locomotion behaviors (e.g., path-finding and obstacle avoidance). A mixed approach consisting of data-driven (e.g., Mo-cap gestures) and procedural animations provides the characters with gaze, gesturing, and sitting behaviors.

The Impulsion Social AI engine<sup>5</sup> combines a number of reactive social behaviors, including those reflecting personal space and group conversational interaction, in a general steering framework inspired by Reynolds [1999]. The engine manages important aspects including the social perception of other agents and the user (in avatar based interactions), locomotion, proxemics, and gaze behavior. Its implementation relies on an event-driven model based on behavior trees and the result is a responsive and continuous steering of agents' body joints [Pedica and Vilhjálmsson 2012].

The SEMAINE System [Schröder 2010] aims at integrating various research technologies for creating a virtual active listener agent. The system adopts a machine

5. <http://secom.ru.is/projects/impulsion/>

**Table 6.1** A summary of the full systems reviewed in this section adopting the different ECA's behavior generation approaches: data-driven (DD), rule-based (RB), and combined (CMB).

System	Year	Approach	SAIBA	Specialty
VIB	2014	CMB	yes	Domain-independent ECAs
ASAP	2014	CMB	yes	Continuous behavior
VHT	2013	CMB	yes	ECA Authoring Environment
Cerebella	2014	CMB	yes	Multi-layer analysis of comm. function
LBody	2009	RB	no	Relational agents
BEAT	2001	RB	no	Synchronized gestures

learning approach for determining the user's turns by analyzing audio-visual features (i.e., prosody, silences, and head gaze behavior). A rule-based approach is then used for synthesize the appropriate back-channel verbal and nonverbal behavior.

The Generic Embodied Conversational Agent (GECA) is a rule-based framework that adopts a scripting XML-like language, named GECA Scenario Mark-up Language (GSML), for defying user-agent scripted interactions [Huang et al. 2008].

The Multimodal Affective and Reactive Characters toolkit (MARC) [Courgeon and Clavel 2013] provides a set of tools for animating in real-time a character's facial and postural animations. It adopts a state-machine's approach for defining an ECA's behavior and it supports BML 1.0 inputs.

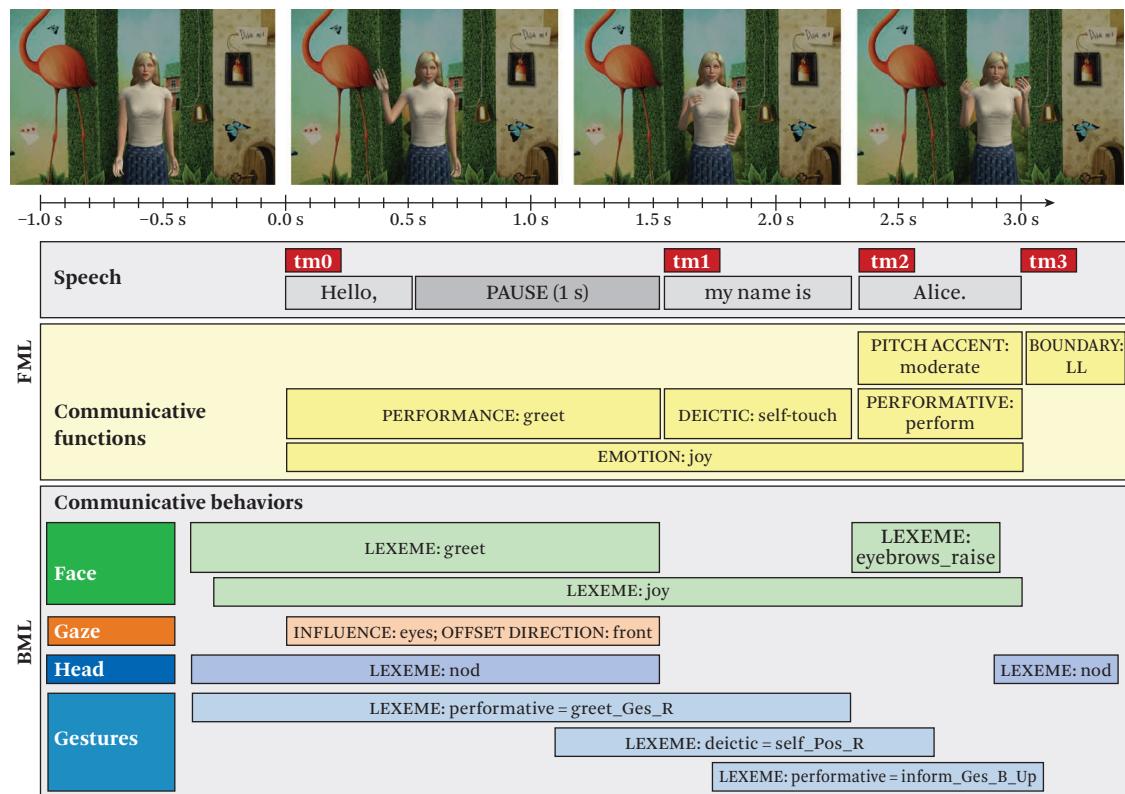
#### 6.4.4 Comparative Summary

A summary of the main system for ECA's behavior generation is provided in Table 6.1. We indicate the year of the latest scientific publication describing it, the approach adopted (i.e., data-driven (DD), rule-based (RB), and combined (CMB)), compliance with SAIBA and the main feature supported by the system (i.e., specialty).

## 6.5 An Annotated Behavior Generation Example in VIB

In the previous section we described the SAIBA framework (Section 6.4.1) and we listed several complete systems for nonverbal behavior generation (Section 6.4.3).

Prior to concluding this chapter, we describe an annotated example of behavior generation within the VIB platform (described in Section 6.4.3.1). The ECA used in this example and depicted in Figure 6.9 is named Alice. It has been deployed



**Figure 6.9** An annotated example of behavior generation within the VIB platform. The FML (i.e., FML-APML) and BML standards are used for describing the ECA's communicative functions (shown under the speech elements) and behaviors. The functions are transformed in behaviors according to a mapping lexicon and probabilistic rules.

within the European H2020 project ARIA-Valuspa,<sup>6</sup> which deals with the creation of affective information retrieval assistants.

For this example, the agent renders through verbal and nonverbal behaviors the following sentence: “Hello, my name is Alice.” In Figure 6.9, the four images depict salient moments where visual behavior is exhibited by Alice in synchrony with speech produced by Cereproc [Aylett and Pidcock 2007a], a TTS engine in real time. At the functional level, she performs two main functions. First, she greets the interlocutor and secondly she delivers an information (i.e., her name). During

6. See more details about ARIA-Valuspa at <http://aria-agent.eu/>

this short interaction, she is in a joyful emotional state. At behavioral level, she accomplishes the greeting with a wave gesture (see second image starting from the left) accompanied with a head nod and an eye gaze directed to the user (the camera). Then, she performs two gestures, she touches herself in synchrony with the synthesized utterance “my name is,” and she exhibits a metaphoric gesture by raising up the hands while uttering her name accompanied with a head nod. The gestures description in this example follow the categorization of [McNeill \[1992\]](#). Finally, joy is expressed through the eyebrows raise and special lexemes associated with facial movements (e.g., smile and cheek raise). Next, we describe in technical terms how these functions are mapped into behaviors that are generated and exhibited by Alice thanks to VIB.

The images are annotated with FML and BML corresponding, respectively, to the communicative functions and behaviors of Alice. The actual scripts that are given in input to VIB (in FML-APML syntax and BML 1.0 standard) are listed in Appendix 6.A. The FML script contains a BML part for the speech definition including the speech elements (i.e., words), prosody (i.e., pitch accent) and several time markers. This speech part is described using the SSML standard of which the pause element is part of. Time markers represent coarse-grained temporal constraints that support the synchronization between communicative functions and speech elements. Those time markers are later transformed from symbolic to concrete values (i.e., time values) by VIB for a more precise synchronization between speech and nonverbal behaviors. The communicative functions are shown underneath the speech part. In the Figure 6.9 they are represented with a label indicating the general category (e.g., performative, emotion) and the specific instance within the category.

The behavior planner in VIB transforms a given FML into BML according to a lexicon, probabilistic rules, and expressivity parameters. The *lexicon* can be seen as a dictionary in which an entry is a communicative function (described with category and type). For each entry (i.e., function), a set of behaviors (involving facial expressions, gaze, gestures, etc . . . ) to accomplish the function is proposed along with several alternatives that are named *behavior sets*. Each behavior set comes with a probability determining the likelihood of being chosen among the others in the entry. A basic lexicon can have, for instance, discrete uniform distributions associated to each behavior set, i.e., each alternative has an equal probability to be chosen with respect to the other alternatives in the same entry. As a result, any given communicative function in the FML script can be accomplished in different ways, due to both the lexicon being used (different lexicons can propose different mappings from functions to behavior sets) and the probabilities defined therein.

In our example, the lexicon that we adopted for the greeting function suggests a wave gesture, an eyebrow raise and a head nod. As an alternative behavior set, only the wave gesture is proposed for this particular function. Once the behaviors are instantiated, the *expressivity parameters* add more variation to the way a behavior can be realized. These parameters, omitted for simplicity from the example in the image, define further aspects of the execution of each behavior such as fluidity and openness (for gestures).

As for the behaviors timing, during the transformation from FML into BML, the time markers assume real values as indicated in the scripts in the Appendix. We summarize here those timings:<sup>7</sup> tm0 (0.0s), tm1 (1.54s), tm2 (2.34s), and tm3 (3.01s).

Time markers, as mentioned earlier, allow the generated behaviors to be in synchrony with the produced speech. The real values, in VIB, are produced by the TTS engine that calculates the time for a specific time marker based on the time needed to synthesize the word(s) preceding it. For the generated behaviors in 6.9, note that some gestures (e.g., *performative = greet\_Ges\_R*) have a negative start time. This is needed by the behavior planner to calculate the precise duration of behaviors. For instance, in the functional specification we indicated that the greeting function must be accomplished between *tm0* and *tm1* in synchrony with “Hello.” The planner foresees the anticipation of Alice’s movements and the result of this computation is the *preparation gesture phases* of the waving gesture starting earlier in order to be in synchrony with the specified part (i.e., *tm0*) at the moment the gesture is in the *stroke gesture phases* (see Sections 6.3.3 and 6.6.1.2 for more details on gestures phases).

Finally, inverse kinematics animation techniques are used by VIB for the realization of the behaviors. Alice’s joints rotations and positions are animated according to the behavior that has been planned (see Huang and Pelachaud [2012]) As part of this animation process, blending of certain behaviors is necessary. Consider the facial lexemes that constitute the expression of joy (e.g., smiling) and the blending with lip movements generated when uttering the synthesized speech. When behavior realization conflicts arise, those are solved through an importance parameter (c.f. the scripts in Appendix 6.A).

---

7. N.b. the behavior elements in Figure 6.9 are aligned to speech elements according to the real timings as well.

## 6.6

### Conclusion and Future Trends

This chapter discussed the roles and properties of nonverbal behaviors, as well as presented existing approaches to build behavior controllers. By reviewing the systems presented in Section 6.4.3 there seems to emerge the need for diverse combined techniques fulfilling specific requirements (e.g., data-driven machine learning models accounting for deep semantics and continuous behavior generation) or addressing some limitations (e.g., rule-based statistical models when there is lack of large datasets).

The adoption of data-driven approaches, for example, using annotated corpora, allows researchers to empirically validate behavior design and generation as exemplified in Cassell's *Study-Model-Build-Test* development cycle (c.f. [Cassell \[2006\]](#)). However, those approaches require the collection of large amount in order to derive patterns of behaviors emerging from concrete instances of human behavior. Furthermore, the collected data require annotations that sometimes are difficult to obtain automatically, although new tools that facilitate the automatic and semi-automatic annotations are becoming available [[Wagner et al. 2013](#), [Baur et al. 2013](#)].

Crowdsourcing represents an alternative solution for obtaining manual annotations (e.g., using Mechanical Turk's workers) but it requires experienced workers (i.e., annotators) if a complex annotation schema is used.

The data-driven approaches based on Mo-cap data produce more believable and natural behaviors compared to rule-based ones. Nevertheless, they require a costly setup involving experienced actors and time-consuming post processing operations on the acquired data [[Menache 2011](#)]. Another downside on Mo-cap-based techniques is that generative modules based on those need to cope with situations for which appropriate Mo-cap data and speech samples are missing [[Rehm and André 2008](#)].

Machine learning techniques for deriving generative rules automatically is a promising approach [[Čereković et al. 2010](#), [GroBekathöfer et al. 2012](#), [Kipp 2006](#), [Bergmann and Kopp 2009b](#), [Bergmann and Kopp 2009a](#), [Chiu et al. 2015](#)]. Unfortunately, this approach requires very large training datasets for deriving generalized sets of rules. [Ding et al. \[2014\]](#) proposed a strategy for addressing this issue in the context of laughter behavior generation. They adopted a combined machine learning approach where a data-driven animation model first learns from a collected laughter corpus of Mo-cap data the relationship linking phonemes duration and lip shapes in a GMM. Then they compute an interpolation function based on Hidden Markov Models (HMMs). Finally, in the synthesis step, the lip shape GMM is used to infer a first lip shape stream from the inputs and it is smoothed by the HMM interpolation resulting in a synthesized real-time lip animation.

### 6.6.1 Challenges to Multimodal Behavior Design

To conclude we report challenges on the design of Embodied Conversational Agents, not from a technical point of view, but rather questioning which human multimodal behaviors properties should be modeled.

#### 6.6.1.1 Behavior Coherency

Most of the existing computational models of ECAs behavior generation rely on fixed lexicon and do not consider the sequence of behaviors. That is they do not consider the meaning conveyed by these behaviors nor their shape [Xu et al. 2014]. However, as Burgoon and Poire [1999] say: “No nonverbal cue is an *island*. It is continually surrounded by a host of nonverbal behaviors which together may delimit and clarify meaning.” The meaning of a behavior is influenced by the surrounding behaviors, emitted either by the locutor or interlocutor (user or ECA, and vice versa). Gesture shape shows also coherency with the surrounding gestures. Calbris [2011] highlights how successive gestures within an *ideational unit* (c.f. [Glossary](#)) share shapes; they are differentiated by an element of the shape (e.g., a hand configuration or movement) allowing to distinguish a new gesture is produced; and this element acts as a demarcative function between successive gestures.

#### 6.6.1.2 Timing and Synchronization

Achieving synchronization across modalities in a virtual character can be difficult, especially in the case of behaviors such as hand gestures that involve relatively large-scale motion and multiple *gesture phases*. Consider a *beat gesture*, a staccato, and a stroke of the hand that can be used to provide emphasis. To perform a downward stroke, the hand must be raised in preparation for the stroke. After the stroke, the hand can be held in a post-stroke hold pose to provide further emphasis, followed by a relaxation to a rest position. This sequence of behaviors occur in alignment with the speech, so there must be sufficient time to prepare for the stroke, plus the stroke and any post-stroke hold must be tightly coordinated with the parts of the dialogue that is being emphasized. The relaxation may also need to take into account any co-articulation required for subsequent gestures to be performed.

Such challenges make the pattern and timing of the behavior animations that accompany utterances unique to the utterance and the state of the character. Manual creation of the behaviors by hand animation and/or Mo-cap are consequently time consuming and costly, as well as requiring considerable expertise from the animator or the Mo-cap performer.

### **6.6.1.3 Interactive and Incremental Unfolding**

The realization of nonverbal behavior must also take into account that face-to-face conversations are highly interactive, with addressees reacting to the speakers with backchannels or interrupting the speaker. Speakers are also altering midstream what they are saying, reacting to their own unfolding thought processes as they incrementally construct the utterance, interpret their own utterance as they speak as well as interpret an addressee's reactions to what has been said. This poses significant challenges on any technologies that hope to replicate such properties in a virtual human. Simple approaches that fully plan a multimodal presentation and then execute it won't suffice in such a highly dynamic interaction that requires fine grain blending of planning and reaction during execution. Adaptation of interactants needs to happen at various levels: at emotional level (e.g., to show empathy) up to behavioral level (e.g., to mimic a smile). Lately, on-the-fly behavioral adaptation relying on neural networks has been proposed [Pecune et al. 2015]. This approach supports the influence on ECA's behavior based on user's behavior. However, more recent work focused on re-planning at functional level (e.g., when an interruption occurs) [Cafaro et al. 2016a].

### **6.6.1.4 Influence of Nonverbal Behaviors**

Now one might justifiably argue that the influence of nonverbal behavior on internal mental states is perhaps less relevant when creating artifacts that exhibit such behaviors since the body to internal mental state relation is not necessary to the design of the artifact. However, the picture is less clear when one shifts the focus to the human interactant in a human-virtual human interaction.

People tend to mirror or mimic each others behaviors, for example facial expressions [Dimberg 1997] and such mimicry has been noticed in human-virtual human interactions. This leads to the potential that by exhibiting certain behaviors, a virtual human could induce mimicry in the human interactant, leading to changes in the human's mental states such as attitudes about the virtual human. Of course, one should be circumspect about the magnitude of such effects, but the designer of an artifact may well want to take them into account, since such mimicry can have positive effects on human-human social interaction. One area where virtual human research has extensively studied this is in the case of rapport [Tickle-Degnen and Rosenthal 1990] between people engaged in a conversation, evidenced by positive emotional displays, mutual attention as well as physical mimicry and synchronization. The use of these behaviors in virtual humans has been shown to elicit rapport in human-virtual human interactions and to positively impact the social interaction [Gratch et al. 2007].

### 6.6.1.5 Yardstick: Naturalistic vs. Effective Nonverbal Performance?

How a person's mental states are encoded into a behavior is not necessarily how an observer will decode that behavior [Gifford 1994]. Put another way, the mapping between mental states and behaviors depicted in Figure 1 is not shared knowledge between participants in a conversation. To draw inferences from the behavior of other participants in the conversation, a person will effectively be using a different mapping, leading to inaccurate inferences. The fact that there are cultural and individual differences complicates this inference further. Even if the mapping was shared knowledge, the many-to-many mapping by itself induces ambiguities.

This raises the question of how a designer should realize a nonverbal performance. Should the designer strive for naturalistic performance in the sense that it is consistent with human behavior or for a performance that is more readily decoded by an observer to ensure it influences the user consistent with the designer's intent. For example, studies of actors [Coats et al. 1999] suggest they tend to emote at an unrealistic rate, the emotions revealed are likely to be more consistent with the situation and show little masking or dissemblance in service of social goals.

The answer to this question often depends on the application. Consider an intelligent tutoring system that is teaching math and uses embodied agents to engage and motivate the learner [Robison et al. 2009]. One will likely want nonverbal behavior more readily decoded by the learner or motivating to the learner, as opposed to naturalistic. However, if one is building an application to teach people how to decode people's nonverbal behavior in real-life situations (cf. Luciew et al. [2011]), one might instead want a more naturalistic performance. One may also want to consider graded approaches. Consider an application designed to teach the reading of facial expressions to children on the autistic spectrum. In such a case, the designer may want to start by using easy to decode expressions and then move to more naturalistic behavior.

This raises the question of what the gold standards are for naturalistic vs. effective vs. easily decoded. Naturalistic performance of course is what people actually do and we can attempt to record what people do and use that data to inform our design. In the case of effective or decodable, we may turn to other sources.

The arts have extensively studied the issue of effective nonverbal behavior. Historically, various acting theories have explored formal, often stylized representations of behavior in terms of what they convey. For instance, Delsarte (see Marsella et al. [2006] for a discussion of Delsarte) characterized facial features, postures and gestures in terms of the meaning they conveyed. He also contrasted the meaning conveyed by the temporal ordering of behaviors. For example, he argued that two parts of the body moving in opposite directions simultaneously suggest expressive force, physical, or emotional power, while parts of the body moving in parallel di-

reactions simultaneously suggest deliberateness, planning, intentionality. Delsarte argued that successive movements also convey meaning depending on the order of movements through the parts of the body. For example, smiling, raising a hand to shake, and then moving towards someone to greet them suggests sincere emotions while the reverse order from extremities to the head suggests insincerity.

The animation arts have also explored nonverbal behavior in terms of what it conveyed and how it influenced an audience. Most notably, the work of Disney animators [Thomas and Johnston 1995] laid out a range of techniques, many of which such as squash and stretch are often used in ways that depart dramatically from naturalistic to the point that the behavior is not feasible for a human body.

Finally, research in animal ethology explored the idea of supranormal stimuli that are actually more effective in eliciting responses than the natural occurring stimuli [Tinbergen and Perdeck 1950]. A moment's thought brings to mind many such examples of such stimuli in human culture, such as using Photoshop to re-craft depictions of fashion models, the exaggerated depiction of human bodies in games, the use of eye makeup to make the eyes appear larger, and the various ways we augment our bodies through surgery.

### **Focus Questions**

- 6.1. How are the multimodal communicative behaviors and functions defined?
- 6.2. What are the mechanisms for synchronizing verbal and nonverbal behaviors?
- 6.3. What are the differences between behavior generation approaches?
- 6.4. What are the existing Embodied Conversational Agent (ECA) platforms and what representation languages do they adopt?
- 6.5. What are the current and future challenges?

## **6.A**

### **A Behavior Generation Example in VIB**

In this section we show the FML (i.e., FML-APML) and BML listings that are used in the full example of behavior generation described in Section 6.5.

The following is the full FML-APML script described in the example.

The FML-APML script described in Figure 6.10 results in the BML script shown below when transformed within the VIB platform (described in Section 6.4.3.1). Note that VIB offers several expressivity parameters for head and gesture behaviors (e.g., spatiality, openness, fluidity, intensity, repetitiveness), but those have been omitted from the listing due to space constraints.

```

<!-- FML-APML script -->
<fm1-apml composition="replace">

    <!-- The Speech (i.e., verbal behavior) must be included within a BML tag -->
    <bml>
        <speech id="s1" start="0.0" language="english" voice="cereproc">
            <!-- Time Markers are anchors for the intentions defined below -->
            <tm id="tm0"/>
                Hello, <break time="1s"/>
            <tm id="tm1"/>
                my name is
            <tm id="tm2"/>
                Alice.
            <tm id="tm3"/>
            <pitchaccent id="pa1" start="s1:tm2" end="s1:tm3"
                level="moderate" type="Hstar" importance="1"/>
            <boundary id="b2" start="s1:tm3" type="LL" />
        </speech>
    </bml>
    <fm1>
        <!-- A greeting intention in the performative category -->
        <performative id="p1" start="s1:tm0" end="s1:tm1"
            type="greet" importance="1.0"/>

        <!-- A self-referring intention -->
        <deictic id="p2" start="s1:tm1" end="s1:tm2"
            type="selftouch" importance="1.0"/>

        <!-- An information giving intention in the performative category -->
        <performative id="p3" start="s1:tm2" end="s1:tm3"
            type="inform" importance="1.0"/>

        <!-- An emotional state -->
        <emotion id="e1" start="s1:tm0" end="s1:tm3"
            type="joy" intensity="0.8" importance="1.0"/>
    </fm1>
</fm1-apml>

```

**Figure 6.10** A full example of a script specifying the communicative functions described in FML-APML.

```

<!-- BML script -->
<bml xmlns="http://www.mindmakers.org/projects/BML"
      id="bml1" character="Alice" composition="replace" >

    <!-- The Speech (i.e., verbal behavior) -->
    <speech id="s1" start="0.0" language="english" voice="cereproc">

        <!-- Time Markers are anchors for the intentions defined below -->
        <tm id="tm0"/>
            Hello, <break time="1s"/>
        <tm id="tm1"/>
            my name is
        <tm id="tm2"/>
            Alice.
        <tm id="tm3"/>

        <pitchaccent id="pa1" start="s1:tm2" end="s1:tm3"
                     level="moderate" type="Hstar" importance="1"/>
        <boundary id="b2" start="s1:tm3" type="LL" />
    </speech>

    <!-- Face behaviors -->
    <face id="face0" start="-0.38" end="1.53" amount="1.00" >
        <lexeme lexeme="greet"/>
    </face>

    <face id="face1" start="-0.30" end="2.99" amount="0.80" >
        <lexeme lexeme="joy"/>
    </face>

    <face id="face2" start="2.34" end="2.89" amount="1.00">
        <lexeme lexeme="eyebrows_raise"/>
    </face>

    <!-- Gaze behaviors -->
    <gaze id="gaze0" start="0.01" end="1.54" influence="eyes"
          offsetAngle="0.0" offsetDirection="front"
          ready="0.52" relax="1.03"
          origin="Alice" target="" />

    <!-- Head behaviors -->
    <head id="head1" start="-0.36" end="1.53" lexeme="nod" />

    <head id="head2" start="2.9" end="3.4" lexeme="nod" />

    <!-- Gesture behaviors -->
    <gesture id="gesture0" start="-0.38" end="2.29" lexeme="greet_Ges_R" />
    <gesture id="gesture1" start="1.12" end="2.62" lexeme="self_Pos_R" />
    <gesture id="gesture2" start="1.61" end="3.11" lexeme="inform_Ges_B_Up" />

</bml>

```

**Figure 6.11** A full example of behavior generation described in BML.

## References

- J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3–4): 273–287. DOI: [10.1007/s10579-007-9061-5](https://doi.org/10.1007/s10579-007-9061-5). 226
- M. Argyle and M. Cook. 1976a. *Gaze and Mutual Gaze*. Cambridge University Press. <https://books.google.com/books?id=ZBR0QgAACAAJ>. 221, 227
- M. Argyle and M. Cook. 1976b. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge. 219
- M. Aylett and C. Pidcock. 2007a. The cerevoice characterful speech synthesiser sdk. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, eds., *Intelligent Virtual Agents*, vol. 4722 of *Lecture Notes in Computer Science*, pp. 413–414. Springer-Verlag, Berlin, Heidelberg. 242
- M. P. Aylett and C. J. Pidcock. 2007b. The cerevoice characterful speech synthesiser sdk. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, vol. 4722 of *LNCS*, pp. 413–414. Springer-Verlag. DOI: [10.1007/978-3-540-74997-4\\_65](https://doi.org/10.1007/978-3-540-74997-4_65). 233
- T. Bänziger and K. R. Scherer. 2010. Introducing the geneva multimodal emotion portrayal (GEMEP) corpus. In K. R. Scherer, T. Bänziger, and E. B. Roesch editors, *Blueprint for Affective Computing: A sourcebook*, pp. 271–294. Oxford University Press, Oxford, England. 226
- T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André. October 2013. Nova: Automated analysis of nonverbal signals in social interactions. In A. A. Salah, H. Hung, O. Aran, and H. Gunes, editors, *Proceedings of the Human Behavior Understanding: 4th International Workshop, HBU 2013*, Barcelona, Spain, pp. 160–171. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-02714-2\\_14](https://doi.org/10.1007/978-3-319-02714-2_14). 245
- J. B. Bavelas. 1994. Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction*, 27(3): 201–221. DOI: [10.1207/s15327973rlsi2703\\_3](https://doi.org/10.1207/s15327973rlsi2703_3). 219, 221
- K. Bergmann and S. Kopp. 2009a. GNetIc—Using bayesian decision networks for iconic gesture generation. In *Intelligent Virtual Agents*, vol. 1, pp. 76–89. <http://www.springerlink.com/index/l754644205211847.pdf>. 232, 245
- K. Bergmann and S. Kopp. 2009b. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems—Volume 1*, pp. 361–368. <http://dl.acm.org/citation.cfm?id=1558062>. 232, 245
- T. Bickmore, D. Schulman, and G. Shaw. 2009. Dtak and litebody: Open source, standards-based tools for building web-deployed embodied conversational agents. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, pp. 425–431. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-04380-2\\_46](https://doi.org/10.1007/978-3-642-04380-2_46). 231, 238

- P. Bourgeois and U. Hess. 2008. The impact of social context on mimicry. *Biological Psychology*, 77(3): 343–352. DOI: [10.1016/j.biopspsycho.2007.11.008](https://doi.org/10.1016/j.biopspsycho.2007.11.008). 225
- C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung. 2013. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1): 82–111. DOI: [10.5898/JHRI.2.1.Breazeal](https://doi.org/10.5898/JHRI.2.1.Breazeal). 231
- H. Bunt. 2014. A context-change semantics for dialogue acts. In H. Bunt, J. Bos, S. Pulman, editors, *Computing Meaning, Text, Speech and Language Technology*, vol. 47, pp. 177–201. Springer, Dordrecht. DOI: [10.1007/978-94-007-7284-7\\_10](https://doi.org/10.1007/978-94-007-7284-7_10). 226
- J. K. Burgoon and B. A. Poire. 1999. Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. *Communication Monographs*, 66(2): 105–124. DOI: [10.1080/03637759909376467](https://doi.org/10.1080/03637759909376467). 246
- A. Cafaro, R. Gaito, and H. H. Vilhjálmsson. 2009. Animating idle gaze in public places. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, pp. 250–256. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/978-3-642-04380-2\\_28](https://doi.org/10.1007/978-3-642-04380-2_28). 231
- A. Cafaro, H. Vilhjálmsson, T. Bickmore, D. Heylen, and C. Pelachaud. 2014. Representing communicative functions in saiba with a unified function markup language. In T. Bickmore, S. Marsella, and C. Sidner, editors, *Intelligent Virtual Agents*, vol. 8637 of *Lecture Notes in Computer Science*, pp. 81–94. Springer International Publishing. DOI: [10.1007/978-3-319-09767-1\\_11](https://doi.org/10.1007/978-3-319-09767-1_11). 230
- A. Cafaro, N. Glas, and C. Pelachaud. 2016a. The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, p. 911–920. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. <http://dl.acm.org/citation.cfm?id=2936924.2937059>. 247
- A. Cafaro, B. Ravenet, M. Ochs, H. H. Vilhjálmsson, and C. Pelachaud. July 2016b. The effects of interpersonal attitude of a group of agents on user's presence and proxemics behavior. *ACM Transactions on Interactaction and Intelligent Systems*, 6(2): 12:1–12:33. DOI: [10.1145/2914796](https://doi.org/10.1145/2914796). 231
- G. Calbris. November 2011. *Elements of Meaning in Gesture* 378 pps. John Benjamins Publishing, Philadelphia, PA. 221, 224, 226, 246
- J. Cassell. 2006. Body language: Lessons from the near-human. In G. Redux, editor, *Genesis Redux: Essays in the History and Philosophy of Artificial Life*, pp. 346–374. University of Chicago Press. DOI: [10.7208/chicago/9780226720838.003.0017](https://doi.org/10.7208/chicago/9780226720838.003.0017). 245
- J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pp. 520–527. ACM, New York. DOI: [10.1145/302979.303150](https://doi.org/10.1145/302979.303150). 228

- J. Cassell, H. H. Vilhjálmsdóttir, and T. Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pp. 477–486. ACM, New York. DOI: [10.1145/383259.383315](https://doi.org/10.1145/383259.383315). 228, 231, 232, 238, 239
- A. Čereković, T. Pejša, and I. S. Pandžić. 2010. A controller-based animation system for synchronizing and realizing human-like conversational behaviors. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, eds., *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland*, March 23–27, 2009, Revised Selected Papers, pp. 80–91. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-12397-9\\_6](https://doi.org/10.1007/978-3-642-12397-9_6). 232, 245
- T. L. Chartrand and J. A. Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6): 893. DOI: [10.1037/0022-3514.76.6.893](https://doi.org/10.1037/0022-3514.76.6.893). 222, 764
- C.-C. Chiu, L.-P. Morency, and S. Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In W.-P. Brinkman, J. Broekens, and D. Heylen, editors, *Intelligent Virtual Agents: 15th International Conference*, IVA 2015, Delft, The Netherlands, August 26–28, 2015, *Proceedings*, pp. 152–166. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-21996-7\\_17](https://doi.org/10.1007/978-3-319-21996-7_17). 232, 245
- M. Chollet, M. Ochs, and C. Pelachaud. 2014. From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In T. Bickmore, S. Marsella, and C. Sidner, eds., *Intelligent Virtual Agents: 14th International Conference*, IVA 2014, Boston, MA, USA, August 27–29, 2014. *Proceedings*, pp. 120–133. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-09767-1\\_15](https://doi.org/10.1007/978-3-319-09767-1_15). 231
- H. Clark. 1996. *Using Language*. Cambridge, Cambridge University Press. DOI: [10.1017/CBO9780511620539](https://doi.org/10.1017/CBO9780511620539). 227
- E. J. Coats, R. S. Feldman, and P. Philippot. 1999. The influence of television on children's nonverbal behavior. In P. Philippot, R. S. Feldman, and E. J. Coats, eds., *The social context of nonverbal behavior*, pp. 156–181. Cambridge University Press, Paris. 248
- W. Condon and W. Osgton. 1967. A segmentation of behavior. *Journal of Psychiatric Research*, 5: 221–235. 225
- M. Courgeon and C. Clavel. 2013. Marc: a framework that features emotion models for facial animation during human-computer interaction. *Journal on Multimodal User Interfaces*, 7(4): 311–319. DOI: [10.1007/s12193-013-0124-1](https://doi.org/10.1007/s12193-013-0124-1). 239, 241
- R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and D. Karpouzis. 2011. Issues in Data Labelling. In: R. Cowie, C. Pelachaud, P. Petta, editors, *Emotion-Oriented Systems, Cognitive Technologies*, pp. 215–244. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-15184-2\\_13](https://doi.org/10.1007/978-3-642-15184-2_13). 227
- C. Darwin. 1872. *The Expression of the Emotions in Man and Animals*. London, UK. John Murray publisher. 225

- E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. July-September 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3): 349–365. DOI: [10.1109/T-AFFC.2012.12](https://doi.org/10.1109/T-AFFC.2012.12). 223, 786
- U. Dimberg. 1997. Facial emg: Indicator of rapid emotional reactions. *International Journal of Psychophysiology*, 25(1): 52–53. <http://www.sciencedirect.com/science/article/pii/S0167876097854839>. DOI: [10.1016/S0167-8760\(97\)85483-9](https://doi.org/10.1016/S0167-8760(97)85483-9). 225, 247
- Y. Ding, K. Prepin, J. Huang, C. Pelachaud, and T. Artières. 2014. Laughter animation synthesis. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pp. 773–780. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. <http://dl.acm.org/citation.cfm?id=2615731.2615856>. 245
- E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In A. C. R. Paiva, R. Prada, R. W. Picard, editors, *Affective Computing and Intelligent Interaction*. ACII 2007. Lecture Notes in Computer Science, vol. 4738. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-540-74889-2\\_43](https://doi.org/10.1007/978-3-540-74889-2_43). 226, 227
- E. Douglas-Cowie, C. Cox, J. C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry, A. Batlinger, and F. Höning. 2011. The HUMAINE database. In R. Cowie, C. Pelachaud, P. Petta, editors, *Emotion-Oriented Systems. Cognitive Technologies*. Springer, Berlin, Heidelberg. 227
- S. Duncan and D. Fiske. 1985. *Interaction Structure and Strategy*. Cambridge University Press, New York, NY. 227
- P. Edelmann and R. B. Zajonc. 1989. Facial Efference and the Experience of Emotion. *Annual Review of Psychology*, 40: 249–280. DOI: [10.1146/annurev.ps.40.020189.001341](https://doi.org/10.1146/annurev.ps.40.020189.001341). 225
- P. Ekman. 1979. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pp. 169–248. Cambridge University Press, Cambridge, UK, New York. 227
- P. Ekman. 2003. *Emotions Revealed*. Times Books (US), New York. Weidenfeld & Nicolson (world), London. 227
- P. Ekman and W. V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1: 49–98. 219
- P. Ekman, W. Friesen, and J. Hager. 2002. *The Facial Action Coding System (2nd Edition)*. Research Nexus eBook, London, Weidenfeld & Nicolson (world), Salt Lake City, UT. 226
- C. Ennis and C. O'Sullivan. May 2012. Perceptually plausible formations for virtual conversers. *Computer Animation and Virtual Worlds*, 23(3–4): 321–329. DOI: [10.1002/cav.1453](https://doi.org/10.1002/cav.1453). 231

- C. Ennis, R. McDonnell, and C. O'Sullivan. July 2010. Seeing is believing: Body motion dominates in multisensory conversations. *ACM Transactions on Graphics*, 29(4): 91:1–91:9. DOI: [10.1145/1778765.1778828](https://doi.org/10.1145/1778765.1778828). 231
- S. Garrod and A. Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2): 181–218. DOI: [10.1016/0010-0277\(87\)90018-7](https://doi.org/10.1016/0010-0277(87)90018-7). 225
- R. Gifford. 1994. A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, 66(2): 398–412. DOI: [10.1037/0022-3514.66.2.398](https://doi.org/10.1037/0022-3514.66.2.398). 248
- H. Giles, N. Coupland, and I. Coupland. 1991. Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, and N. Coupland, editors, *Studies in Emotion and Social Interaction. Contexts of Accommodation: Developments in Applied Sociolinguistics*, pp. 1–68. New York, NY. Cambridge University Press, Paris, France. Editions de la Maison des Sciences de l'Homme. DOI: [10.1017/CBO9780511663673.001](https://doi.org/10.1017/CBO9780511663673.001). 225
- N. Glas and C. Pelachaud. 2015. Definitions of engagement in human-agent interaction. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 944–949. IEEE. DOI: [10.1109/ACII.2015.7344688](https://doi.org/10.1109/ACII.2015.7344688). 222, 779
- J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. 2007. Creating rapport with virtual agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Intelligent Virtual Agents*, pp. 125–138. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-540-74997-4\\_12](https://doi.org/10.1007/978-3-540-74997-4_12). 247
- U. GroBekathöfer, N.-C. Wöhler, T. Hermann, and S. Kopp. 2012. On-the-fly behavior coordination for interactive virtual agents: A model for learning, recognizing and reproducing hand-arm gestures online. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems—Volume 3*, AAMAS '12, pp. 1177–1178. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. 232, 245
- A. Hartholt, D. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch. 2013. *All Together Now. Introducing the Virtual Human Toolkit*, pp. 368–381. Springer, Berlin, Heidelberg. 234
- A. Heloir and M. Kipp. 2009. Embr—a realtime animation engine for interactive embodied agents. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, eds., *Intelligent Virtual Agents: 9th International Conference*, IVA 2009 Amsterdam, The Netherlands, September 14–16, 2009 Proceedings, pp. 393–404. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-04380-2\\_43](https://doi.org/10.1007/978-3-642-04380-2_43). 232
- D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. H. Vilhjálmsson. 2008. The next step towards a function markup language. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, IVA '08, pp. 270–280. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/978-3-540-85483-8\\_28](https://doi.org/10.1007/978-3-540-85483-8_28). 230

- A. B. Hostetter and M. W. Alibali. 2008. *Psychonomic Bulletin and Review*, 15: 495. DOI: [10.3758/PBR.15.3.495](https://doi.org/10.3758/PBR.15.3.495). 225
- H.-H. Huang, T. Nishida, A. Cerekovic, I. S. Pandzic, and Y. Nakano. 2008. The design of a generic framework for integrating eca components. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent System—Volume 1*, AAMAS '08, pp. 128–135. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. <http://dl.acm.org/citation.cfm?id=1402383.1402406>. 241
- J. Huang and C. Pelachaud. 2012. An efficient energy transfer inverse kinematics solution. In *International Conference on Motion in Games*, pp. 278–289. Springer. DOI: [10.1007/978-3-642-34710-8\\_26](https://doi.org/10.1007/978-3-642-34710-8_26). 244
- L. Huang, L.-P. Morency, and J. Gratch. 2011. Virtual rapport 2.0. In *International Workshop on Intelligent Virtual Agents*, pp. 68–79. Springer. DOI: [10.1007/978-3-642-23974-8\\_8.223](https://doi.org/10.1007/978-3-642-23974-8_8.223), 783
- K. Jokinen and C. Pelachaud. 2013. From Annotation to Multimodal Behavior. In M. Rojc, N. Campbell, editors, *Coverbal Synchrony in Human-Machine Interaction*. Boca Raton, CRC Press. DOI: [10.1201/b15477](https://doi.org/10.1201/b15477). 227
- A. Kendon. 2000. Language and gesture: Unity or duality. In D. McNeill, editor, *Language and Gesture*, number 2 in *Language, Culture and Cognition*, pp. 47–63. Cambridge University Press. 219, 224
- A. Kendon. 2002. Some uses of the head shake. *Gesture*, 2(2): 147–183. 223
- A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press. DOI: [10.1017/CBO9780511807572](https://doi.org/10.1017/CBO9780511807572). 226, 227
- M. Kipp. 2006. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, pp. 230–242. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/11821830\\_19](https://doi.org/10.1007/11821830_19). 245
- S. Kopp and I. Wachsmuth. March 2004. Synthesizing multimodal utterances for conversational agents: Research articles. *Computer Animation and Virtual Worlds*, 15(1): 39–52. DOI: [10.1002/cav.v15:1](https://doi.org/10.1002/cav.v15:1). 236
- S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: the behavior markup language. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, pp. 205–217. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/11821830\\_17](https://doi.org/10.1007/11821830_17). 229, 230, 233
- S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal of Multimodal User Interfaces*, 8(1): 97–108. <http://dblp.uni-trier.de/db/journals/jmui/jmui8.html#KoppWYB14>. DOI: [10.1007/s12193-013-0130-3](https://doi.org/10.1007/s12193-013-0130-3). 235

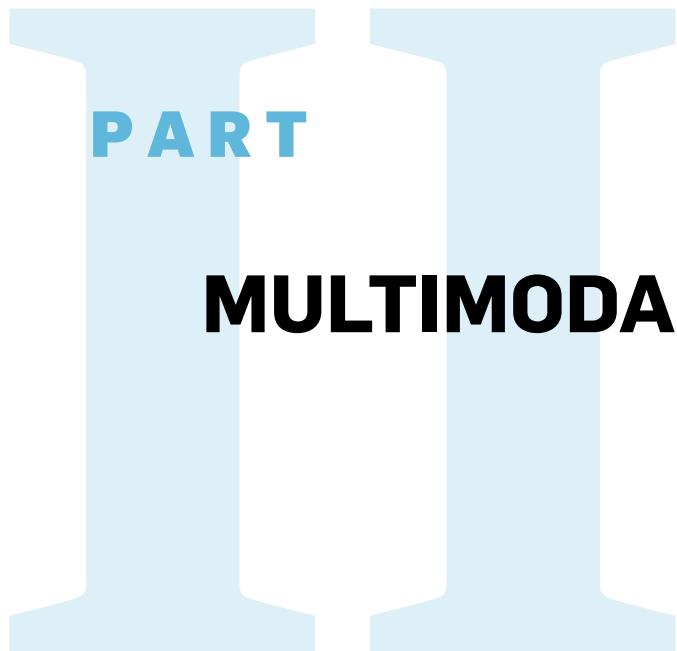
- J. L. Lakin, V. A. Jefferis, C. M. Cheng, and T. L. Chartrand. 2003. Chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Nonverbal Behavior*, 27(3): 145–162. DOI: [10.1023/A:1025389814290](https://doi.org/10.1023/A:1025389814290). 225
- J. Lee and S. Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, pp. 243–255. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/11821830\\_20](https://doi.org/10.1007/11821830_20). 231, 234
- M. Lhommet, Y. Xu, and S. Marsella. 2015. Cerebella: Automatic generation of nonverbal behavior for virtual humans. 237
- D. Luciew, J. Mulkern, and R. Punako Jr. 2011. Finding the Truth: Interview and Interrogation Training Simulations. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Paper No. 11186. 248
- M. Mancini and C. Pelachaud. 2008. The FML-APML language. In *Why Conversational Agents do What They Do. Workshop on Functional Representations for Generating Conversational Agents Behavior at AAMAS*. 233
- S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, pp. 25–35. ACM, New York. DOI: [10.1145/2485895.2485900](https://doi.org/10.1145/2485895.2485900). 232, 237
- S. C. Marsella, S. M. Carnicke, J. Gratch, A. Okhmatovskaia, and A. Rizzo. 2006. An exploration of Delsarte's structural acting system. In *Intelligent Virtual Agents*, pp. 80–92. Springer. DOI: [10.1007/11821830\\_7](https://doi.org/10.1007/11821830_7). 248
- D. Matsumoto. 2006. Culture and nonverbal behavior. In V. Manusov and M. L. Patterson editors,, *The Sage Handbook of Nonverbal Communication*, (pp. 219–235). Thousand Oaks, CA. Sage Publications, Inc. DOI: [10.4135/9781412976152.n12](https://doi.org/10.4135/9781412976152.n12). 224
- E. Z. McClave. June 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7): 855–878. <http://www.sciencedirect.com/science/article/pii/S037821669900079X>. DOI: [10.1016/S0378-2166\(99\)00079-X](https://doi.org/10.1016/S0378-2166(99)00079-X). 223
- D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago. 221, 222, 226, 227, 243, 763, 773
- A. Menache. 2011. 4 - The motion data. In A. Menache, editor, *Understanding Motion Capture for Computer Animation*. 2nd ed., The Morgan Kaufmann Series in Computer Graphics, pp. 135–185. Morgan Kaufmann, Boston. <http://www.sciencedirect.com/science/article/pii/B9780123814968000044>. DOI: [10.1016/B978-0-12-381496-8.00004-4](https://doi.org/10.1016/B978-0-12-381496-8.00004-4). 245
- J. Nadel and G. Butterworth. 1999. *Imitation in Infancy*, vol. 16. Cambridge University Press Cambridge. 222, 775
- J. Nadel and H. Tremblay-Leveau. 1999. Early perception of social contingencies and interpersonal intentionality: Dyadic and triadic paradigms. *Early social Cognition: Understanding Others in the First Months of Life*, pp. 189–212. 225

- R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud. 2009. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems—Volume 2*, AAMAS '09, pp. 1399–1400. International Foundation for Autonomous Agents and Multiagent Systems. DOI: [10.1145/1558109.1558314](https://doi.org/10.1145/1558109.1558314). [228](#)
- A. Ortony, G. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotions*. New York, NY. Cambridge University Press. [227](#)
- A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobrepeerez, S. Woods, C. Zoll, and L. Hall. 2004. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems—Volume 1*, pp. 194–201. IEEE Computer Society. [222](#), [770](#)
- F. Pecune, A. Cafaro, M. Chollet, P. Philippe, and C. Pelachaud. 2014. Suggestions for extending saiba with the vib platform. In *Workshop on Architectures and Standards for IVAs, held at the '14th International Conference on Intelligent Virtual Agents*, IVA 2014, p. 16–20. Bielefeld eCollections. [233](#)
- F. Pecune, B. Biancardi, Y. Ding, C. Pelachaud, M. Mancini, G. Varni, A. Camurri, and G. Volpe. 2015. Lol-laugh out loud. In *AAAI*, pp. 4309–4310. [232](#), [247](#)
- C. Pedica and H. H. Vilhjálmsson. 2012. Lifelike interactive characters with behavior trees for social territorial intelligence. In *ACM SIGGRAPH 2012 Posters*, SIGGRAPH '12, pp. 32:1–32:1. ACM, New York, NY. DOI: [10.1145/2342896.2342938](https://doi.org/10.1145/2342896.2342938). [231](#), [232](#), [240](#)
- M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02): 169–190. [222](#), [225](#), [761](#)
- I. Poggi. 2007. *Mind, Hands, Face and Body. A Goal and Belief View of Multimodal Communication*, volume Körper, Zeichen, Kultur. Weidler Verlag. [225](#), [227](#)
- K. Prepin, M. Ochs, and C. Pelachaud. 2013. Beyond backchannels: co-construction of dyadic stance by reciprocal reinforcement of smiles between virtual agents. In *Proceedings of CogSci Annual Meeting of the Cognitive Science Society*. [232](#)
- B. Ravenet, M. Ochs, and C. Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, editors, *Intelligent Virtual Agents: 13th International Conference*, IVA 2013, Edinburgh, UK, August 29–31, 2013 *Proceedings*, pp. 263–274. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-40415-3\\_23](https://doi.org/10.1007/978-3-642-40415-3_23). [231](#)
- M. Rehm and E. André. 2008. From annotated multimodal corpora to simulated human-like behaviors. In *Proceedings of the Embodied Communication in Humans and Machines, 2Nd ZiF Research Group International Conference on Modeling Communication with Robots and Virtual Humans*, ZiF'06, pp. 1–17. Springer-Verlag, Berlin, Heidelberg. <http://dl.acm.org/citation.cfm?id=1794517.1794518>. DOI: [10.1007/978-3-540-79037-2\\_1](https://doi.org/10.1007/978-3-540-79037-2_1). [230](#), [245](#)

- C. Reynolds. 1999. Steering behaviors for autonomous characters. In *Proceedings of the Game Developers Conference*, pp. 763–782. Miller Freeman Game Groups, San Francisco, CA. 240
- J. Robison, S. McQuiggan, and J. Lester. September 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6. DOI: [10.1109/ACII.2009.5349555](https://doi.org/10.1109/ACII.2009.5349555). 248
- R. Rojas. 1996. *Neural Networks—A Systematic Introduction*. Springer-Verlag, Berlin, New-York. 232
- B. Rossen and B. Lok. Apr. 2012. A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies*, 70(4): 301–319. DOI: [10.1016/j.ijhcs.2011.11.004](https://doi.org/10.1016/j.ijhcs.2011.11.004). 231, 239
- J. A. Russell and J. M. Fernadez-Dols. 1997. What does a facial expression mean. In J. A. Russell and J. M. Fernadez-Dols, editors, *The Psychology of Facial Expression*. Cambridge University Press, New York, NY. 223
- K. Scherer. 2000. Emotion. In M. Hewstone and W. Stroebe, editors, *Introduction to Social Psychology: A European Perspective*, pp. 151–191. Blackwell, Oxford, UK. 227
- K. R. Scherer and G. Ceschi. 1997. Lost luggage: a field study of emotion–antecedent appraisal. *Motivation and Emotion*, 21(3): 211–235. DOI: [10.1023/A:1024498629430](https://doi.org/10.1023/A:1024498629430). 226
- M. Schröder. January 2010. The semaine api: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction*. DOI: [10.1155/2010/319406](https://doi.org/10.1155/2010/319406). 239, 240
- M. Schröder, P. Baggio, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato. 2011. Emotionml—an upcoming standard for representing emotions and related states. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 316–325. Springer. DOI: [10.1007/978-3-642-24600-5\\_35](https://doi.org/10.1007/978-3-642-24600-5_35). 226
- A. Shoulson, N. Marshak, M. Kapadia, and N. I. Badler. 2013. Adapt: The agent development and prototyping testbed. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D ’13, pp. 9–18. ACM, New York, NY, USA. DOI: [10.1145/2448196.2448198](https://doi.org/10.1145/2448196.2448198). 239
- W. R. Swartout, J. Gratch, R. W. Hill Jr, E. Hovy, S. Marsella, J. Rickel, and D. Traum. 2006. Toward virtual humans. *AI Magazine*, 27(2): 96–108. 235
- M. ter Maat and D. Heylen. 2011. Flipper: An information state component for spoken dialogue systems. In H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, editors, *Intelligent Virtual Agents: 10th International Conference*, IVA 2011, Reykjavik, Iceland, September 15–17, 2011. *Proceedings*, pp. 470–472. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-23974-8\\_67](https://doi.org/10.1007/978-3-642-23974-8_67). 235
- M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. 2008. Smartbody: behavior realization for embodied conversational agents. In *Proceedings of the 7th International*

- Joint Conference on Autonomous Agents and Multiagent Systems—Volume 1, AAMAS '08*, pp. 151–158. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. <http://dl.acm.org/citation.cfm?id=1402383.1402409>. DOI: [10.1145/1402383.1402409](https://doi.org/10.1145/1402383.1402409). 235
- F. Thomas and O. Johnston. 1995. *The Illusion of Life: Disney Animation*. Hyperion, New York. 249
- L. Tickle-Degnen and R. Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4): 285–293. 225, 247
- N. Tinbergen and Perdeck. 1950. On the stimulus situation releasing the begging response in the newly hatched herring gull chick (*Larus argentatus argentatus* Pont). *Behavior*, 3: 1–39. DOI: [10.1163/156853951X00197](https://doi.org/10.1163/156853951X00197). 249
- B. Tversky and B. M. Hard. 2009. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110(1): pp. 124–129. 224
- R. B. Van Baaren, R. W. Holland, B. Steenaert, and A. van Knippenberg. 2003. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39(4): 393–398. DOI: [10.1016/S0022-1031\(03\)00014-3](https://doi.org/10.1016/S0022-1031(03)00014-3). 222, 779
- H. van Welbergen, D. Reidsma, Z. M. Ruttkay, and J. Zwiers. 2010. Elckerlyc - A BML realizer for continuous, multimodal interaction with a virtual human. *Journal on Multimodal User Interfaces*, 3(4): 271–284. DOI: [10.1007/s12193-010-0051-3](https://doi.org/10.1007/s12193-010-0051-3). 236
- H. van Welbergen, R. Yaghoubzadeh, and S. Kopp. 2014. *AsapRealizer 2.0: The Next Steps in Fluent Behavior Realization for ECAs*, pp. 449–462. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-09767-1\\_56](https://doi.org/10.1007/978-3-319-09767-1_56). 232, 235
- H. H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thórisson, H. Welbergen, and R. J. Werf. 2007. The behavior markup language: Recent developments and challenges. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, IVA '07, pp. 99–111. Springer-Verlag, Berlin, Heidelberg. DOI: [10.1007/978-3-540-74997-4\\_10](https://doi.org/10.1007/978-3-540-74997-4_10). 230, 233
- H. H. Vilhjálmsson. 2005. Augmenting online conversation through automated discourse tagging. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)—Track 4—Volume 04*, HICSS '05, pp. 109.1–. IEEE Computer Society, Washington, DC. DOI: [10.1109/HICSS.2005.109](https://doi.org/10.1109/HICSS.2005.109). 228
- H. H. Vilhjálmsson. 2009. Representing communicative function and behavior in multimodal communication. In A. Esposito, A. Hussain, M. Marinaro, and R. Martone, eds., *Multimodal Signals: Cognitive and Algorithmic Issues*, vol. 5398 of *Lecture Notes in Computer Science*, pp. 47–59. Springer, Berlin, Heidelberg. DOI: [10.1007/978-3-642-00525-1\\_4](https://doi.org/10.1007/978-3-642-00525-1_4). 229
- J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. 2013. The social signal interpretation (ssi) framework: multimodal signal processing and recognition

- in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia, MM '13*, pp. 831–834. ACM, New York, NY. DOI: [10.1145/2502081.2502223](https://doi.org/10.1145/2502081.2502223). 233, 245
- M. Walker and C. Trimboli. 1983. The expressive function of the eye flash. *Journal of Nonverbal Behavior*, 8(1): 3–13. 227
- G. L. Wells and R. E. Petty. 1980. The effects of over head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1(3): 219–230. DOI: [10.1207/s15324834basp0103\\_2](https://doi.org/10.1207/s15324834basp0103_2). 225
- Y. Xu, C. Pelachaud, and S. Marsella. Sept. 2014. Compound gesture generation: A model based on ideational units. In *Intelligent Virtual Agents*. [http://www.ccs.neu.edu/~marsella/publications/Yuyu\\_IVA2014.pdf](http://www.ccs.neu.edu/~marsella/publications/Yuyu_IVA2014.pdf). DOI: [10.1007/978-3-319-09767-1\\_58](https://doi.org/10.1007/978-3-319-09767-1_58). 237, 246
- R. Zhao, A. Papangelis, and J. Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*, pp. 514–527. Springer. DOI: [10.1007/978-3-319-09767-1\\_62](https://doi.org/10.1007/978-3-319-09767-1_62). 223, 783



**PART**

**MULTIMODAL BEHAVIOR**



# Ergonomics for the Design of Multimodal Interfaces

Alexis Heloir, Fabrizio Nunnari, Myroslav Bachynskyi

## 7.1

### Introduction

There are many ways a machine can infer a user intention or her/his cognitive and affective states: voice, voluntary movements, skin conductivity, eye movement, or muscle activation, to name a few. Voluntary movement, however, is still the privileged input channel for multimodal interfaces: it can be a button press, a mouse-mediated aimed movement, a direct touch on a screen, a mid-air gesture, or a full-body movement. The recent development of touch and motion-sensing technology broadens the interaction space by extending the number of input effectors: not only the fingertips but also the whole body now have the potential to support future input strategies. Indeed, touchscreens, inertial measurement units (IMU), as well as RGB, stereo, and time of flight (ToF) camera sensors will eventually become standard components of ubiquitous multimodal systems. However, as the time-to-market is continuously shrinking, it becomes more and more difficult for user experience (UX) designers to account for the usability and ergonomics of novel multimodal devices entering the market.

In the past, the task of interaction design was split into two fields: *industrial design*, dealing with the design and development of hardware and physical input artifacts together with their appropriate physical ergonomics assessment; and *human-computer interaction*, dealing with cognitive and information processing aspects of computer input in software as well as its usability in terms of effectiveness, efficiency, and satisfaction. Today, new interactive products hit the market at a pace which is too fast for applying traditional ergonomic assessment methods before launch phases. It is, therefore, the early adopters who practically assess the relevance and ergonomics of new consumer devices.

That said, device manufacturers and interaction designers would all benefit from generic guidelines obtained through principled ergonomics assessment conducted on post-desktop multimodal interfaces. Generic guidelines are indeed compatible with the tight timelines experienced in the industry. More precisely, application designers may largely benefit from biomechanical simulation when accounting for ergonomics early on during the design phase of consumer range products and applications. This chapter presents an overview of ergonomic studies in multimodal interfaces, starting with the main challenges posed by post-desktop interfaces and introducing the importance of ergonomic studies. Then, physical ergonomics is introduced. Next, the focus is placed on motion capture-based biomechanical simulation as a universal method for assessing multimodal interaction design. Finally, open questions and prospects are discussed.

## 7.2 The Generic Design Process

Over the last 40 years, HCI has developed a number of design methods, approaches, and processes, the most effective of which are documented as international standards [ISO/IEC 2009, ISO/IEC 1999, ISO/IEC 1998]. A more modern approach is user-centered design (UCD) [Norman 2013, Nielsen 1994, Abras et al. 2004].

Because it deeply involves end-users throughout the whole period of shaping and development of a product, UCD makes it possible to define and meet multiple design goals: match the user's conceptual model, knowledge, skills, and capabilities. UCD also guarantees relevant feedback, as well as robustness to errors and simplicity of use.

UCD ensures that a product can be used by the end-users to achieve their goals with effectiveness, efficiency, and satisfaction in the specified context of use, or in other words, UCD ensures good usability of the product. It is one of the most effective design processes, and thus it has become an international standard [ISO/IEC 1999, ISO/IEC 2009]. As can be observed in Figure 7.1, UCD consists of multiple iterations of four activities [ISO/IEC 1999, Norman 2013]: understand and specify the context of use, specify the user requirements, produce design solutions, and evaluate designs against requirements.

While UCD is a widely used and effective design process, it is far from being perfectly fitted to multimodal post-desktop interfaces and the pace of the tech industry. For example, the cost of prototypes and user studies can be high, the participants of user studies may poorly represent the user population, and the design process can take a considerable amount of time, increasing time-to-market and posing a

### Glossary

**Degrees of freedom (DOF)** of a mechanical system describes how a normalized mechanical junction can move. It usually defines a set of rotational axes together with rotation boundaries.

**Electromyography measurements (EMG)** provide quantitative data on real muscle recruitment. They were often used in physical ergonomics assessment of desktop interfaces, in particular mouse and keyboard. However, they are ineffective and inefficient for post-desktop interface evaluation, as they are limited to only close-to-surface muscles, and they suffer from cross-talk and muscle belly drift in dynamic movements, thus providing unreliable data.

**Ergonomics** is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance.

**Goniometer.** A goniometer is an instrument which measures an angle.

**Industrial design** deals with design and development of hardware and physical input artifacts, for example mouse, keyboard or joysticks, and their appropriate physical ergonomics assessment.

**Inertia matrices** describe inertial properties such as total mass and mass distribution of the rigid segments of the human body, and are involved in computation of joint moments for given kinematics through Newton's law.

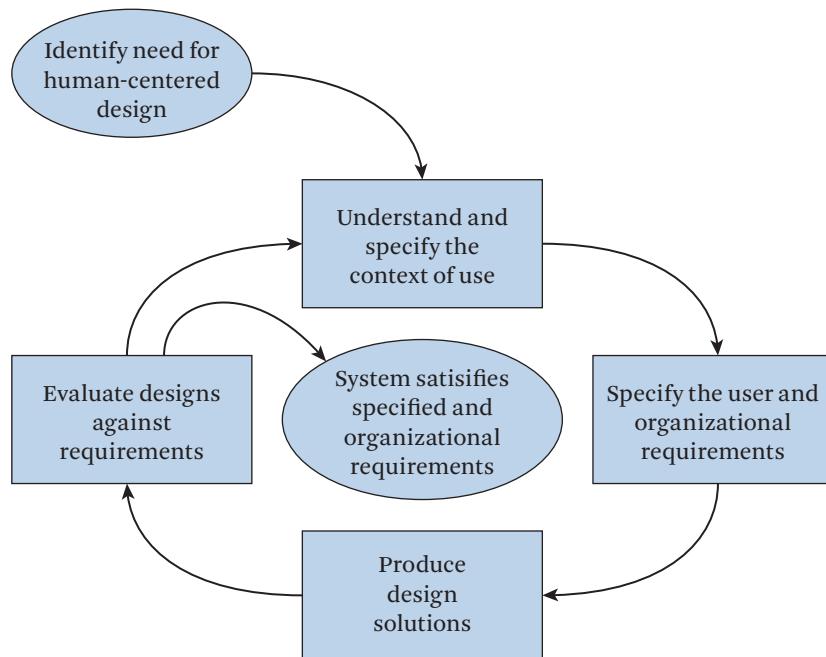
**Inverse dynamics** is a method for computing forces and/or moments of force (torques) based on the kinematics (motion) of a body and the body's inertial properties (mass and moment of inertia).

**Inverse kinematics** is an algorithm which resolves joint parameters for a skeletal structure given a specific set of constraints: for example, in the context of biomechanics, joint angles are resolved given kinematics of a set of markers attached to the body segments and recorded by the motion capture system.

**Physical ergonomics** is a subfield of ergonomics which considers human anatomy, physiology, and biomechanics in relation to physical activity.

**Repetitive strain injury (RSI)** is an “injury to the musculoskeletal and nervous systems that may be caused by repetitive tasks, forceful exertions, vibrations, mechanical compression, or sustained or awkward positions”: definition taken from [van Tulder et al. \[2007\]](#).

**Static optimization** is a method to resolve the joint moments computed by inverse dynamics to forces and activations of individual muscles.



**Figure 7.1** User-centered design process from ISO/IEC [1999].

risk of being overtaken by competitors. For this reason, considerable effort was invested by the research community into model-based interface design and development of generalized human models, which can provide cheaper, more accessible, and information-rich alternative to informing input method design [Paterno 2012, Duffy 2008, Puerta 1997].

Additionally, physical ergonomics problems cannot be discovered during a short design process or short user studies. Today, most consumer electronic products are sold to individual customers and used in environments as diverse as the couch, office/school desk, library, airplane, bed, or subway. Since most devices, when sold, do not ship with any user manual in their packaging anymore, the user is left alone with the responsibility to figure out on his own which postures and gestures will lower his chances to develop repetitive strain injury (RSI) in the distant future. It is paramount that potential effects of prolonged use of post-desktop multimodal input methods are studied and understood before industrial production and global public adoption.

## 7.3

### Physical Ergonomics: Data Collection

Physical ergonomics is a subfield of ergonomics which considers human anatomy, physiology, and biomechanics in relation to physical activity. Physical ergonomics principles and methods are widely used in industrial design, but not yet in the design of multimodal interfaces. In the following, we summarize existing physical ergonomics practices, tools, and methods used in research and design. We highlight possibilities and deficiencies of these methods for applications within HCI and in particular for post-desktop interface design.

Personal computers became an object of interest for a multitude of physical ergonomics works with their penetration into the work environment [Ming et al. 2004]. Each work covers either interaction with personal computers in general [Korhonen et al. 2003, Jensen et al. 2002, Blatter and Bongers 2002], or is focused on musculoskeletal health effects of interaction using a particular computer form-factor [Szeto and Lee 2002, Asundi et al. 2010, Moffet et al. 2002] or input device such as a keyboard [Gerr et al. 2006, Blangsted et al. 2004, Zecevic et al. 2000], mouse [Andersen et al. 2003, Aarås et al. 2002, Visser et al. 2004], or touchpad [Sommerich et al. 2002]. A large fraction are questionnaire-based and often cover an impressively large user population, from a few hundred [Schlossberg et al. 2004, Ortiz-Hernández et al. 2003], up to a few thousand [Andersen et al. 2003, Hakala et al. 2006]. Fewer works are based on user observations [Breen et al. 2007] and objective measurements using videometry [Moffet et al. 2002, Cook et al. 2004], goniometers [Moffet et al. 2002, Keir et al. 1999], EMG [Bjarne Laursen et al. 2002, Visser et al. 2004, Harvey and Peper 1997], grip and interaction forces [Feuerstein et al. 1997, Visser et al. 2004], and even pressure inside the carpal tunnel [Keir et al. 1999].

Decades of ergonomics research concerned with traditional input methods have resulted in the desktop workspace as we know it today, optimized according to the established ergonomic recommendations, which can significantly reduce risks of various musculoskeletal disorders. Traditional input methods were centered around physical artifacts, where the movement and design space were small enough for fast evaluation of usability within the field of physical ergonomics and the evaluation of performance within the field of HCI. It was possible to analyze the small design and movement space using a wide range of methods, as reported in tens of relevant user studies. However, with post-desktop input methods the situation radically changes: the movement space is huge and both performance and ergonomics have to be considered within the field of HCI; as a result, the traditional methods are inefficient and additionally there is a lack of ergonomics

expertise. Therefore, the amount of research in assessment of post-desktop input methods is much smaller than for traditional input devices, which is reflected in the reduced number of related works when moving from traditional input methods to touch-based input methods [Hinckley 2017] and further to mid-air input methods [Hornung et al. 2019]. In particular for mid-air input methods, this results in poor designs, a number of large industrial failures, and current stagnation in the field.

### 7.3.1 Discomfort Questionnaire Methods

Discomfort questionnaire-based methods provide rough qualitative data about the risks at the workplace and are the most straightforward to apply. It is generally accepted that discomfort at the workplace is the first warning sign of musculoskeletal injury. If the discomfort is ignored, after prolonged exposure it can lead to experience of pain caused by minor trauma. If further ignored, it can lead to serious musculoskeletal injury or disease, e.g., repetitive strain injury, carpal tunnel syndrome, and arthritis [Stanton et al. 2004]. As discomfort is a subjective experience of a user, it can only be assessed by asking the user about it, either in an interview or by filling out a questionnaire.

All questionnaires are based on previous work in the ergonomics field and reflect the state of a particular society's requirements for musculoskeletal safety at the workplace. The most prominent discomfort questionnaires are the Standardised Nordic Questionnaire [Kuorinka et al. 1987], PLIBEL [Kemmlert 1995], NIOSH [Bernard et al. 1993], and the Dutch Musculoskeletal Survey [Hildebrandt et al. 2001]. They consist of a set of questions systematically covering postures, movement types, and their temporal characteristics. The health risk areas are identified based on all responses, and are then investigated in detail by observation or measurement-based methods.

The questionnaires contain various types of questions: binary, categorical, and ordinal. While binary or categorical results are straightforward to interpret, ordinal results need special analysis and interpretation. Ordinal questions commonly assess perceived levels of exposure, discomfort, exertion, workload, stress, etc. on a rating scale. The most common in all areas are 5- or 7-level Likert scales [Likert 1932], and within ergonomics the Borg Ratings of Perceived Exertion (RPE) and Borg Category-Ratio 10 (CR10) scales [Borg 1998]. In contrast to purely ordinal Likert scales, Borg RPE and Borg CR10 provide mappings of verbal anchors to numerical values on a linear scale, allowing application of standard statistical methods.

While a variety of questionnaires are often applied in multimodal interaction studies, discomfort questionnaires are poorly suited in this context. The reason

is that these methods are oriented toward workers regularly exposed to the risk factors for a long enough period to develop a discomfort, while traditional studies are not prolonged enough, making the results unreliable. Additionally, even with long exposures the results are subjective, need a large number of participants to be statistically significant, and lack validity and reliability.

### 7.3.2 Posture Observation Methods

Posture observation-based methods allow expert ergonomists to gather objective qualitative data about the risks at the workplace without disturbing the workers or influencing their activities. These methods are based on the fact that the observable human posture reflects the musculoskeletal activity of the whole body. It is assumed that there exists a safe “neutral” posture, and that deviating from this safe posture may strain the musculoskeletal system proportionally to the angle, frequency, and time spent in the wrong posture. In contrast to discomfort questionnaires, this method allows possible risks to be identified even before the discomfort can be perceived, providing applicability in short-term studies for interface design. The assessment is performed by ergonomics experts either by direct observation of the workplace or by analysis of video recordings [Fallentin et al. 2001].

The most important and widely used methods from this category are Rapid Upper Limb Assessment (RULA) [McAtamney and Corlett 1993] and Rapid Entire Body Assessment (REBA) [Hignett and McAtamney 2000]. Similar to questionnaires, they systematically cover regions of interest (RULA) or the whole body (REBA) and allow ergonomics experts to quickly perform event-based assessments of users’ postures. Additionally, they support the ergonomists in rating the postures by visualizations of postural schemes. Furthermore, all posture segment ratings are summarized through table computations into a single risk assessment score.

Another method, the Strain Index (SI) [Moore and Garg 1995], includes as well as the assessment, possible postures of the arms below the elbow, also the observed exertion levels in each posture and temporal features of the activity, such as duration of exertion, percentage of task cycle in exertion, frequency of exertions, and the total time on the task per day. While part of the data necessary for SI is quantitative, it is still split into few categories and provides only qualitative results.

A number of other methods within this category were developed and used for ergonomical analyses of workplaces, for example the Quick Exposure Checklist [Li and Buckle 2000], Ovako Working posture Analysis System (OWAS) [Karhu et al. 1977], posture distribution-based RULA [Bao et al. 2007], Portable Ergonomics Observation (PEO) method [Fransson-Hall et al. 1995], OCRA Index ([Colombini

[etal. 2001](#)]), etc. However, they are similar to the above-described RULA, and REBA, or SI, use the same principles and assumptions, so we do not describe them here.

In contrast to discomfort questionnaires, the observation-based methods can be applied in multimodal interaction for injury risk assessment in research and in early stages of input method design. In this way, major design faults, in particular for gestural interfaces [[Hayward 2010](#)], can be avoided at relatively low cost. The downside is that application of these methods demand specific knowledge and skills in the field of biomechanics, which are commonly out of scope of the experimenters' expertise. Although the data avoids subjective variability of each individual worker, it is still subjective with respect to an ergonomics expert performing the analysis. As a result, these methods provide only imprecise qualitative results concerning presence or absence of the health risks and usually require following more detailed analyses to reduce the risk. While health risk assessment is of course important for success of input methods, the observation-based methods still lack the power to provide information about effort and fatigue, which are particularly important for HCI.

### 7.3.3 Direct Measurement Methods

Direct measurement-based methods are the most comprehensive, informative, and accurate, albeit the most costly in terms of time, equipment, and required competences. They provide rich quantitative data, measured directly in the human body, which describes most physiological and even some cognitive and emotional states and processes [[Zhou et al. 2018](#)]. Depending on the type of data recorded in the experiment, the direct measurement methods can be further split into three broad categories:

1. methods which consider *mechanical processes* inside the human body, for example computer vision (CVMC) [[Moeslund and Granum 2001](#), [Moeslund et al. 2006](#), [Cappelli and Duffy 2006](#), [Gagnon and Gagnon 1992](#)], electromagnetic (EMMC) [[EMTS 2016](#)], mechanical motion capture (MMC) [[MetaMotion 2016](#), [Kirchner et al. 2019](#)] and inertial motion unit motion capture (IMUMC) [[Kim and Nussbaum 2013](#)] human motion capture (HMC), electronic goniometry (EG) [[Costello et al. 1999](#), [Franko et al. 2008](#), [Dennerlein and Johnson 2006b](#)], hand kinematics recording by CyberGlove (CG) [[Chao et al. 2000](#), [Dipietro et al. 2008](#)], trunk kinematics recording with Lumbar Motion Monitor (LMM) [citeparenMarras199275](#), force recording within the musculoskeletal system with force transducers (FT) [[Schuind et al. 1992](#)], isometric [[Essendrop et al. 2001](#), [Coldwells et al. 1994](#)], isotonic

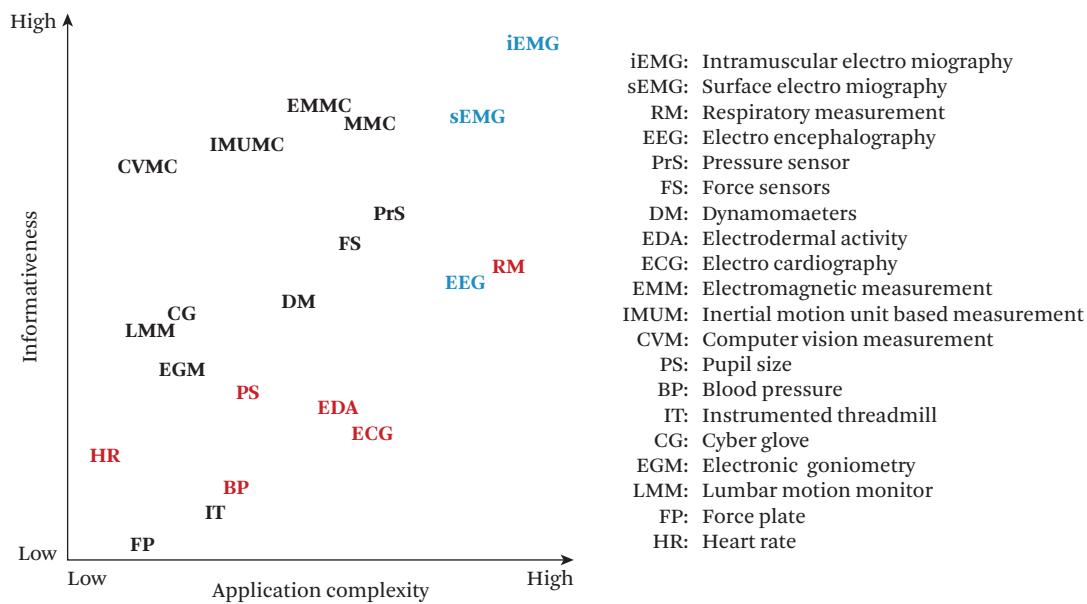
[Stauber et al. 2000] and isokinetic [Cabri 1991, Baltzopoulos and Brodie 1989] dynamometers (DM), as well as external force recording with force plates (FP) [Gagnon and Gagnon 1992], instrumented treadmills (IT) [Belli et al. 2001], force sensors (FS) [Dennerlein and Johnson 2006b, Radwin et al. 1992], and pressure sensors (PrS) [Ashruf 2002, Tan et al. 2001, Gurram et al. 1995];

2. methods which consider *electrical processes* inside the human body, for example surface (sEMG) [Kumar and Mital 1996, Dennerlein and Johnson 2006a, Gurram et al. 1995, Andreoni et al. 2002] and intramuscular electromyography (iEMG) [Zennaro et al. 2003, Thorn et al. 2002] electromyography (EMG), and electroencephalography (EEG) [Weber et al. 1980, Jap et al. 2009]; and
3. methods which consider *physiological processes* inside the human body and their effects, for example electrodermal activity (EDA) [Boucsein 2012, Boucsein and Thum 1997, Kuhmann 1989], electrocardiography (ECG) [Egelund 1982, Weber et al. 1980, Melamed et al. 1989], pupil size (PS) [Pomplun and Sunkara 2003, Iqbal et al. 2004], blood pressure (BP) [Hong et al. 2000, Hjortskov et al. 2004], heart rate (HR) [Sayers 1973, Hong et al. 2000, Hjortskov et al. 2004], respiratory measurement (RM) [Macfarlane 2001, Petrofsky and Lind 1978, Poole et al. 1988], etc.

In contrast to questionnaires and observation-based methods, most of the direct measurement methods cannot be applied directly at a workplace, and necessitate special equipment and often also a laboratory setting. The complexity of experimental data collection depends on a particular method and can be relatively low for some methods and very high for the others, as we summarize in Figure 7.2. Multiple methods are also very invasive, which restricts possible application scenarios and could lead to unnatural behaviors during the experiment. We give more details on the most important methods and their applicability in the paragraphs below.

The first category of methods concerns physical ergonomics most directly, as any mechanical activity or body movement is captured in this category. The mechanical methods can be split further into 3 categories, which complement each other in detailed analysis: kinematics, internal forces, and external forces measurements.

Kinematics measurements include positions, velocities, and accelerations of given points on the human body (CVMC, EMMC, IMUMC) or angular equivalents at given skeletal joints (MMC, EGM, CG, LMM), and is performed on a small segment of the body (EGM, CG, LMM) or on the whole body (CVMC, MMC, IMUMC). All



**Figure 7.2** Summary of direct measurement methods' informativeness vs. application complexity. Informativeness takes into account type of data, its detail, and scope with respect to the whole human body. Application complexity takes into account complexity and time to set up a measurement system, complexity of individual subject preparations, and level of expertise required from experimenters. The values were subjectively assessed through pairwise comparisons of each method's combination with respect to each variable. Black is used for mechanical, blue for electrical, and red for physiological methods.

measurements are non-invasive to the human body and some of them are also non-intrusive to human activity. For example, as can be seen in Figure 7.3, CVMC necessitates only wearing a special skin-tight suit [PhaseSpace 2016], or in the case of markerless motion capture, even specifies no other requirements besides keeping a line of sight between the user and cameras [Stoll et al. 2011, Sridhar et al. 2013]. Motion capture data is used for analyses of postures and human movement over the whole activity duration. All necessary analyses can be performed on it without manual frame-by-frame data inspection. For deeper insights about processes inside the human body, motion capture data can be used as an input to biomechanical simulation.

Kinematic data alone does not provide any insights about actual muscle and skeleton tissue loads inside the human body, which are essential for ergonomics



**Figure 7.3** Optical motion capture and external force recording during mid-air gestural interaction.

assessment. That is why kinematic data is usually complemented by external force measurement, and sometimes also internal force measurement. Internal muscle and skeleton tissue loads are defined by kinematics of movement considered in the context of inertial properties of the body, as well as external forces acting on the body, such as gravity, ground reaction force, chair reaction force, object weight, reaction force, etc. External forces are measured between contact points of the human body and the external world or object by introducing an intermediate force-sensing layer. For ground reaction forces, this layer is represented by force plates or an instrumented treadmill; for other external forces, specific force or pressure sensors are installed. In some studies, external forces are considered as a stand-alone data source for ergonomics analysis, for example when looking at grasp force during mouse interaction [Johnson et al. 2000], but more often they are considered jointly with motion capture data and in the context of biomechanical simulation.

In contrast to external forces, internal forces are much harder and more invasive to measure. Measurement of internal forces within an activity of interest is possible only by inserting special force transducers into corresponding tendons or muscles. Internal muscle forces are sometimes estimated in an additional experiment from

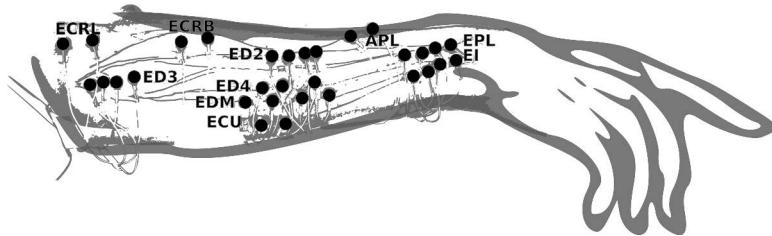
measurements by isometric, isotonic, and isokinetic dynamometers in order to tune muscle parameters in a musculoskeletal model. This allows more accurate simulation of an activity of interest.

In general, the group of mechanical process methods provide the best application complexity vs. informativeness trade-off for physical ergonomics assessment. As can be seen in Figure 7.2, most methods from this group (black abbreviations) are in the left central and upper segments of the chart, which corresponds to low to average application complexity and average or higher informativeness.

The second category of methods analyzes the electrical signals running through the human body. These processes belong to deeper levels in the system than mechanical processes and can be understood as control signals from the central nervous system (CNS) to the mechanical plant of the musculoskeletal system. The methods of this group consider electrical signals at each muscle (sEMG and iEMG), or more centrally, activations of various brain areas (EEG).

The human neural system transmits action potentials from the brain motor cortex to muscles similar to an electrical current. The action potential arrives to muscle at a motor end-plate, and then propagates along muscle fibers forcing them to contract and as a result generate mechanical force between opposite ends of fibers. EMG measures the difference in electrical potentials between two points along the muscle fiber located on the same side from the motor end-plate using either a pair of electrodes inserted into the muscle next to muscle fibers of interest (iEMG), or a pair of electrodes attached to the skin surface over the muscle of interest (sEMG). The force depends both on amplitude of the action potential and frequency, so there is a linear relationship between the linear envelope of EMG and force exerted by a muscle. To estimate muscle force from EMG an additional measurement is necessary: in parallel with EMG recording an isometric, isotonic, or isokinetic dynamometer is applied during a similar type of muscle contraction. This makes it possible to approximate the relationship between the EMG envelope value and force output, and use the relationship further for estimation of force. A second additional measure of EMG is during maximum voluntary contraction (MVC), which is used as a basis for normalization between experiment sessions or participants. EMG can also be used to estimate fatigue, which is reflected in a frequency shift of the power spectrum.

While EMG methods are very informative, their complexity, limitations, and invasiveness prevent wide adoption in ergonomics and multimodal interface design. They demand special expertise from the experimenters, in particular accurate muscle identification. In particular, iEMG necessitates precise needle insertion, is very invasive, and cannot be applied outside a clinical setting. Additionally, needles can



**Figure 7.4** Electrode setup for EMG of ten forearm muscles landmark data based on [Leijnse et al. \[2008\]](#).

cause discomfort in muscles during dynamic movements and prevent natural behavior. sEMG is not invasive, but it is limited to close-to-the-surface muscles, and is still intrusive to activities and naturalness of human behavior due to the cables. sEMG data is also not very reliable, as it suffers from muscular cross-talk, or muscle drift during dynamic contractions. It contains a large amount of variability between participants and experiments due to natural day-to-day variation in skin conductivity. Simultaneous recording of a large number of muscles makes both surface and intramuscular EMG setup very cumbersome, complex, and time consuming, and impacts naturalness of movements; for example, Figure 7.4 displays the complexity of the setup for 10 muscles out of the 630 present in the human body.

EEG records electrical activity of the brain sensed by electrodes attached to the scalp. Unlike EMG, it cannot provide a fine-grained signal related to activation of a particular muscle, and gives only a high-level summary of brain area activity. In ergonomics, it is used as an aggregated signal to detect general fatigue and sleepiness [[Weber et al. 1980](#), [Jap et al. 2009](#)]. In multimodal interface research it is used as an input method in brain-computer interfaces [[Choi and Cichocki 2008](#)] rather than as an ergonomics measurement instrument.

The third category of methods considers physiological processes in the human body, such as heart rate, blood pressure, skin conductivity, pupil size, respiratory measurement, oxygen intake, etc. These processes reflect whole-body aggregated variables and are used for assessment of energy expenditure. They require special equipment and can be applied in a laboratory setting, but they do not demand extensive specialized expertise from the experimenters. For multimodal input systems, these methods can be used for general evaluation of an input method, but not for detailed analysis to inform design. With respect to informativeness these methods are better than questionnaires or observations as they provide objective quantitative data, but worse than other direct measurement methods as

they do not provide enough details, despite having almost the same application complexity.

## 7.4

### Physical Ergonomics: Experimental Models

Ergonomic experiments and empirical studies provide a large amount of data describing human activity from a variety of possible perspectives. To make sense of such a large amount of data, dedicated mathematical and statistical models are applied. A number of such models have been developed for physical ergonomics and human factors. A large fraction of existing models directly correspond to—and were developed in tight coupling with—a specific data collection method, mostly questionnaires and observation-based methods.

For example, within the REBA anatomical scheme [[Hignett and McAtamney 2000](#)], the experimenter observes and rates the occurrence of events related to the postures of six body segments. Then, the assigned ratings are used to compute two composite ratings for body posture (legs, trunk, and neck) and the arm's posture (shoulder, elbow, and wrist) using Table 7.1 (top and middle). Further, the two composite ratings are combined into a single grand rating using Table 7.1 (bottom). The grand rating is then interpreted into severity of risks and priority of action on workplace improvement. Most other questionnaires and observation-based methods use similar models, encoded as scoring tables (QEC, RULA) or as a linear formula (DMS, SI, OCRA Risk Index) for convenience.

More interesting and complex models were developed for direct measurement data. They can be classified into six groups, although the boundaries between them are not always sharp: direct health risk estimation models; exposure-effect models; anthropometric posture prediction models; posture-based skeletal load prediction models; models which predict physiologic measures; and models of muscular fatigue and recovery.

#### 7.4.1 Direct Health Risk Estimation Models

The first group, health risk estimation models, is the broadest and, similar to the questionnaire and observation-based models, they take into account all collected data and directly output health risks. These models provide either high-level generic results, for example the model of overexertion [[Kumar 1994](#)], or alternatively consider a small part of the human body and provide a more detailed result for it, for example a risk model of carpal tunnel syndrome [[Tanaka and McGlothlin 1993](#), [Liu et al. 2003](#)].

**Table 7.1** REBA scoring tables [Stanton et al. 2004].

Body Posture Scoring													
Neck / Legs													
Trunk	1				2				3				
	1	2	3	4	1	2	3	4	1	2	3	4	
1	1	2	3	4	1	2	3	4	3	3	5	6	
2	2	3	4	5	3	4	5	6	4	5	6	7	
3	2	4	5	6	4	5	6	7	5	6	7	8	
4	3	5	6	7	5	6	7	8	6	7	8	9	
5	4	6	7	8	6	7	8	9	7	8	9	9	

Arms Posture Scoring														
Lower Arm / Wrist														
Shoulder	1			2			3			1			2	
	1	2	3	1	2	3	1	2	3	1	2	3	2	3
1	1	2	2	2	3	3	1	2	3	1	2	3		
2	1	2	2	3	3	3	2	3	3	2	3	4		
3	3	4	4	5	5	5	4	5	5	4	5	5		
4	4	5	5	5	5	5	5	6	6	5	6	7		
5	6	7	7	8	8	8	7	8	8	7	8	8		
6	7	8	8	8	8	8	8	9	9	8	9	9		

Grand Scoring													
Arms Posture Score													
Body posture score	1	2	3	4	5	6	7	8	9	10	11	12	
1	1	1	1	2	3	3	4	5	6	7	7	7	
2	1	2	2	3	4	4	5	6	6	7	7	8	
3	2	3	3	3	4	5	6	7	7	8	8	8	
4	3	4	4	4	5	6	7	8	8	9	9	9	
5	4	4	4	5	6	7	8	8	9	9	9	9	
6	6	6	6	7	8	8	9	9	10	10	10	10	
7	7	7	7	8	9	9	9	10	10	10	11	11	
8	8	8	8	9	10	10	10	10	10	10	11	11	
9	9	9	9	10	10	10	11	11	11	11	12	12	
10	10	10	10	11	11	11	11	12	12	12	12	12	
11	11	11	11	11	12	12	12	12	12	12	12	12	
12	12	12	12	12	12	12	12	12	12	12	12	12	

Other models were developed for assessment of lower back risks [Marras et al. 1993, Marras et al. 1995, Marras et al. 2000, Waters et al. 1993, Waters et al. 1994], neck and upper extremity risks [Armstrong et al. 1993, Korhonen et al. 2003], arm and hand risks [Sperling et al. 1993], carpal tunnel syndrome risk [Tanaka and McGlothlin 1993, Liu et al. 2003], overexertion [Kumar 1994], etc. In most cases, models are inferred using a regression fit of a low-degree polynomial relating risk levels to a list of independent variables present in the data, which, besides other variables, usually includes force level and duration of the exertion. The drawback of such models is that they do not provide any information on internal loads imposing the risk, and in this way break the logical chain of load propagation, reduce interpretability, and limit detailed analysis.

#### **7.4.2 Exposure-effect Models**

The second group of models generalizes the models from the first group and describe the load propagation chain from external loads, through internal loads to acute and chronic effects. These models provide a conceptual framework for inclusive ergonomics risk assessment in the presence of data from various sources; in particular they not only allow an assessment of health risks from directly collected external loads data, for some body segment, but also perform a whole-body assessment based on internal physiological loads [Winkel and Mathiassen 1994, Rohmert 1984]. These models need physiological loads as input, so they can be applied only after the internal loads are measured or computed.

#### **7.4.3 Anthropometric Posture Prediction Models**

The third group of models estimates probable user postures for a variety of workplace setups [Das and Behara 1995, Jung et al. 1995, Pheasant and Haslegrave 2005], or physical properties of designed artifacts [Garneau and Parkinson 2008, Endo et al. 2008] based on anthropometric data of user population. Instead of user studies, they can be used on early stages of design as inputs to the models which predict internal loads based on posture.

#### **7.4.4 Posture-based Skeletal Load Prediction Models**

The fourth group of models predicts internal musculoskeletal loads based on measured user postures and external forces. These models use an inverse approach for computation of internal loads, in which they consider the whole kinematic chain, its inertial properties, and applied external loads as inputs. The models belonging to this group range in complexity from simple link-segment models computing rough joint angles and moments in two dimensions [de Looze et al. 1992, Corlett

et al. 2003, Kingma et al. 1996b, Kingma et al. 1996a] to high-fidelity, full-body digital human models [Freivalds et al. 1984, Kromodihardjo and Mital 1986, Kingma et al. 1996a] able to compute joint angles, forces inside joints, and moments at the joints. However, biomechanical models currently used in ergonomic computations do not predict muscular loads and activations by the neural system, which are important for the analysis of ergonomics, energy expenditure, and fatigue.

#### **7.4.5 Models Predicting Physiologic Measures**

The fifth group of models predicts internal physiologic loads based on non-postural inputs. The most important and widely used inputs for physiological loads estimation are EMG signals [Hagberg and Hagberg 1989, Mathiassen and Winkel 1990, Gerdle et al. 1998, Mathiassen et al. 1995]. They enable the computation of the actual forces exerted by muscles and the generated joint moments. Although these models can provide deep insights based on the physically measured data of muscle activity, they are also constrained by limitations inherent in both intramuscular and sEMG data collection: low reliability for dynamic movements, limited number of accessible muscles or excessive invasiveness, high between-session and between-subject variability, and high complexity and cost of experiments. Within this group there are also models which consider other types of inputs, for example the mass-spring-damper model of grip [Lin et al. 2001], pendulum model of walking [Holt et al. 1990], model of force distribution in the shoulder [Karlsson and Peterson 1992], and more. However, they provide less accurate and insightful results than EMG-based models or the posture-based models from the fourth group.

#### **7.4.6 Models Predicting Muscular Fatigue**

The sixth group of models predicts muscular fatigue, recovery, and endurance during physical activity. They take as input the level of force exertion, its duration and repetitiveness, the fraction of exerted force to the maximum voluntary contraction force, and sometimes also the percentage of fast-twitch and slow-twitch muscle fibers, and compute the fatigue-recovery state of a muscle and its endurance in the context of a particular power output. Multiple models are developed within this group, for example the muscle fatigue-recovery model [Rodgers 1992, Rohmert 1973b, Rohmert 1973a], work-rest model of discomfort and endurance [Dul et al. 1994], critical power and power-time to exhaustion model [Morton 1996], etc. These models complement the models predicting internal physiological loads in the physical ergonomics assessment by quantifying the muscle state and potential work-recovery cycle and duration.

Often, data collection and analyses in ergonomics are not limited to a single type of data or a particular model, and combine results of a combination of methods which support and complement each other. In particular, this is common for modern ergonomics assessment software and digital human simulations, which often require the acquisition of the user's movement in 3D using motion capture.

## 7.5

### **Motion Capture-based Biomechanical Simulation**

The field of ergonomics has a large number of sophisticated methods for physical ergonomics assessment in a workplace and in controlled experiments. Multiple types of data can be collected to describe the physical ergonomics, and multiple models can be used on top of the data to expand and interpret it. However, most data collection types are not suitable for design, and in particular for post-desktop input methods, either because of unreliability, or because of application limits, due to being too intrusive or needing too much expertise. Optical motion capture using a marker suit and external force recording provide the best trade-off between application complexity and information quality provided by the data. However, all models previously used in ergonomics can only poorly interpret the data.

Motion capture-based biomechanical simulation is an experimental computational method developed in the field of biomechanics and widely used in the fields of medicine, rehabilitation, and sports research. It makes it possible to precisely analyze natural human movements and corresponding mechanical processes inside the human body, for example to assess movement deficiency sources, musculoskeletal risk factors in sports, and outcomes of potential surgery. The method principally consists of four parts:

- measurement of the necessary musculoskeletal properties of a particular patient or athlete, to adjust a generalized musculoskeletal model to match the body of the specific subject;
- a user study accurately recording the movement of interest of the participant with a motion capture system;
- application of the biomechanical simulation pipeline to produce joint angles and moments, muscle forces, activations, and excitations developed within the movement;
- analysis of the simulation outputs, their comparison to “normal” ranges and patterns, and identification of problematic spots.

The biomechanics community takes special care to minimize effects of measurement and modeling errors during all steps and has developed corresponding

practices and recommendations. We describe them and their possible application within HCI while staying realistic about HCI goals, expertise, resources, and experimental settings.

### 7.5.1 Optical Motion Capture

The key experimental input for biomechanical simulation and analysis is motion capture data describing human movements. This data consists of sequences of the 3D spatial coordinates of markers attached to the human body or angular coordinates of joints within the body. It can be recorded by a variety of methods which can be grouped into five categories: marker-based computer vision (Vicon [[Vicon 2016](#)], OptiTrack [[OptiTrack 2016b](#)], PhaseSpace [[PhaseSpace 2016](#)], Qualisys [[Qualisys 2016](#)], Metria Innovation [[Metria 2016](#)]), markerless and depth-based computer vision (The Captury [[Captury 2016](#)], Microsoft Kinect [[Microsoft 2016](#)], Apple PrimeSense), electromagnetic (Polhemius [[EMTS 2016](#)]), mechanical (MetaMotion Gypsy [[MetaMotion 2016](#)]), and inertial measurement unit (IMU)-based (Xsens [[Xsens 2016](#)]).

Each motion capture category has its own advantages and limitations which affect its applicability to biomechanical simulation within multimodal interfaces. Mechanical and IMU systems are the cheapest and they do not limit tracking volume. However, they do not provide precision suitable for biomechanical simulation. Additionally, they are cumbersome to apply and calibrate. When applied in ideal lab conditions, electromagnetic systems are exempt from electromagnetic perturbations and might thus be precise enough for biomechanical simulation studies involving ample limb movements.

Markerless computer vision and depth-based systems are the cheapest, easiest-to-apply solutions, and thus have a large potential in multimodal interfaces. However, at the current stage of development they provide limited accuracy, poorly suited to biomechanical simulation. The most widely adopted in conjunction with biomechanical simulation are the marker-based computer vision systems, as they provide sub-millimeter precision and reliable data. Their main limitation, however, is that each marker can be tracked only if it is within the field of view of at least three cameras, which might be an issue if the task accomplished by the recorded subject implies frequent visual occlusions.

That said, optical motion capture technology has not only matured, but also become significantly cheaper. At the time of writing (last quarter 2016), it is possible to order and set up a fully-functional optical motion capture system with covered volume of  $3 \times 3 \times 2.5$  with sub-millimeter tracking precision for as low as

\$20,000 [OptiTrack 2016a]. This allows for wide adoption of motion capture systems and opens the possibility for more applications. In particular, it becomes more accessible for laboratories, both in academia and industry.

### 7.5.2 Musculoskeletal Models

At the core of any biomechanical simulation is a model of human musculoskeletal system. This model describes the human body as a mechanical multi-body system, consisting of passive elements constraining a movement (bones and joints) and active elements which generate forces and energy for movement (muscles) based on a particular control input (neural signals). Musculoskeletal systems can be described according to three levels of abstraction:

**Kinematic** aspects describe rigid skeletal geometry and osteo- or even arthro-kinematics of joints between the rigid segments. In particular, high-fidelity musculoskeletal models can describe both 3D translation and 3D rotation components of a 1 DoF joint such as the knee, while low-fidelity models describe such a joint as a simple hinge with 1 axis of rotation. These aspects allow estimation of joint angles and the whole posture matching motion capture data.

**Dynamic** aspects describe inertial properties of each skeletal segment, like mass and the inertia matrix. Most biomechanical models describe each skeletal segment as completely rigid with constant inertia, while in real life soft tissues have a significant effect on resulting joint moments and forces and corresponding wobbling mass models are promising to reach more realism [Gruber et al. 1998]. The dynamic aspects are necessary for estimation of joint moments for a particular kinematics.

**Muscular** aspects describe active force generation within the human body by musculo-tendon units. The musculo-tendon units in musculoskeletal models are commonly represented by Hill-type models [Hill 1938, Martins et al. 1998] consisting of three components: the active element, serial, and parallel elastic elements. Such models are described by a force-length-velocity relationship (active), tendon stiffness, and force-length relationship (passive). Current models differ mostly in how the above-mentioned relationships are numerically described, which directly affects feasibility, efficiency, and accuracy of computations. Furthermore, musculoskeletal models differ in quality of muscle model parameters of each muscle; in some cases the parameters are derived from cadavers, while in other cases they are derived from healthy adults or sports students.

Mathematically, a biomechanical simulation model can be interpreted as a set of non-linear differential equations. Biomechanical simulation fits this model to a user and to the motion capture data of particular movements in order to output

joint angles. Then, by deeper interpretation, the simulation fits joint moments and forces and finally muscle forces and activations. Development of biomechanical models started from simple one-joint kinematic models, adding inertial properties toward dynamic models and adding muscle properties towards musculoskeletal models. All simple models were developed using direct-measurement experimental data from cadavers, X-rays, and joint moment or EMG measurements of regular humans. The most complex models commonly integrate parameters of multiple simpler models, sometimes with adjustments if the source models are based on different types of subject populations (cadavers, students, adults, etc.), and compare the final high-level outputs with corresponding data. While these models can be considered “Frankenstein-like,” they can provide realistic results [[Bachynskyi et al. 2014](#)]. The research effort in the biomechanics community has been unevenly distributed between studying lower extremities, upper extremities, and trunk, resulting in uneven quality of full-body musculoskeletal models. While current lower body models describe the human body in impressive detail, upper extremity models are less developed, and trunk models are the weakest part.

### 7.5.3 Biomechanical Simulation

One goal of biomechanical simulation is to infer neuromuscular activation signals from the captured kinematics of a human. To achieve this, a set of algorithms use a musculoskeletal model as a prior to fit the captured motion together with the measured external forces. Once fitted, the system outputs a sequence of joint moments, muscle forces, and activations for the analyzed movement. While fitting a musculoskeletal model to the captured data, the following steps apply.

1. **Model Scaling** adjusts a generic musculoskeletal model to match the anthropometric parameters of a particular person: size of his skeletal segments, total body mass and mass distribution, and musculotendon properties. This step outputs a model which is faithful to the proportions of the studied subject, his/her weight, mass distribution, and musculotendon parameters.
2. **Marker Adjustment** is performed after the model scaling to improve spatial correspondence between the markers attached to a model and to a participant. To improve the placement of the model markers with respect to the ones attached to a participant, modified inverse kinematics is performed on one averaged frame of static posture data with known minimal drift of the markers close to anatomical landmarks, which are considered by the

algorithm as fixed, and shifting other markers locations to match their correspondences in the data [OpenSim 2016]. After this step, the markers on the model and markers on the participant are in close correspondence.

3. **Inverse Kinematics** fits the posture of the musculoskeletal model to match the recorded motion capture data frame by frame. This algorithm corresponds to an optimization problem subject to kinematic constraints of the musculoskeletal model and minimizing an energy function of total squared error between virtual and physical markers and, if known, between externally computed and model generalized coordinates, by adjusting model generalized coordinates as parameters:

$$\min_q \left[ \sum_{i \in \text{markers}} w_i \left( \vec{x}_i^{\text{data}} - \vec{x}_i^{\text{model}}(q^{\text{model}}) \right)^2 + \sum_{j \in \text{coordinates}} w_j \left( q_j^{\text{data}} - q_j^{\text{model}} \right)^2 \right],$$

where  $q$  denotes generalized coordinates,  $x$  is the marker 3D locations, and  $w_i$  and  $w_j$ , are the weights of a particular marker or a particular joint coordinate. If the simulation inputs contain solely marker data, the second part of the energy function vanishes. This step outputs sequences of model-generalized coordinates closely matching the motion capture data within each frame [Lu and O'Connor 1999, Delp et al. 2007, OpenSim 2016].

4. **Inverse Dynamics** computes total joint moments and forces emerging within the movements described by the kinematics data. According to Newton's second law, point acceleration is directly proportional to the sum of forces acting on it and inversely proportional to its mass. Assuming that the human body consists of rigid segments of a particular mass and inertia, and measuring external forces acting on it, it is straightforward to apply the laws of classical mechanics, separate known forces and rearrange equation components to derive the forces and moments inside the human body:

$$\tau = M(q)\ddot{q} + C(q, \dot{q}) + G(q),$$

where  $q$ ,  $\dot{q}$ , and  $\ddot{q} \in R^N$  are the vectors of generalized coordinates, their velocities, and accelerations,  $M(q) \in R^{N \times N}$  is the skeletons mass matrix,  $C(q, \dot{q}) \in R^N$  is the vector of the Coriolis and centrifugal forces,  $G(q) \in R^N$  is the gravity vector, and  $\tau \in R^N$  is the vector representing all generalized forces, namely, total joint moments and total forces for translational joints [OpenSim 2016]. All components on the right side of the equation are known from measure-

ments of a previous simulation step resulting in direct and straightforward computation. This step outputs joint moments and forces that must be applied at each DoF of the skeleton in each frame to produce the specified kinematics.

5. **Static Optimization** computes the forces applied by each muscle and the corresponding activations necessary to produce the total joint moments computed for each frame by the inverse dynamics. It resolves muscle redundancy based on the assumption that humans recruit muscles in an efficient manner. The muscle recruitment strategy can be approximated using an appropriate objective function. Multiple objective functions have been studied in the past and many optimality criteria were proposed, for example total muscle force, total squared muscle stress, total squared muscle activation, mechanically based metabolic energy, biochemical muscle energy consumption, etc. [Prilutsky and Zatsiorsky 2002, Tsirakos et al. 1997, Praagman et al. 2006]. It has been shown that squared muscle activation, while being a simple and computationally cheap objective function, provides high correlation (0.85 in Praagman et al. [2006]) between predicted muscle cost and recorded metabolic cost ( $O_2$  consumption); thus, it is the most widely used objective function in biomechanical simulation. The problem is formulated as the minimization of an objective function:

$$J = \sum_{m \in \text{muscles}} (a_m)^2$$

subject to a set of constraints describing the muscle force-length-velocity physiological relationship and relating muscle forces with total joint moments:

$$\sum_{m \in \text{muscles}} [a_m f(F_m^0, l_m, v_m)] r_{m,j} = \tau_j,$$

where  $a_m$  is the activation of a muscle,  $F_m^0$  is the muscle's maximum isometric force,  $l_m$  is the muscle's fiber length,  $v_m$  the muscle fiber shortening velocity,  $f(F_m^0, l_m, v_m)$  the force-length-velocity surface of a muscle,  $r_{m,j}$  the moment arm of a muscle at a particular joint, and  $\tau_j$  the total moment at the joint [OpenSim 2016]. While the static optimization performs computations frame by frame, ignoring activation dynamics (activation value of each frame is not influenced by the values of preceding frames) and contraction dynamics (the tendon is considered as rigid and the parallel elastic element in muscles is ignored), it can produce remarkably similar muscle activations

and joint reaction forces compared to the ones produced by computationally intensive dynamic simulation [[Anderson and Pandy 2001](#)]. This step outputs muscle forces and activations necessary to produce the joint moments following the recorded kinematics.

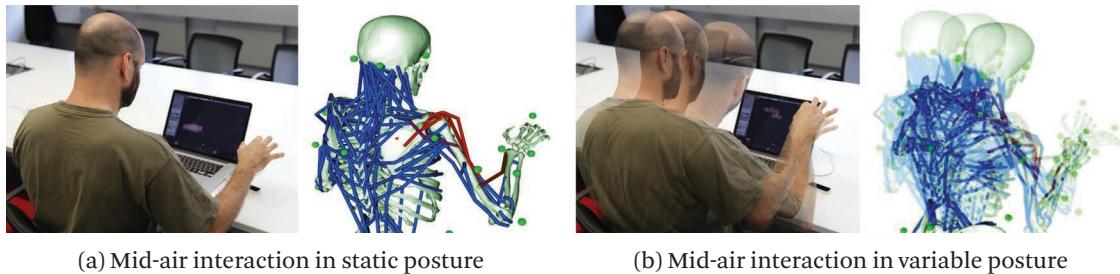
6. **Computed Muscle Control**, similar to the static optimization, computes muscle forces, activations, and additionally muscle excitations by the neural system. Unlike static optimization, CMC performs dynamically consistent simulation of movement taking into account contraction and activation dynamics, while still being computationally tractable, in contrast to full dynamic optimization. CMC integrates static optimization and forward simulation into a feedback loop with proportional-derivative control. Simply put, the muscle activations computed by static optimization are input into forward dynamics and integrated, and then the difference between the result from forward dynamics kinematics and the required kinematics is used to adjust the next static optimization input [[Theelen et al. 2003](#), [Theelen and Anderson 2006](#), [Delp et al. 2007](#)]. The output of the CMC algorithm are muscle forces, activations, and neural excitations producing the given kinematics with the musculoskeletal model.

The described algorithms represent the state of the art in biomechanical simulation and most of them or their variations are implemented in biomechanical software such as OpenSim [[Delp et al. 2007](#)], SIMM [[SIMM 2016](#)], AnyBody [[AnyBody 2016](#)], LifeModeler [[LifeModeler 2016](#)], and SantosHuman [[SantoHuman 2016](#)]. Their effectiveness and efficiency has been showcased in the biomechanics, rehabilitation, and sports fields. However, biomechanical simulation has only recently been used in the field of HCI. In the HCI setting, the simulation paradigm shifts from an application to a particular movement of a single subject in the context of a small body segment and tuning of hundreds of algorithm parameters, to a batch processing of experiment population data with the same set of algorithm parameters, covering a large variety of movements and simulating the whole human body.

## 7.6

### **Summary and Future Research Directions**

In the most recent work, motion capture-based biomechanical simulation has been introduced to multimodal interfaces [[Bachynskyi et al. 2014](#), [Bachynskyi et al. 2015a](#), [Bachynskyi et al. 2015b](#), [Nunnari et al. 2016](#)]. It was shown that biomechanical simulation, after a small set of adaptations [[Bachynskyi 2016](#)], can be successfully applied to a wide range of tasks [[Bachynskyi et al. 2014](#)], that it pro-



**Figure 7.5** Depiction of the setup of a user study comparing a conventional mid-air interface against a mid-air interface encouraging postural changes. Optical motion capture recorded the user's posture and movements and further biomechanical simulation-derived musculoskeletal loads.

duces valid results [Bachynskyi et al. 2014], and that they provide new insights into performance and ergonomics of an input method. For instance, Bachynskyi et al. [2015a] assessed the benefit of postural variability of upper body on muscular fatigue and the casually called “Gorilla arm effect.” Motion capture-based biomechanical simulation was used to understand differences in joint loads, muscle forces and activations. Figure 7.5 shows the two conditions which were studied: in static posture the load is all the time concentrated on the same muscles and they quickly get fatigued, in variable posture the loaded muscles change over time which gives more time for them to recover. The study shows a 15–25% decrease in average muscle loads on the principle muscles of the shoulder complex in the modified condition comparing to the baseline. Additionally, based on the data generated by the method, the large movement space of post-desktop interfaces can be summarized in a small set of homogeneous regions based on the underlying physiology executing the movement, namely based on muscle activation patterns [Bachynskyi et al. 2015b].

The experimental procedure described in the work above needs only slight overhead compared to common user performance studies: 10–15 minutes are necessary for putting on the suit with markers and calibrating the motion capture system. This method can be used within a UCD process to accelerate the assessment of design alternatives. The validity of the data is high enough for multimodal tasks, as it has been shown that correlation between the data generated by the deepest simulation step (static optimization) and actual EMG recording was positive with median 50% for a wide user population and large set of aimed movements [Bachynskyi et al.

2014]. Additionally, based on collected data, a summarization of the large movement space reachable by the arm has been proposed using muscle co-activation clustering [Bachynskyi et al. 2015b].

Currently, the method is applicable to analysis of movements larger than 4 cm, which is not suitable for analysis of small finger gestures. In order to be applicable to post-desktop interfaces, biomechanical simulation must account for the fine motor movement of fingers. Furthermore, data processing reveals a number of challenges: first, on a desktop computer computation of inverse dynamic, inverse kinematic, and static optimization takes time: 15, 50, and 1800 s, respectively, for the 500 ms movement of a single arm. Faster simulation methods exist [Murai et al. 2010, van den Bogert et al. 2013], but they have not yet been validated for multimodal interfaces. Even the fastest simulation software cannot make up for the time-consuming setup of markers and other manual inspections and interventions that are required in the preprocessing phase. To better support practitioners, future work should seek to streamline the method. A promising direction is *markerless* motion capture (e.g., Stoll et al. [2011]). It allows unhindered mobility for the subjects, and the setup effort is reduced to a minimum. However, markerless motion capture has lower accuracy. It remains to be examined for which tasks this lower accuracy is still sufficient.

Present body models work best with a middle-aged male. More work is needed in statistical body modeling that can account for statistical variation in body shapes and mass distributions. This way, the method will be applicable to a wider range of subjects. The model also makes simplifying anatomical assumptions, such as joints being “hinges,” although, for example, the thumb has a saddle joint. Improving the accuracy of the model will improve the permissible size of movements. Moreover, more studies are needed to understand limitations with different demographics. Existing validation study of biomechanical simulation in HCI [Bachynskyi et al. 2014, Nunnari et al. 2016] showed no real relationship between predicted muscle activations and self-reports of stress and tension. The experience of the user is obviously important for researchers, but it is not unusual to find a low correlation between objective and subjective measures in multimodal interaction [Hornbæk and Law 2007]. It is suspected that the reason for the low correlation is that the subjects were not tired enough for stress to emerge. Future work should examine whether muscle fatigue can be estimated when more muscle parameters are known [Neumann and Rowan 2002]. While it had been shown that the complex biomechanical dataset can be summarized by a small set of equivalence classes [Bachynskyi et al. 2015b], a principled neuromuscular control model has yet to be proposed. Such a model would allow a researcher to test new types of movements, outside of the reference dataset.

## Focus questions

- 7.1. What are the challenges and requirements for modern UX design?
- 7.2. What are the limitations of user-centered design (UCD) with respect to the requirements of modern interface design?
- 7.3. List and compare the advantages and disadvantages of the six categories of physical ergonomic models.
- 7.4. Describe the general organization of the musculoskeletal model commonly used in ergonomic studies. What are the three levels of these models, and how do they relate to each other?
- 7.5. What are the steps commonly used to perform a biomechanical simulation? What is the input and output of each step?
- 7.6. How precise is biomechanical simulation? In which context might current biomechanical simulations be problematic?
- 7.7. How would you assess the muscle activation and the ergonomics of a virtuoso pianist performing a 2-hour recital?
- 7.8. How would recent advances in non-supervised machine learning overcome the inherent limitations of current biomechanical simulation? Before answering this question, we recommend reading [Baltrušaitis et al. \[2018\]](#).
- 7.9. Imagine you could record both the kinematics and the muscle activation of all bones and muscles in the human body. Which HCI-related insights would you infer from this data?
- 7.10. Perform a bibliography search on the ergonomic guidelines booklets distributed with new desktop computers in the 1990s and Compare it to the guidelines provided with today's electronic consumer devices. How would you explain the difference?

## References

- A. Aaråas, M. Dainoff, O. Ro, and M. Thoresen. 2002. Can a more neutral position of the forearm when operating a computer mouse reduce the pain level for {VDU} operators? *International Journal of Industrial Ergonomics*, 30(4–5): 307–324. Musculoskeletal disorders in computer users. DOI: [10.1016/S0169-8141\(02\)00133-6](https://doi.org/10.1016/S0169-8141(02)00133-6). [269](#)
- C. Abras, D. Maloney-Krichmar, and J. Preece. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction*. Thousand Oaks: Sage Publications, 37(4): 445–456, 2004. [266](#)

- J. Andersen, J. Thomsen, and E. O. E. 2003. Computer use and carpal tunnel syndrome: A 1-year follow-up study. *JAMA*, 289(22): 2963–2969. DOI: [10.1001/jama.289.22.2963](https://doi.org/10.1001/jama.289.22.2963). **269**
- F. C. Anderson and M. G. Pandy. 2001. Static and dynamic optimization solutions for gait are practically equivalent. *Journal of Biomechanics*, 34(2): 153–161. **288**
- G. Andreoni, G. C. Santambrogio, M. Rabuffetti, and A. Pedotti. 2002. Method for the analysis of posture and interface pressure of car drivers. *Applied Ergonomics*, 33(6): 511–522. DOI: [10.1016/S0003-6870\(02\)00069-8](https://doi.org/10.1016/S0003-6870(02)00069-8). **273**
- AnyBody. 2016. <http://www.anybodytech.com/> Accessed July 1, 2016. **288**
- T. J. Armstrong, P. Buckle, L. J. Fine, M. Hagberg, B. Jonsson, A. Kilbom, I. A. Kuorinka, B. A. Silverstein, G. Sjogaard, and E. R. Viikari-Juntura. 1993. A conceptual model for work-related neck and upper-limb musculoskeletal disorders. *Scandinavian Journal of Work, Environment & Health*, 19(2): 73–84. **280**
- C. Ashruf. 2002. Thin flexible pressure sensors. *Sensor Review*, 22(4): 322–327. DOI: [10.1108/02602280210444636](https://doi.org/10.1108/02602280210444636). **273**
- K. Asundi, D. Odell, A. Luce, and J. T. Dennerlein. 2010. Notebook computer use on a desk, lap and lap support: Effects on posture, performance and comfort. *Ergonomics*, 53(1): 74–82. DOI: [10.1080/00140130903389043](https://doi.org/10.1080/00140130903389043). **269**
- M. Bachynskyi. 2016. Biomechanical models for human-computer interaction. Ph.D. thesis, Saarland University, Saarbrücken, Germany. **288**
- M. Bachynskyi, A. Oulasvirta, G. Palmas, and T. Weinkauf. 2014. Is motion capture-based biomechanical simulation valid for HCI studies?: Study and implications. In *Proceedings of the CHI*, pp. 3215–3224. ACM. DOI: [10.1145/2556288.2557027](https://doi.org/10.1145/2556288.2557027). **285, 288, 289, 290**
- M. Bachynskyi, G. Palmas, A. Oulasvirta, J. Steimle, and T. Weinkauf. 2015a. Performance and ergonomics of touch surfaces: A comparative study using biomechanical simulation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI ’15, pp. 1817–1826, New York. DOI: [10.1145/2702123.2702607](https://doi.org/10.1145/2702123.2702607). **288, 289**
- M. Bachynskyi, G. Palmas, A. Oulasvirta, and T. Weinkauf. January 2015b. Informing the design of novel input methods with muscle coactivation clustering. *ACM Transactions of Computer-Human Interactions*, 21(6):30: 1–30:25. DOI: [10.1145/2687921](https://doi.org/10.1145/2687921). **288, 289, 290**
- T. Baltrušaitis, C. Ahuja, and L.-P. Morency. 2018. Multimodal machine learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. **291**
- V. Baltzopoulos and D. Brodie. 1989. Isokinetic dynamometry. *Sports Medicine*, 8(2): 101–116. DOI: [10.2165/00007256-198908020-00003](https://doi.org/10.2165/00007256-198908020-00003). **273**

- S. Bao, N. Howard, P. Spielholz, and B. Silverstein. 2007. Two posture analysis approaches and their application in a modified rapid upper limb assessment evaluation. *Ergonomics*, 50(12): 2118–2136. [271](#)
- A. Belli, P. Bui, A. Berger, A. Geyssant, and J.-R. Lacour. 2001. A treadmill ergometer for three-dimensional ground reaction forces measurement during walking. *Journal of Biomechanics*, 34(1): 105–112. DOI: [10.1016/s0021-9290\(00\)00125-1](#). [273](#)
- B. Bernard, S. Sauter, and L. Fine. 1993. Hazard evaluation and technical assistance report: Los Angeles Times, Los Angeles, CA, Cincinnati, Oh: US Department of Health and Human Services. *Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, NIOSH Report No. HHE*, 90: 013–2277. [270](#)
- A. K. Blangsted, K. Søgaard, H. Christensen, and G. Sjøgaard. 2004. The effect of physical and psychosocial loads on the trapezius muscle activity during computer keying tasks and rest periods. *European Journal of Applied Physiology*, 91(2): 253–258. DOI: [10.1007/s00421-003-0979-z](#). [269](#)
- B. Blatter and P. Bongers. 2002. Duration of computer use and mouse use in relation to musculoskeletal disorders of neck or upper limb. *International Journal of Industrial Ergonomics*, 30(4–5): 295–306. DOI: [10.1016/S0169-8141\(02\)00132-4](#). [269](#)
- G. Borg. 1998. *Borg's Perceived Exertion and Pain Scales*. Human Kinetics, Champaign, IL. [270](#)
- W. Boucsein. 2012. *Electrodermal Activity*. Springer Science & Business Media. [273](#)
- W. Boucsein and M. Thum. 1997. Design of work/rest schedules for computer work based on psychophysiological recovery measures. *International Journal of Industrial Ergonomics*, 20(1): 51–57. DOI: [10.1016/S0169-8141\(96\)00031-5](#). [273](#)
- R. Breen, S. Pyper, Y. Rusk, and S. Dockrell. 2007. An investigation of children's posture and discomfort during computer use. *Ergonomics*, 50(10): 1582–1592. DOI: [10.1080/00140130701584944](#). [269](#)
- J. Cabri. 1991. Isokinetic strength aspects in human joints and muscles. *Applied Ergonomics*, 22(5): 299–302. DOI: [10.1016/0003-6870\(91\)90384-T](#). [273](#)
- T. M. Cappelli and V. G. Duffy. 2006. Motion capture for job risk classifications incorporating dynamic aspects of work. Technical report, SAE Technical Paper. DOI: [10.4271/2006-01-2317](#). [272](#)
- A. Chao, A. J. Kumar, K. Nagarajarao, and H. You. 2000. An ergonomic evaluation of cleco pliers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 4, p. 441. Human Factors and Ergonomics Society. DOI: [10.1177/154193120004402805](#). [272](#)
- K. Choi and A. Cichocki. Control of a wheelchair by motor imagery in real time. In C. Fyfe, D. Kim, S.-Y. Lee, and H. Yin, editors, *Intelligent Data Engineering and Automated Learning—IDEAL 2008*, vol. 5326 of *Lecture Notes in Computer Science*, pp. 330–337. Springer Berlin Heidelberg. DOI: [10.1007/978-3-540-88906-9\\_42](#). [277](#)

- A. Coldwells, G. Atkinson, and T. Reilly. 1994. Sources of variation in back and leg dynamometry. *Ergonomics*, 37(1): 79–86. DOI: [10.1080/00140139408963625](https://doi.org/10.1080/00140139408963625). 272
- D. Colombini, E. Occhipinti, N. Delleman, N. Fallentin, A. Kilbom, and A. Grieco. 2001. Exposure assessment of upper limb repetitive movements: a consensus document. *International Encyclopaedia of Ergonomics and Human Factors*, pp. 52–66. 272
- C. Cook, R. Burgess-Limerick, and S. Papalia. 2004. The effect of upper extremity support on upper extremity posture and muscle activity during keyboard use. *Applied Ergonomics*, 35(3): 285–292. The Occlusion Technique. DOI: [10.1016/j.apergo.2003.12.005](https://doi.org/10.1016/j.apergo.2003.12.005). 269
- E. N. Corlett, J. R. Wilson, and I. Manenica. April 1995. *Ergonomics Of Working Postures: Models, Methods And Cases: The Proceedings of the First International Occupational Ergonomics Symposium*, Zadar, Yugoslavia, 15–17. CRC Press, Boca Raton, FL. 281
- K. Costello, C. Nair, C. Roth, and D. Mitchell. October 1999. Portable electronic data collection apparatus for monitoring musculoskeletal stresses. US Patent 5,964,719. 272
- B. Das and D. N. Behara. 1995. Determination of the normal horizontal working area: a new model and method. *Ergonomics*, 38(4): 734–748. DOI: [10.1080/00140139508925145](https://doi.org/10.1080/00140139508925145). 280
- M. de Looze, I. Kingma, J. Bussmann, and H. Toussaint. 1992. Validation of a dynamic linked segment model to calculate joint moments in lifting. *Clinical Biomechanics*, 7(3): 161–169. DOI: [10.1016/0268-0033\(92\)90031-X](https://doi.org/10.1016/0268-0033(92)90031-X). 280
- S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, and A. Habib. 2007. OpenSim: Open-source software to create and analyze dynamic simulations of movement. *IEEE Trans. Biomedical Engineering*, 54(11): 1940–1950. DOI: [10.1109/TBME.2007.901024](https://doi.org/10.1109/TBME.2007.901024). 286, 288
- J. T. Dennerlein and P. W. Johnson. 2006a. Changes in upper extremity biomechanics across different mouse positions in a computer workstation. *Ergonomics*, 49(14): 1456–1469. DOI: [10.1080/00140130600811620](https://doi.org/10.1080/00140130600811620). 273
- J. T. Dennerlein and P. W. Johnson. 2006b. Different computer tasks affect the exposure of the upper extremity to biomechanical risk factors. *Ergonomics*, 49(1): 45–61. DOI: [10.1080/00140130500321845](https://doi.org/10.1080/00140130500321845). 272, 273
- L. Dipietro, A. Sabatini, and P. Dario. July 2008. A survey of glove-based systems and their applications. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(4): 461–482. DOI: [10.1109/TSMCC.2008.923862](https://doi.org/10.1109/TSMCC.2008.923862). 272
- V. G. Duffy. 2008. *Handbook of Digital Human Modeling: Research for Applied Ergonomics and Human Factors Engineering*. CRC Press, Boca Raton, FL. 268
- J. Dul, M. Douwes, and P. Smitt. 1994. Ergonomic guidelines for the prevention of discomfort of static postures based on endurance data. *Ergonomics*, 37(5): 807–815. DOI: [10.1080/00140139408963690](https://doi.org/10.1080/00140139408963690). 281
- N. Egelund. 1982. Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics*, 25(7): 663–672. DOI: [10.1080/00140138208925026](https://doi.org/10.1080/00140138208925026). 273

- Electromagnetic motion tracking system. 2016. <http://polhemus.com/motion-tracking/overview>. Accessed on July 1, 2016. 272, 283
- Y. Endo, S. Kanai, N. Miyata, M. Kouchi, M. Mochimaru, J. Konno, M. Ogasawara, and M. Shimokawa. 2008. Optimization-based grasp posture generation method of digital hand for virtual ergonomics assessment. *SAE International Journal of Passenger Cars—Electronic and Electrical Systems*, 1(2008-01-1902): 590–598. DOI: [10.4271/2008-01-1902](https://doi.org/10.4271/2008-01-1902). 280
- M. Essendrop, B. Schibye, and K. Hansen. 2001. Reliability of isometric muscle strength tests for the trunk, hands and shoulders. *International Journal of Industrial Ergonomics*, 28(6): 379–387. DOI: [10.1016/S0169-8141\(01\)00044-0](https://doi.org/10.1016/S0169-8141(01)00044-0). 272
- N. Fallentin, B. Juul-Kristensen, S. Mikkelsen, J. H. Andersen, J. P. Bonde, P. Frost, and L. Endahl. 2001. Physical exposure assessment in monotonous repetitive work—the prim study. *Scandinavian Journal of Work, Environment & Health*, 27(1): 21–29. 271
- M. Feuerstein, T. Armstrong, P. Hickey, and A. Lincoln. December 1997. Computer keyboard force and upper extremity symptoms. *Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine*, 39(12): 1144–1153. 269
- O. I. Franko, S. Lal, T. Pauyo, M. Alexander, D. Zurakowski, and C. Day. 2008. Validation of an objective device for assessing circumductive wrist motion. *The Journal of Hand Surgery*, 33(8): 1293–1300. DOI: [10.1016/j.jhsa.2008.03.012](https://doi.org/10.1016/j.jhsa.2008.03.012). 272
- C. Fransson-Hall, R. Gloria, Å. Kilbom, J. Winkel, L. Karlqvist, and C. Wiktorin. 1995. A portable ergonomic observation method (peo) for computerized on-line recording of postures and manual handling. *Applied Ergonomics*, 26(2): 93–100. DOI: [10.1016/0003-6870\(95\)00003-U](https://doi.org/10.1016/0003-6870(95)00003-U). 271
- A. Freivalds, D. B. Chaffin, A. Garg, and K. S. Lee. 1984. A dynamic biomechanical evaluation of lifting maximum acceptable loads. *Journal of Biomechanics*, 17(4): 251–262. DOI: [10.1016/0021-9290\(84\)90136-2](https://doi.org/10.1016/0021-9290(84)90136-2). 281
- D. Gagnon and M. Gagnon. 1992. The influence of dynamic factors on triaxial net muscular moments at the l5s1 joint during asymmetrical lifting and lowering. *Journal of Biomechanics*, 25(8): 891–901. 272, 273
- C. J. Garneau and M. B. Parkinson. 2008. Optimal product sizing through digital human models. Technical Report 2008-01-1921, SAE Technical Paper. DOI: [10.4271/2008-01-1921](https://doi.org/10.4271/2008-01-1921). 280
- B. Gerdle, N.-E. Eriksson, and C. Hagberg. 1998. Changes in the surface electromyogram during increasing isometric shoulder forward flexions. *European Journal of Applied Physiology and Occupational Physiology*, 57(4): 404–408. 281
- F. Gerr, C. P. Monteilh, and M. Marcus. 2006. Keyboard use and musculoskeletal outcomes among computer users. *Journal of Occupational Rehabilitation*, 16(3): 259–271. DOI: [10.1007/s10926-006-9037-0](https://doi.org/10.1007/s10926-006-9037-0). 269

- K. Gruber, H. Ruder, J. Denoth, and K. Schneider. 1998. A comparative study of impact dynamics: wobbling mass model versus rigid body models. *Journal of Biomechanics*, 31(5): 439–444. DOI: [10.1016/S0021-9290\(98\)00033-5](https://doi.org/10.1016/S0021-9290(98)00033-5). 284
- R. Gurram, S. Rakheja, and G. J. Gouw. 1995. A study of hand grip pressure distribution and emg of finger flexor muscles under dynamic loads. *Ergonomics*, 38(4): 684–699. DOI: [10.1080/00140139508925140](https://doi.org/10.1080/00140139508925140). 273
- C. Hagberg and M. Hagberg. April 1989. Surface emg amplitude and frequency dependence on exerted force for the upper trapezius muscle: a comparison between right and left sides. *European Journal of Applied Physiology and Occupational Physiology*, 58(6): 641–645. DOI: [10.1007/BF00418511](https://doi.org/10.1007/BF00418511). 281
- P. T. Hakala, A. H. Rimpelä, L. A. Saarni, and J. J. Salminen. Frequent computer-related activities increase the risk of neck–shoulder and low back pain in adolescents. *The European Journal of Public Health*, 16(5): 536–541. DOI: [10.1093/eurpub/ckl025](https://doi.org/10.1093/eurpub/ckl025). 269
- R. Harvey and E. Peper. Surface electromyography and mouse use position. *Ergonomics*, 40(8): 781–789. DOI: [10.1080/001401397187775](https://doi.org/10.1080/001401397187775). 269
- A. Hayward. December 2010. Deca sports freedom review. <http://www.gamesradar.com/deca-sports-freedom-review/>. 272
- S. Hignett and L. McAtamney. 2000. Rapid entire body assessment (reba). *Applied Ergonomics*, 31(2): 201–205. 271, 278
- V. H. Hildebrandt, P. M. Bongers, F. J. H. van Dijk, H. C. G. Kemper, and J. Dul. 2001. Dutch musculoskeletal questionnaire: description and basic qualities. *Ergonomics*, 44(12): 1038–1055. DOI: [10.1080/00140130110087437](https://doi.org/10.1080/00140130110087437). 270
- A. V. Hill. 1938. The heat of shortening and the dynamic constants of muscle. In *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 126(843): 136–195. DOI: [10.1098/rspb.1938.0050](https://doi.org/10.1098/rspb.1938.0050). 284
- K. Hinckley. 2017. A background perspective on touch as a multimodal (and multisensor) construct. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3015783.3015789](https://doi.org/10.1145/3015783.3015789). 270
- N. Hjortskov, D. Rissén, A. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1–2): 84–89. DOI: [10.1007/s00421-004-1055-z](https://doi.org/10.1007/s00421-004-1055-z). 273
- K. G. Holt, J. Hamill, and R. O. Andres. 1990. The force-driven harmonic oscillator as a model for human locomotion. *Human Movement Science*, 9(1): 55–68. DOI: [10.1016/0167-9457\(90\)90035-C](https://doi.org/10.1016/0167-9457(90)90035-C). 281
- Y. Hong, J. X. Li, A. S. K. Wong, and P. D. Robinson. 2000. Effects of load carriage on heart rate, blood pressure and energy expenditure in children. *Ergonomics*, 43(6): 717–727. DOI: [10.1080/001401300404698](https://doi.org/10.1080/001401300404698). 273

- K. Hornbæk and E. L.-C. Law. 2007. Meta-analysis of correlations among usability measures. In *Proceedings CHI*, pp. 617–626. DOI: [10.1145/1240624.1240722](https://doi.org/10.1145/1240624.1240722). 290
- R. Hornung, N. Chen, and P. van der Smagt. 2019. Early integration for movement modeling in latent spaces. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 270
- S. T. Iqbal, X. S. Zheng, and B. P. Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA'04, pp. 1477–1480, New York. ACM. DOI: [10.1145/985921.986094](https://doi.org/10.1145/985921.986094). 273
- ISO/IEC. 1998. Ergonomic requirements for office work with visual display terminals (VDTs)—part 11: Guidance on usability. The International Organization for Standardization. New York, NY. 266
- ISO/IEC. 1999. Human-centred design processes for interactive systems. The international Organization for Standardization. New York, NY. 266, 268
- ISO/IEC. 2009. Ergonomics of human system interaction—part 210: Human-centred design for interactive systems. International Standardization Organization (ISO). New York, NY. 266
- B. T. Jap, S. Lal, P. Fischer, and E. Bekiaris. 2009. Using {EEG} spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2, Part 1): 2352–2359. DOI: [10.1016/j.eswa.2007.12.043](https://doi.org/10.1016/j.eswa.2007.12.043). 273, 277
- C. Jensen, L. Finsen, K. Søgaard, and H. Christensen. 2002. Musculoskeletal symptoms and duration of computer and mouse use. *International Journal of Industrial Ergonomics*, 30(4–5): 265–275. DOI: [10.1016/S0169-8141\(02\)00130-0](https://doi.org/10.1016/S0169-8141(02)00130-0). 269
- P. W. Johnson, M. Hagberg, E. W. Hjelm, and D. Rempel. 2000. Measuring and characterizing force exposures during computer mouse use. *Scandinavian Journal of Work, Environment & Health*, 26(5): 398–405. 275
- E. S. Jung, D. Kee, and M. K. Chung. 1995. Upper body reach posture prediction for ergonomic evaluation models. *International Journal of Industrial Ergonomics*, 16(2): 95–107. DOI: [10.1016/0169-8141\(94\)00088-K](https://doi.org/10.1016/0169-8141(94)00088-K). 280
- O. Karhu, P. Kansi, and I. Kuorinka. 1977. Correcting working postures in industry: A practical method for analysis. *Applied Ergonomics*, 8(4): 199–201. 271
- D. Karlsson and B. Peterson. 1992. Towards a model for force predictions in the human shoulder. *Journal of Biomechanics*, 25(2): 189–199. 281
- P. J. Keir, J. M. Bach, and D. Rempel. 1999. Effects of computer mouse design and task on carpal tunnel pressure. *Ergonomics*, 42(10): 1350–1360. DOI: [10.1080/001401399184992](https://doi.org/10.1080/001401399184992). 269
- K. Kemmlert. 1995. A method assigned for the identification of ergonomic hazards—PLIBEL. *Applied Ergonomics*, 26(3): 199–211. DOI: [10.1016/0003-6870\(95\)00022-5](https://doi.org/10.1016/0003-6870(95)00022-5). 270

- S. Kim and M. A. Nussbaum. 2013. Performance evaluation of a wearable inertial motion capture system for capturing physical exposures during manual material handling tasks. *Ergonomics*, 56(2): 314–326. DOI: [10.1080/00140139.2012.742932](https://doi.org/10.1080/00140139.2012.742932). 272
- I. Kingma, M. P. de Looze, H. M. Toussaint, H. G. Klijnsma, and T. B. Bruijnen. 1996a. Validation of a full body 3-d dynamic linked segment model. *Human Movement Science*, 15(6): 833–860. DOI: [10.1016/S0167-9457\(96\)00034-6](https://doi.org/10.1016/S0167-9457(96)00034-6). 281
- I. Kingma, H. M. Toussaint, M. P. D. Looze, and J. H. V. Dieen. 1996b. Segment inertial parameter evaluation in two anthropometric models by application of a dynamic linked segment model. *Journal of Biomechanics*, 29(5): 693–704. DOI: [10.1016/0021-9290\(95\)00086-0](https://doi.org/10.1016/0021-9290(95)00086-0). 281
- E. A. Kirchner, S. H. Fairclough, and F. Kirchner. 2019. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 272
- T. Korhonen, R. Ketola, R. Toivonen, R. Luukkonen, M. Häkkänen, and E. Viikari-Juntura. 2003. Work related and individual predictors for incident neck pain among office employees working with video display units. *Occupational and Environmental Medicine*, 60(7): 475–482. 269, 280
- S. Kromodihardjo and A. Mital. 1986. Kinetic analysis of manual lifting activities: Part i—development of a three-dimensional computer model. *International Journal of Industrial Ergonomics*, 1(2): 77–90. DOI: [10.1016/0169-8141\(86\)90012-0](https://doi.org/10.1016/0169-8141(86)90012-0). 281
- W. Kuhmann. 1989. Experimental investigation of stress-inducing properties of system response times. *Ergonomics*, 32(3): 271–280. DOI: [10.1080/00140138908966087](https://doi.org/10.1080/00140138908966087). 273
- S. Kumar. 1994. A conceptual model of overexertion, safety, and risk of injury in occupational settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 36(2): 197–209. DOI: [10.1177/001872089403600202](https://doi.org/10.1177/001872089403600202). 278, 280
- S. Kumar and A. Mital. *Electromyography in Ergonomics*. CRC Press, Boca Raton, FL. 273
- I. Kuorinka, B. Jonsson, A. Kilbom, H. Vinterberg, F. Biering-Sørensen, G. Andersson, and K. Jørgensen. 1987. Standardised nordic questionnaires for the analysis of musculoskeletal symptoms. *Applied Ergonomics*, 18(3): 233–237. DOI: [10.1016/0003-6870\(87\)90010-X](https://doi.org/10.1016/0003-6870(87)90010-X). 270
- B. Laursen, B. R. Jensen, A. H. Garde, and A. H. Jørgensen. 2002. Effect of mental and physical demands on muscular activity during the use of a computer mouse and a keyboard. *Scandinavian Journal of Work, Environment & Health*, 28(4): 215–221. DOI: [10.5271/sjweh.668](https://doi.org/10.5271/sjweh.668). 269
- J. Leijnse, N. H. Campbell-Kyureghyan, D. Spektor, and P. M. Quesada. 2008. Assessment of individual finger muscle activity in the extensor digitorum communis by surface emg. *Journal of Neurophysiology*, 100(6): 3225–3235. DOI: [10.1152/jn.90570.2008](https://doi.org/10.1152/jn.90570.2008). 277

- G. Li and P. Buckle. 2000. Evaluating change in exposure to risk for musculoskeletal disorders—a practical tool. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(30): 5–407–5–408. DOI: [10.1177/154193120004403001](https://doi.org/10.1177/154193120004403001). 271
- LifeModeler. 2016. <http://www.lifemodeler.com/> Accessed on July 1, 2016. 288
- R. Likert. June 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140). 270
- J.-H. Lin, R. G. Radwin, and T. G. Richard. 2001. Dynamic biomechanical model of the hand and arm in pistol grip power handtool usage. *Ergonomics*, 44(3): 295–312. DOI: [10.1080/00140130118167](https://doi.org/10.1080/00140130118167). 281
- C.-W. Liu, C.-H. Chen, C.-L. Lee, M.-H. Huang, T.-W. Chen, and M.-C. Wang. 2003. Relationship between carpal tunnel syndrome and wrist angle in computer workers. *The Kaohsiung Journal of Medical Sciences*, 19(12): 617–622. DOI: [10.1016/S1607-551X\(09\)70515-7](https://doi.org/10.1016/S1607-551X(09)70515-7). 278, 280
- T.-W. Lu and J. O'Connor. 1999. Bone position estimation from skin marker co-ordinates using global optimisation with joint constraints. *Journal of Biomechanics*, 32(2): 129–134. 286
- D. Macfarlane. 2001. Automated metabolic gas analysis systems. *Sports Medicine*, 31(12): 841–861. DOI: [10.2165/00007256-200131120-00002](https://doi.org/10.2165/00007256-200131120-00002). 273
- W. Marras, F. Fathallah, R. Miller, S. Davis, and G. Mirka. 1992. Accuracy of a three-dimensional lumbar motion monitor for recording dynamic trunk motion characteristics. *International Journal of Industrial Ergonomics*, 9(1): 75–87. DOI: [10.1016/0169-8141\(92\)90078-E](https://doi.org/10.1016/0169-8141(92)90078-E).
- W. S. Marras, S. A. Lavender, S. E. Leurgans, S. L. Rajulu, W. G. Allread, F. A. Fathallah, and S. A. Ferguson. 1993. The role of dynamic three-dimensional trunk motion in occupationally-related low back disorders: The effects of workplace factors, trunk position, and trunk motion characteristics on risk of injury. *Spine*, 18(5): 617–628. 280
- W. S. Marras, S. A. Lavender, S. E. Leurgans, F. A. Fattalah, S. A. Ferguson, W. G. Allread, and S. L. Rajulu. 1995. Biomechanical risk factors for occupationally related low back disorders. *Ergonomics*, 38(2): 377–410. DOI: [10.1080/00140139508925111](https://doi.org/10.1080/00140139508925111). 280
- W. S. Marras, W. G. Allread, D. L. Burr, and F. A. Fathallah. 2000. Prospective validation of a low-back disorder risk model and assessment of ergonomic interventions associated with manual materials handling tasks. *Ergonomics*, 43(11): 1866–1886. DOI: [10.1080/00140130050174518](https://doi.org/10.1080/00140130050174518). 280
- J. Martins, E. Pires, R. Salvado, and P. Dinis. 1998. A numerical model of passive and active behavior of skeletal muscles. *Computer Methods in Applied Mechanics and Engineering*, 151(3–4): 419–433. DOI: [10.1016/S0045-7825\(97\)00162-X](https://doi.org/10.1016/S0045-7825(97)00162-X). 284
- S. Mathiassen, J. Winkel, and G. Hägg. 1995. Normalization of surface {EMG} amplitude from the upper trapezius muscle in ergonomic studies—a review. *Journal of Electromyography and Kinesiology*, 5(4): 197–226. DOI: [10.1016/1050-6411\(94\)00014-X](https://doi.org/10.1016/1050-6411(94)00014-X). 281

- S. E. Mathiassen and J. Winkel. 1990. Electromyographic activity in the shoulder-neck region according to arm position and glenohumeral torque. *European Journal of Applied Physiology and Occupational Physiology*, 61(5): 370–379. [281](#)
- L. McAtamney and E. N. Corlett. 1993. Rula: a survey method for the investigation of work-related upper limb disorders. *Applied Ergonomics*, 24(2): 91–99. [271](#)
- S. Melamed, J. Luz, T. Najenson, E. Jucha, and M. Green. Ergonomic stress levels, personal characteristics, accident occurrence and sickness absence among factory workers. *Ergonomics*, 32(9): 1101–1110. DOI: [10.1080/00140138908966877](#). [273](#)
- MetaMotion Gypsy Motion Capture System. <http://www.metamotion.com/gypsy/gypsy-motion-capture-system.htm> Accessed July 1, 2016. [272](#), [283](#)
- Metria Innovation Moire Motion Capture System. <http://www.metriainnovation.com/about> Accessed July 1, 2016. [283](#)
- Microsoft Kinect. <https://dev.windows.com/en-us/kinect> Accessed July 1, 2016. [283](#)
- Z. Ming, M. Närhi, and J. Siivola. 2004. Neck and shoulder pain related to computer use. *Pathophysiology*, 11(1): 51–56. DOI: [10.1016/j.pathophys.2004.03.001](#). [269](#)
- T. B. Moeslund and E. Granum. 2001. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3): 231–268. DOI: [10.1006/cviu.2000.0897](#). [272](#)
- T. B. Moeslund, A. Hilton, and V. Krüger. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2–3): 90–126. DOI: [10.1016/j.cviu.2006.08.002](#). [272](#)
- H. Moffet, M. Hagberg, E. Hansson-Risberg, and L. Karlqvist. 2002. Influence of laptop computer design and working position on physical exposure variables. *Clinical Biomechanics*, 17(5): 368–375. DOI: [10.1016/S0021-9290\(02\)00062-3](#). [269](#)
- J. S. Moore and A. Garg. 1995. The strain index: A proposed method to analyze jobs for risk of distal upper extremity disorders. *American Industrial Hygiene Association Journal*, 56(5): 443–458. DOI: [10.1080/15428119591016863](#). [271](#)
- R. H. Morton. 1996. A 3-parameter critical power model. *Ergonomics*, 39(4): 611–619. DOI: [10.1080/00140139608964484](#). [281](#)
- A. Murai, K. Kurosaki, K. Yamane, and Y. Nakamura. 2010. Musculoskeletal-see-through mirror: Computational modeling and algorithm for whole-body muscle activity visualization in real time. *Progress in Biophysics & Molecular Biology*, 103(2): 310–317. DOI: [10.1016/j.pbiomolbio.2010.09.006](#). [290](#)
- D. A. Neumann and E. E. Rowan. 2002. *Kinesiology of the Musculoskeletal System: Foundations for Physical Rehabilitation*. Mosby, St. Louis, MO. [290](#)
- J. Nielsen. 1994. *Usability Engineering*. Elsevier. [266](#)
- D. A. Norman. 2013. *The design of Everyday Things: Revised and Expanded Edition*. Basic Books. [266](#)
- F. Nunnari, M. Bachynskyi, and A. Heloir. 2016. Introducing postural variability improves the distribution of muscular loads during mid-air gestural interaction. In *Proceedings*

- of the 9th International Conference on Motion in Games, MIG '16, pp. 155–160, New York. ACM. DOI: [10.1145/2994258.2994278](https://doi.org/10.1145/2994258.2994278). 288, 290
- OpenSim. 2016. OpenSim documentation. <http://simtk-confluence.stanford.edu:8080/display/OpenSim/User>. Accessed May 3, 2016. 286, 287
- OptiTrack Motion Capture System Configurator. 2016. <http://www.optitrack.com/systems/#motive-body/prime-13/8>. Accessed July 1, 2016. 284
- OptiTrack Motion Capture Systems. 2016. <http://www.optitrack.com/motion-capture-biomechanics/>. Accessed July 1, 2016. 283
- L. Ortiz-Hernández, S. Tamez-González, S. Martínez-Alcántara, and I. Méndez-Ramírez. 2003. Computer use increases the risk of musculoskeletal disorders among newspaper office workers. *Archives of Medical Research*, 34(4): 331–342. DOI: [10.1016/S0188-4409\(03\)00053-5](https://doi.org/10.1016/S0188-4409(03)00053-5). 269
- F. Paterno. 2012. *Model-based Design and Evaluation of Interactive Applications*. Springer Science & Business Media. New York, NY. 268
- J. S. Petrofsky and A. R. Lind. 1978. Metabolic, cardiovascular, and respiratory factors in the development of fatigue in lifting tasks. *Journal of Applied Physiology*, 45(1): 64–68. DOI: [10.1152/jappl.1978.45.1.64](https://doi.org/10.1152/jappl.1978.45.1.64). 273
- PhaseSpace Impulse. 2016. <http://www.phasespace.com/impulse-motion-capture.html>. Accessed July 1, 2016. 274, 283
- S. Pheasant and C. M. Haslegrave. 2005. *Bodyspace: Anthropometry, Ergonomics and the Design of Work*. CRC Press, Boca Raton, FL. 280
- M. Pomplun and S. Sunkara. 2003. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*. 273
- D. C. Poole, S. A. Ward, G. W. Gardner, and B. J. Whipp. 1988. Metabolic and respiratory profile of the upper limit for prolonged exercise in man. *Ergonomics*, 31(9): 1265–1279. DOI: [10.1080/00140138808966766](https://doi.org/10.1080/00140138808966766). 273
- M. Praagman, E. Chadwick, F. van der Helm, and H. Veeger. 2006. The relationship between two different mechanical cost functions and muscle oxygen consumption. *Journal of Biomechanics*, 39(4): 758–765. DOI: [10.1016/j.jbiomech.2004.11.034](https://doi.org/10.1016/j.jbiomech.2004.11.034). 287
- B. I. Prilutsky and V. M. Zatsiorsky. 2002. Optimization-based models of muscle coordination. *Exercise and Sport Sciences Reviews*, 30(1): 32. 287
- A. Puerta. July 1997. A model-based interface development environment. *Software, IEEE*, 14(4): 40–47. DOI: [10.1109/52.595902](https://doi.org/10.1109/52.595902). 268
- Qualisys Motion Capture Systems. <http://www.qualisys.com/> on 07.01.2016. 283
- R. G. Radwin, S. Oh, T. R. Jensen, and J. G. Webster. External finger forces in submaximal five-finger static pinch prehension. *Ergonomics*, 35(3): 275–288. DOI: [10.1080/00140139208967813](https://doi.org/10.1080/00140139208967813). 273
- S. Rodgers. 1992. A functional job analysis technique. *Occupational Medicine*, 7(4): 679–711. 281

- W. Rohmert. 1973a. Problems of determination of rest allowances part 2: Determining rest allowances in different human tasks. *Applied Ergonomics*, 4(3): 158–162. DOI: [10.1016/0003-6870\(73\)90166-X](https://doi.org/10.1016/0003-6870(73)90166-X). 281
- W. Rohmert. 1973b. Problems in determining rest allowances. *Applied Ergonomics*, 4(2): 91–95. DOI: [10.1016/0003-6870\(73\)90082-3](https://doi.org/10.1016/0003-6870(73)90082-3). 281
- W. Rohmert. 1984. Das belastungs-beanspruchungs-konzept. *Zeitschrift für Arbeitswissenschaft*, 38(4): 193–200. 280
- SantosHuman. 2016. <http://www.santoshumaninc.com/>. Accessed July 1, 2016. 288
- B. Sayers. 1973. Analysis of heart rate variability. *Ergonomics*, 16(1): 17–32. DOI: [10.1080/00140137308924479](https://doi.org/10.1080/00140137308924479). 273
- E. B. Schlossberg, S. Morrow, A. E. Llosa, E. Mamary, P. Dietrich, and D. M. Rempel. 2004. Upper extremity pain and computer use among engineering graduate students. *American Journal of Industrial Medicine*, 46(3): 297–303. DOI: [10.1002/ajim.20071](https://doi.org/10.1002/ajim.20071). 269
- F. Schuind, M. Garcia-Elias, W. P. C. III, and K.-N. An. 1992. Flexor tendon forces: In vivo measurements. *The Journal of Hand Surgery*, 17(2): 291–298. DOI: [10.1016/0363-5023\(92\)90408-H](https://doi.org/10.1016/0363-5023(92)90408-H). 272
- SIMM. <http://www.musculographics.com/html/products/simm.html>. Accessed July 1, 2016. 288
- C. M. Sommerich, H. Starr, C. A. Smith, and C. Shivers. 2002. Effects of notebook computer configuration and task on user biomechanics, productivity, and comfort. *International Journal of Industrial Ergonomics*, 30(1): 7–31. DOI: [10.1016/S0169-8141\(02\)00075-6](https://doi.org/10.1016/S0169-8141(02)00075-6). 269
- L. Sperling, S. Dahlman, L. Wikström, Å. Kilbom, and R. Kadefors. 1993. Special issue hand tools for the 1990s a cube model for the classification of work with hand tools and the formulation of functional requirements. *Applied Ergonomics*, 24(3): 212–220. 280
- S. Sridhar, A. Oulasvirta, and C. Theobalt. December 2013. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2456–2463. DOI: [10.1109/ICCV.2013.305](https://doi.org/10.1109/ICCV.2013.305). 274
- N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, and H. W. Hendrick. 2004. *Handbook of human factors and ergonomics methods*. CRC Press, Boca Raton. 270, 279
- W. T. Stauber, E. R. Barill, R. E. Stauber, and G. R. Miller. 2000. Isotonic dynamometry for the assessment of power and fatigue in the knee extensor muscles of females. *Clinical Physiology*, 20(3): 225–233. 273
- C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. November 2011. Fast articulated motion tracking using a sums of gaussians body model. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 951–958. DOI: [10.1109/ICCV.2011.6126338](https://doi.org/10.1109/ICCV.2011.6126338). 274, 290
- G. P. Szeto and R. Lee. 2002. An ergonomic evaluation comparing desktop, notebook, and subnotebook computers. *Archives of Physical Medicine and Rehabilitation*, 83(4): 527–532. 269

- H. Tan, L. Slivovsky, and A. Pentland. September 2001. A sensing chair using pressure distribution sensors. *Mechatronics, IEEE/ASME Transactions on*, 6(3): 261–268. DOI: [10.1109/3516.951364](https://doi.org/10.1109/3516.951364). 273
- S. Tanaka and J. D. McGlothlin. 1993. A conceptual quantitative model for prevention of work-related carpal tunnel syndrome (cts). *International Journal of Industrial Ergonomics*, 11(3): 181–193. DOI: [10.1016/0169-8141\(93\)90107-O](https://doi.org/10.1016/0169-8141(93)90107-O). 278, 280
- The Captury Markerless Motion Capture Systems. 2016. <http://www.thecaptury.com/>. Accessed July 1, 2016. 283
- D. G. Thelen and F. C. Anderson. 2006. Using computed muscle control to generate forward dynamic simulations of human walking from experimental data. *Journal of Biomechanics*, 39(6): 1107–1115. DOI: [10.1016/j.jbiomech.2005.02.010](https://doi.org/10.1016/j.jbiomech.2005.02.010). 288
- D. G. Thelen, F. C. Anderson, and S. L. Delp. 2003. Generating dynamic simulations of movement using computed muscle control. *Journal of Biomechanics*, 36(3): 321–328. 288
- S. Thorn, M. Forsman, Q. Zhang, and K. Taoda. 2002. Low-threshold motor unit activity during a 1-h static contraction in the trapezius muscle. *International Journal of Industrial Ergonomics*, 30(4–5): 225–236. DOI: [10.1016/S0169-8141\(02\)00127-0](https://doi.org/10.1016/S0169-8141(02)00127-0). 273
- D. Tsirakos, V. Baltzopoulos, and R. Bartlett. Inverse optimization: functional and physiological considerations related to the force-sharing problem. *Critical Reviews in Biomedical Engineering*, 25(4–5): 371–407, 1997. 287
- A. J. van den Bogert, T. Geijtenbeek, O. Even-Zohar, F. Steenbrink, and E. C. Hardin. 2013. A real-time system for biomechanical analysis of human movement and muscle function. *Medical & Biological Engineering & Computing*, pp. 1–9. DOI: [10.1007/s11517-013-1076-z](https://doi.org/10.1007/s11517-013-1076-z). 290
- M. van Tulder, A. Malmivaara, and B. Koes. May 2007. Repetitive strain injury. *The Lancet*, 369(9575): 1815–1822. 267, 784
- Vicon Optical Motion Capture Systems. <http://www.vicon.com/what-is-motion-capture>. Accessed July 1, 2016. 283
- B. Visser, M. P. D. Looze, M. P. D. Graaff, and J. H. V. Dieën. 2004. Effects of precision demands and mental pressure on muscle activation and hand forces in computer mouse tasks. *Ergonomics*, 47(2): 202–217. DOI: [10.1080/00140130310001617967](https://doi.org/10.1080/00140130310001617967). 269
- T. R. Waters, V. Putz-Aenderson, A. Garg, and L. J. Fine. 1993. Revised niosh equation for the design and evaluation of manual lifting tasks. *Ergonomics*, 36(7): 749–776. DOI: [10.1080/00140139308967940](https://doi.org/10.1080/00140139308967940). 280
- T. R. Waters, V. Putz-Anderson, and A. Garg. 1994. *Applications Manual for the Revised NIOSH Lifting Equation*. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Biomedical and Behavioral Science. 280

- A. Weber, C. Fussler, J. F. O'Hanlon, R. Gierer, and E. Grandjean. 1980. Psychophysiological effects of repetitive tasks. *Ergonomics*, 23(11): 1033–1046. DOI: [10.1080/00140138008924812](https://doi.org/10.1080/00140138008924812). 273, 277
- J. Winkel and S. E. Mathiassen. 1994. Assessment of physical work load in epidemiologic studies: concepts, issues and operational considerations. *Ergonomics*, 37(6): 979–988. DOI: [10.1080/00140139408963711](https://doi.org/10.1080/00140139408963711). 280
- Xsens Motion Capture Systems. 2016. <https://www.xsens.com/products/mvn-biomech/>. Accessed July 1, 2016. 283
- A. Zecevic, D. I. Miller, and K. Harburn. 2000. An evaluation of the ergonomics of three computer keyboards. *Ergonomics*, 43(1): 55–72. DOI: [10.1080/001401300184666](https://doi.org/10.1080/001401300184666). 269
- D. Zennaro, T. Läubli, D. Krebs, A. Klipstein, and H. Krueger. 2003. Continuous, intermittent and sporadic motor unit activity in the trapezius muscle during prolonged computer work. *Journal of Electromyography and Kinesiology*, 13(2): 113–124. 273
- J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad. 2018. Multimodal behavioural and physiological signals as indicators of cognitive load. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108002](https://doi.org/10.1145/3107990.3108002). 272



# Early Integration for Movement Modeling in Latent Spaces

Rachel Hornung, Nutan Chen, Patrick van der Smagt

## 8.1

### Introduction

In this chapter, we will show how techniques of advanced machine and deep learning can be used for multimodal integration in movement modeling. Using techniques that represent the data in lower-dimensional latent spaces, the algorithms' abilities to identify important information and inherent feature calculation are exploited, rather than building upon specialists' knowledge. The algorithms are provided raw sensory data. They process it automatically to return the desired output. The methods we suggest are suitable for generating new samples of data and scale to an arbitrary number of modalities. We will show that these methods are applicable to different settings and that they are broadly deployable rather than just suited to a specific user group or application. The data can stem from any source, including but not limited to, body-machine interfaces, optical tracking, or microphones.

Although observing only a single physical movement, it can be recorded through various channels. While the *visual appearance* of motion may be captured by video cameras, depth cameras, or tracking devices, the *sound* of the same movement may provide further information. For human movement, biometric measurements—e.g., *electromyography (EMG)* or *electroencephalography (EEG)*—can provide another augmenting channel of data.

We are concerned with the means of computing models of physical motion from multimodal observations of that motion. While we focus on human motion models

in the context of this chapter, the same principles may be applied to animal and robotic motion.

After introducing the benefits of movement models and early integration for these models, we define a use case for better comprehension of the concepts. In Section 8.2 we present current approaches to motion modeling, including a description of the different spaces of representation. Section 8.3 explains how *early fusion* can be applied to movement modeling, starting from applicable algorithms and ending with possible fusions thereof, also taking deep learning into account. Finally, we elaborate how the concepts introduced in Section 8.3 may be applied to the previously introduced use case. In order to aid comprehension we provide [Focus Questions](#) and a [Glossary](#).

### 8.1.1 Application of Movement Models

Movement models are useful for recognition, categorization, and prediction. A typical use is, for instance, for medical or rehabilitation purposes: one may compare observed motion with a general, normed movement model for diagnostic purposes. Alternatively, comparing the movement model of one subject over time may prove changes throughout a period of rehabilitation or training, or progress of a disease. Using a motion model this change can be evaluated objectively rather than by subjective observation. This use of movement models is related to medical cyber-physical systems and medical decision support as introduced in Chapter 11.

Related to companion technologies introduced in Chapter 11 and referenced in Chapter 13, the models may also be used for control, e.g., to adjust a rehabilitation robot exactly to the needs of a patient and only support motion execution if the patient is actively participating. As Chapters 5 and 9 from Volume 2 pointed out, implicit information can convey valuable information: by monitoring the stress level in accordance with muscular activity the optimal amount of motion support may be provided to the patient. For a hemiparetic patient the motion of the non-affected arm can be used to guide the motion of the affected arm. And for astronauts or weightlifters, the same information may be used to increase resistance, i.e., automatically adjust the weights in a leg press, and thus improve the training effect.

Motion models are not limited to medical and analytical applications. They can also be used to control any other type of robot, and to, e.g., teach it new trajectories that should be executed in an industrial setting. It may be desirable to use the models to generate new data for rendering naturally looking motion of characters, in spite of a new body shape or motion speed. The latter is of great interest, e.g., for computer games, and is also applicable to the motion of animals [[Kang et al. 2006](#), [Taylor et al. 2007](#)].

## Glossary

**An autoencoder (AE)** is an unsupervised neural network that has the same output as input. Using a latent layer with fewer dimensions than the original data, the network is forced to compress the data and find a lower-dimensional representation for the data. The net can be arbitrarily deep and transformations between layers can be nonlinear.

**Deep learning** refers to a subgroup of machine learning algorithms. It comprises methods that use several processing layers with (non)linear transformations. Neural networks that include at least one hidden layer belong to this group. The methods learn different representations of the data autonomously. These representations differ based on the task they are intended to solve and the architecture selected.

**A dynamic movement primitive (DMP)** is a nonlinear dynamic system trained from demonstration of a trajectory. Using a point attractor in a second-order dynamic system a parametrizable description of motion can be obtained.

**Dropout** is a technique applied in neural network learning. Some of the in- or ouput neurons are artificially and randomly set to zero. This way the algorithm has to learn to handle corrupted data and prevents coadaptation of neurons, while improving the detection of correlations. Dropout can also be used to learn correlations between different modalities.

**Early fusion.** After possibly preprocessing data of different modalities, the data is merged and a model for the mixed data is calculated from it.

**A Gaussian process (GP)** is a stochastic process defined by its mean and covariance. Assuming that similar inputs behave similarly, test data targets will be similar to closely located input data targets, based on multidimensional Gaussian distributions.

**Gaussian process dynamical models (GPDMs)** are based on the *Gaussian process latent variable models (GPLVM)*. By incorporating temporal dependency, they enforce a smooth latent space.

**Gaussian process latent variable models (GPLVMs)** are based on the *Gaussian process (GP)*. They generate a low-dimensional representation of the input space.

**Late fusion.** Data of multiple modalities is processed for each modality individually. The decisions for each modality are merged based on their probabilities or other means of comparison or by applying a further learning algorithm for merging. The final result is based on data of all modalities. However, correlations between modalities are not exploited.

**Glossary (continued)**

**Latent representation** substitutes data in a lower dimensional space than the original data. It is capable of explaining the changes in the data. If the mapping between the independent factors accountable for changes in the observed data, and the observed data is known or can be learned, a latent representation should have that intrinsic dimensionality as it is the minimal description of the data.

**Variational autoencoders (VAEs)** are a variational implementation of [autoencoders \(AEs\)](#). They can be trained via stochastic gradient descent.

Gesture recognition also depends on models of the motions to be detected. Applications range from manipulation, e.g., interaction with virtual objects through controlling, as in human-robot teleoperation, to communication, e.g., understanding sign language. For instance, in self-driving cars it is essential to correctly classify the gestures of traffic officers to avoid fatal crashes.

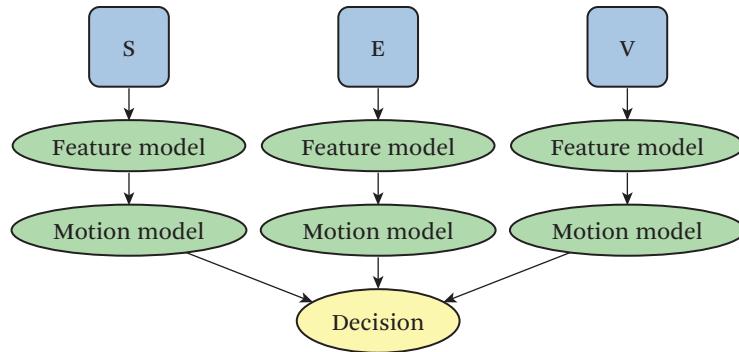
Today, most motion modeling is performed using only a single modality (cf. Section 8.2). If multiple data sources are consulted, they are typically used to recognize (one of a few) motions, rather than to find an integrated representation of that motion.

### 8.1.2 Different Levels of Sensor Fusion

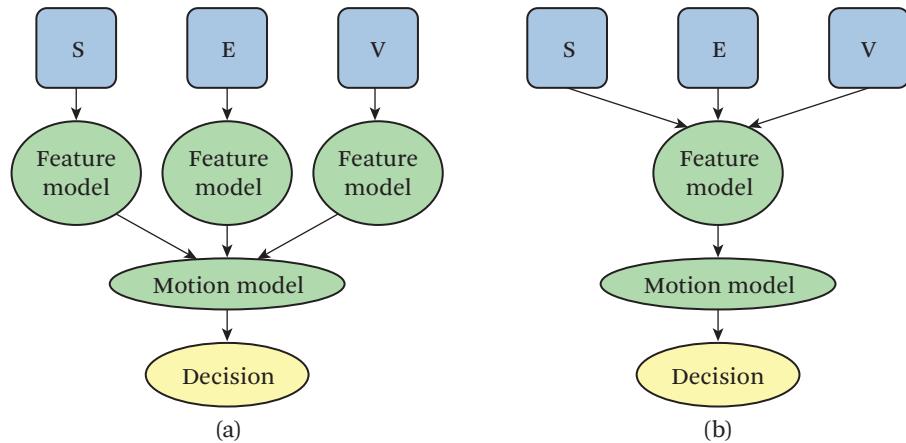
Multimodal integration for motion modeling can take place at different levels. Three such levels are distinguished by [Ross and Govindarajan \[2005\]](#):

- sensor level;
- feature level; and
- match score/rank/decision level.

While match score, rank, and decision level fusion use the model performance of each modality to form a final decision, they are distinguished by the different outputs to which they are applied. Sometimes they are also referred to as semantic level fusion [[Snoek et al. 2005](#)]. In contrast, in sensor-level and feature-level fusion a model that encompasses all modalities at once is formed. In sensor-level fusion the mixed model is fed with raw data; for feature-level fusion raw data is preprocessed and features are extracted which in turn are fed to the mixed model. The term [late fusion](#) refers to decision-level integration, as depicted in Figure 8.1, while



**Figure 8.1** Late fusion: Fusion at the decision level. Features can be calculated, but direct processing of sensor data is possible, as well. V, E, and S indicate different source types, e.g., video, EMG, and sound.



**Figure 8.2** Different setups for early multimodal fusion. V, E, and S indicate different source types, e.g., video, EMG, and sound. (a) Feature-level (early) fusion. Features may be expert designed or automatically generated. (b) Sensor-level (early) fusion. If features are desired, they have to be generated automatically.

early fusion refers to multimodal integration at the sensor level or the feature level (cf. Figure 8.2).

In most cases, motion is not entirely observable through a single modality. For example, a video of a moving arm will give no information on the muscular activity. In this context, late fusion is especially helpful from an engineering perspective:

modeling is addressed via a devide-and-conquer approach; each modality is handled separately; and the final model is a tradeoff between the different inputs. For example, [Vinciarelli and Esposito \[2018\]](#) apply late fusion in order to identify emotions. However, correlations between modalities may remain undiscovered in late fusion and information that can only be decoded after merging will be discarded as noise. We assume that by jointly processing all modalities, i.e. using early fusion, the different modalities can help to disambiguate each other. Therefore, while still difficult, we advocate the use of early fusion and the omission of (expert-designed) features.

In Section 8.3 we introduce a concept for early fusion for motion modeling. The underlying approaches are based on machine learning, as modeling all complexities of dynamics and estimating the real physical parameters is close to intractable [[Taylor et al. 2007](#)]. Our main focus is on using raw sensor data. Nevertheless, the general concept is applicable to data features as well.

### 8.1.3 Benefits of Early Multimodal Integration for Motion Modeling

During unimodal processing of the data valuable information may be lost and intermodal correlations will remain undiscovered [[Snoek et al. 2005](#)]. Besides, the information available for fusion is richer at the sensor or feature level than at the decision level [[Ross and Govindarajan 2005](#)]. Further, [Snoek et al. \[2005\]](#) state that late fusion is a two-level approach with an expensive learning routine as each modality and fusion have to be handled separately. Thus, we advocate the use of early fusion approaches that do not depend on expert-designed feature calculation.<sup>1</sup> Instead, we suggest using methods that are able to identify relevant information from the combined sources autonomously. Nevertheless, the concepts described in Section 8.3.2 can be used for featured data, as well.

Disadvantages attributed to early fusion, according to [Oviatt and Cohen \[2015\]](#), are as follows:

- Long processing times: The data space of early fusion approaches is comparatively large and may suffer from the curse of dimensionality [[Bellman 1957, Bellman 2003, Keogh and Mueen 2010](#)].
- Inability to model reliability of each modality under different conditions: for different situations different sources may contain different levels of information, e.g., the worse the illumination of a scene, the less information can

---

<sup>1</sup>. Section 8.2.3 describes the benefits and disadvantages of expert-designed features in more detail.

be extracted from an RGB camera, while an audio stream is easier to decode if the recording takes place in a quiet office room compared to a crowded market place.

However, using a latent space representation as introduced in Section 8.2.2, parallelization of processing steps, and using GPU calculation the problem of long processing times can be alleviated. Besides, it is of greater importance that a model is applicable online, while model learning often may take place offline, if necessary. The techniques we introduce in the following are able to differentiate the reliabilities of the different modalities. Provided that the method is trained with sufficient amounts of training data from different conditions, the system can learn to adjust automatically.

As a good example for skepticism towards early integration, [Ross and Govindarajan \[2005\]](#) mention that the model design may need to be more complex for early fusion cases. In the following we will provide concepts that consider that the modalities differ from each other, but allows for combined training of a mixed model. Thus, this early fusion approach is not more complex than having to train a model for each modality and a fusion model.

In contrast, late fusion is considered to be:

- tailorabile: the best known classifier or regressor can be used for each signal type;
- synchronous: while the data itself may not be time-synchronous, the decisions can be;
- efficient: the data size is limited, thus the computational requirements are lower;
- accurate: up-to-date late fusion mostly outperforms early fusion; and
- simple: black box recognizers can be used, thus allowing for a large choice of methods and swapping of these within a single interface [[Oviatt and Cohen 2015](#)].

Yet, tailoring the system requires much domain knowledge, and will take a great amount of fine tuning for each modality. The question of synchrony is addressed by the algorithms we suggest in the following. While compared to raw data or the full features, the amount of required memory may be lower for late fusion approaches. Yet these systems usually do not consider latent space (cf. Section 8.2.2) representations, and calculations are required for each source plus the final decision level. Late fusion approaches have been studied more extensively, thus a better

performance is not surprising. Regarding simplicity, the use of different black box recognizers also implies that numerous different approaches should be evaluated, while the user may not have an understanding of the underlying principles and thus may not be able to take justified decisions. Our suggestions, in contrast, leave most of the decisions to the model and take much of the engineering burden from the user.

A further reason against late fusion is that the late combination of modalities often assumes conditional independence of the modalities [Oviatt and Cohen 2015], which does not have to be the case. The higher the correlation between modalities, the stronger the call for early fusion [Oviatt and Cohen 2015]. According to Srivastava and Salakhutdinov [2014] it should be possible to reach a joint representation even though single modalities may be missing after training, and that missing modality values could be inferred from the available ones. In order to achieve this it is inevitable to use a mixed-model approach as introduced here.

#### 8.1.4 Substance of Multimodal Motion Models

Modalities we consider reasonable for motion modeling include, among others, RGB video recordings, depth videos, tracking data, EMG, and EEG signals. It is possible to include data that do not directly reflect the motion but may be related, like the heart rate giving information about the effort, or the galvanic skin response indicating psychological activation (see also Kirchner et al. [2018]). Apart from multimodality, we must also consider multisensory data e.g., multiple cameras observing the same movement, or multiple EMG electrodes. While this might impact a late fusion approach, the difference for the methods introduced in Section 8.3 is negligible.

In Section 8.2, we will present current approaches to motion modeling and modality fusion. Section 8.3 will outline a concept for early fusion in motion modeling.

Motion models as defined in this chapter are not implicitly capable of classifying specific action types or activity understanding, respectively, gesture recognition. It is possible to combine the methods described in Section 8.3 with other methods to achieve these goals. However, these additional tasks will not be addressed in the following text.

#### 8.1.5 Definition of a Use Case for Motion Modeling

In order to make the following explanations more tangible, we assume a concrete motion modeling example: We want to model temporal motion of a single human

arm in free space. The model is intended to be used as a control signal for a robotic device that executes the task at a remote location. The available data is:

- an RGB video of the torso and arm of the subject with a resolution of  $640 \times 480$  pixels, leading to  $640 \times 480 \times 3 = 921,600$  dimensions, with a sampling rate of 30 Hz;
- a recording of eight surface EMG electrodes placed on the respective arm and shoulder of the subject; using a sliding window the data is bundled in blocks of 50 time steps, resulting in  $8 \times 50 = 400$  dimensions; the sampling rate of a single recording is 2000 Hz, so every 25 ms the sample does not contain any of the previous information; and
- an optical tracking recording of 5 different markers, each specified by their position and orientation in 3D, as the latter is represented in angle-axis each marker has seven dimensions, thus resulting in  $5 \times 7 = 35$  dimensions with a sampling rate of 100 Hz.

We assume that at each time step  $t$  we get the latest information from each modality. If the sampling rate  $t$  is bigger than a modality's sampling rate we will retrieve repeated information—it is also possible to extrapolate the data, while placing a greater computational load on the system—, if it is smaller than a modality's sampling rate, we will disregard the additional data. We choose  $t$  to have the same rate as the smallest sampling rate for entirely new data, i.e., 30 Hz.

## 8.2

### State of the Art for Motion Modeling

Most state-of-the-art approaches focus on late fusion. This is due to the fact that often fusion of raw data is computationally very intensive, and on top of that raw data is expressed in the domain of the sensor, where engineering models often fail for all but the simplest sensors. In order to be able to understand the challenges in multimodal fusion for motion modeling, we will give a brief overview of concepts for motion modeling, spaces applicable for motion modeling, motion modeling in low-dimensional spaces, and multimodal fusion approaches applied to motion modeling.

#### 8.2.1 Basic Motion Modeling Principles

For the applications indicated in Section 8.1.1, we need to generate movement models. However, current motion modeling approaches are unsatisfactory, as we will expose below.

Typical approaches are based on optimization of cost functions. An expert uses, for example, equations of dynamics to define the structure of the model. Only if these mathematical functions match the principles of the motion of interest closely, will the optimization result in a useful representation thereof. To achieve sufficient accuracy in articulated objects with several joints, the mathematical models are complex and will cause large computational efforts [Safanova et al. 2004, Taylor et al. 2007, Havoutis 2007]. If the modality of interest does not represent individual joints, but an image of the motion, as in the previously introduced example, optimization of the motion on the raw data becomes intractable.

Adding physical constraints to the optimization task can help generate more realistic motion models. Constraint-based motion optimization has been used in computer graphics since 1988 [Safanova et al. 2004, Witkin and Kass 1988]. While sufficient time and computational power facilitate optimization of very complex motion models—for example, Anderson and Pandy [2001] model muscle activations that minimize the metabolic energy during a single step using a human model consisting of 10 segments, 23 *degrees of freedom* (DoF), and 54 actuating muscles—months of processing time on a supercomputer are not feasible for deploying motion models. In practice, optimization techniques are hardly applicable for larger numbers of DoF, if physically correct motion is desired [Safanova et al. 2004]. Further, optimization-based modeling does not provide for instantaneous generation of new motion, as highlighted by Kang et al. [2006]. Generative modeling is of especial importance when dealing with transitions between motions and if new data has to be generated as in computer animation.

Motion graphs as introduced by Kovar et al. [2002] split recorded motion into subsequences presenting specific parts of a motion. For example, walking can be separated into forward motion, right turn, and left turn. The motion graph allows for easy transition between the different (sub-)motions, however it is limited to motions that have been recorded, requires the user to specify a transition threshold where it is possible to move from one sub-motion to another, and may not be applicable for large data sets. Motion graphs can easily be applied to the RGB and tracking data of our example. Contrary, for EMG data, the definition of transition thresholds will be very difficult, as the raw data at single points in time does not provide sufficient information about the current system state.

Motion models that are strictly based on demonstrations (cf., e.g., Schaal [1997]), have predefined start and end points, and the motion is attributed fixed time frames. Thus, they are hard to adjust and difficult to reuse in new settings [Havoutis 2007]. Putting this to our example, reducing the speed will require the extension of the sequence; in an RGB recording this might be solved by interpo-

lation between the different frames; from a recording of raw EMG data it is fairly impossible to generate an appropriate signal.

In contrast, it is possible to use dynamical systems to model motion. A *dynamic movement primitive (DMP)* [Schaal 2003, Ijspeert et al. 2013] is a nonlinear dynamic system. DMPs are generally trained from a demonstration of a trajectory in a specific modality, commonly state space of a robot or Cartesian space of a robot end effector (cf. Section 8.2.2). Subsequently, they can be reproduced and adjusted. A DMP is a point attractor using a second-order dynamic system:

$$\tau \ddot{\mathbf{y}} = \alpha_z(\beta_z(\mathbf{g} - \mathbf{y}) - \dot{\mathbf{y}}) + \mathbf{f},$$

where  $\tau$  is a time constant and  $\alpha_z$  and  $\beta_z$  are damping constants. The difference term  $(\mathbf{g} - \mathbf{y})$  attracts the trajectory to the goal position  $\mathbf{g}$ —the final frame of the demonstration—where  $\mathbf{y}$  is the trajectory. The parameters can be adjusted to speed up, slow down, and change the start, as well as the goal state.

The dynamic system can force a robot or a humanoid model to follow the demonstrated trajectories using a specified force term  $\mathbf{f}$ . This force term has two formalizations: a discrete attractor or a rhythmic attractor; applied according to the type of movement.

A DMP generates smooth kinematic control policies from recorded movements, and it can generate new movements by changing parameters. For example, after training a single demonstration, a robot can replace a cup to various locations without re-training. There are many extensions to this basic DMP framework, e.g., object avoidance during movements [Park et al. 2008]. The main drawback of DMPs is the increased learning difficulty as the complexity of the controlled plant rises, as each DoF of the plant has to be controlled by an independent control policy [Havoutis 2007, Bitzer and Vijayakumar 2009]. For our use case, we might be able to model the motion using tracking data, but for the RGB image, as well as the EMG recording, the DMP approach is inappropriate.

Table 8.1 provides a short overview of the introduced motion modeling approaches. Beware that each of these methods can only handle single modalities at a time.

### 8.2.2 Different Spaces to Represent Motion

Very often it makes sense to model data in a different space than the one the data is recorded in; this is especially true for data from sparse or complex sensors such as cameras. An overview of the terminology for spaces we discuss is given in Table 8.2.

**Table 8.1** Selected methods to model motion from data

Approach	Short description	Pros	Cons
(constraint-based) optimization	model the motion data using, e.g., equations of dynamics; additional constraints can be included in the optimization task	physically meaningful	complex definition of models, large computational efforts, generation of new motions is not possible
motion graphs	recorded motion is split into subsequences; different transitions can be defined	easy to set up	limited to recorded motion, definition of transition thresholds
demonstrations	a recording of a motion	easy to set up	predefined start and end positions, fixed time frame
DMP	nonlinear dynamic system describing motion of a single DoF	smooth, easy to adapt to new conditions	learning difficulty rises with the number of DoF of the model

An observation contains all accessible information about a system. It is usually augmented with noise. A specific task, in contrast, is given in *task space*. The *task space* describes what is to be done where. The *task space*, e.g., can be the Cartesian space of the end effector. However, if a task is to be executed, e.g., by a robot, the instructions will not be provided in Euclidean coordinates, at least not internally. The commands may contain specific joint configurations that are to be attained; this is referred to as *state space*.

In order to instruct an articulated object to go to a specific configuration, typically higher-dimensional space descriptions addressing each joint are required. Nevertheless, likely there is a space that has fewer dimensions, yet can explain all the variance in the original space. This intrinsic dimensionality is called the *latent space*. High-dimensional and noisy sensor data of a motion recording will be influenced by just a few factors [Havoutis 2007, Wang et al. 2012], even if comprising dynamic human behavior [Safanova et al. 2004]. Factors are independent of each other and account for the variance in the data. For example, two factors suffice to describe an image sequence of a swinging pendulum: angle and angular velocity [Karl et al. 2016]. In the following these latent dimensions will often be referred to as *latents*.

It is possible to transform one space into another. This transformation needs not be unique, it can be difficult to identify, and it can be highly nonlinear. To illustrate this, data recorded with motion capture technology is high-dimensional, and the relationship between the marker positions to the joint angles of the articulated

**Table 8.2** Overview of different spaces

Name	Short description
task space	what is to be done where
state space	information with respect to states of a system, e.g., joint configurations
latent space	low-dimensional description of original data
input space	data before processing
output space	data after processing
hidden space	intermediate representations of the data during processing
feature space	calculated features of the data

object is nonlinear and governed by sine and cosine functions [Taylor et al. 2007]. Often methods of machine learning are applied to find this transformation. Then the initial data, i.e., the observation, is called the *input space*, while anything that the method returns will be the *output space*. Possibly, the method needs additional internal representations of the data, referred to as the *hidden space*. The *latent space* is a subgroup of the *hidden space*, yet in the context of this text we will use *latent space* only for very compact representations. Alternatively, one can use expert-designed features to create an easier to handle *feature space*. The quality of motion models depends on the data quality [Oviatt and Cohen 2015], and thus the space representation used.

In the previous section we introduced methods that handle recorded motion data directly. In the following section we will describe methods that combine the identification of a latent representation and modelling the motion therein.

### 8.2.3 Motion Modeling in Latent Space

In many cases, the data from which motion models are inferred are high-dimensional: video data being in that category; marker-based optical tracking of full human body motion, with typically over 50 DoF, is problematic for many approaches; even tracking of single finger phalanx movement can prove to be too high-dimensional.

Moreover, in most cases the intrinsic dimensionality of the movement is much lower (cf. Section 8.2.2). Dimensionality reduction techniques can reduce the high-dimensional state space to a lower-dimensional latent space, which may not have

a direct physical interpretation, but may be better suited for movement recognition or prediction—in general, for modeling. Some applicable methods for low-dimensional motion modeling are summarized in Table 8.3.

To reduce the dimensionality of raw data, it is common practice to calculate expert-designed features, like spatial filters in image processing or bandpass-filtered data for EMG. Typically, these features are fast to compute, greatly reduce

**Table 8.3** Selected methods to model motion from data in lower-dimensional space

Approach	Short description	Pros	Cons
expert-designed features	calculation of predefined features from data	easy to calculate, known relation to input data	possibly discarding relevant data, diverse feature properties, additional mapping to motion model required
constraint-based optimization after PCA	reduction of the input space using PCA	closed-form solution	only linear
GPLVM	reduction of the input space using GPs; can be extended by including DMPs to the latent space	generalizes well with little training data, high performance for static data	limited to smaller data sets, mainly for static data, assumes data to be independent, does not consider temporal continuity, sensitive to the initial guess
GPDM	reduction of the input space with dynamics and observation mappings using GPs	generalizes well with little training data, capable of filling in missing information	limited to smaller data sets, sensitive to the initial guess
AE-DMP	combination of autoencoder to reduce the dimensionality and DMPs for each latent dimension to smoothly model the motion	generative, robust to noisy and corrupted input, reconstruction of missing information possible, capable to handle large data sets	hyper-parameters have to be tuned
bottleneck CNN	model the image data using a CNN and mapping this to a latent space	very good at image processing	only applicable to data where translational variance does not encode relevant information

the number of dimensions, and extract information with an easy-to-understand relation to the raw data. But this can also lead to removal of relevant information. Using expert-designed features, only information which the expert considers relevant survives the feature calculation process. For instance, the number of zero crossings in EMG data provides some references to the frequency of the signal, but completely discards information about the signal amplitude. To overcome this issue, it is possible to combine multiple features. First, this increases the dimensionality of the feature space. Second, it does not ensure that all helpful information is retained: the expert will consider only attributes relevant to their perception, while they might be unaware of other factors influencing the data. Last, the properties of the different features will differ, similar to the differences between modalities.

On the contrary, *principal component analysis (PCA)* is a data-driven methodology to project a set of data onto a new basis, such that the largest variance of the data is along the first axis, the next largest variance on the second axis, etc. The projection is linear and can be implemented by eigenvalue decomposition of the data covariances. PCA can be used for dimensionality reduction: since the data is transformed to have its principal components sorted along the basis axes, discarding later dimensions removes less relevant information. The remaining dimensions form the latent space of PCA. This method was combined with constraint-based optimization by [Safonova et al. \[2004\]](#) to generate natural-looking and physically feasible motion for human movement with 60 DoF. As apparent in Figure 8.5, for the example introduced in Section 8.1.5, PCA is not sufficient to reduce the EMG data of our example to a processable dimensionality.

Other approaches are based on GPs [[Rasmussen and Williams 2005](#)]. A GP is a stochastic process given by its mean  $m(\mathbf{x})$  and covariance  $k(\mathbf{x}, \mathbf{x}')$  functions:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \quad (8.1)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \quad (8.2)$$

Without loss of generality, we assume the GP has a zero mean function. In GPs the *squared exponential covariance (SE)* is widely used for  $k$ . SE is a smooth function and measures the closeness or similarity of the inputs. GPs assume closely located inputs to behave similarly. Thus, test inputs have similar target values as the target values of close training inputs. The hyper-parameters of the GP can be obtained by maximizing the log-likelihood function under the training data.

Similar to PCA, Gaussian process latent variable model (GPLVM) [[Lawrence 2003](#), [Lawrence 2005](#), [Titsias and Lawrence 2010](#)] is an unsupervised learning

method for learning low-dimensional representations of the data. A GP smoothly maps the latent space to the observations. After an initialization with, e.g., PCA, it optimizes its hyper-parameters by maximizing the likelihood of the GPs with respect to the latent representation. For mapping new data to the latent space, the latent representation of the novel observed data has to be optimized. GPLVM is able to autonomously determine the dimensionality of the latent variables. Both GP and GPLVM have high performance for static data. They assume that data are independent, and they do not consider temporal continuity of data.

Bitzer et al. [2008] reduce the dimensions of static movements via GPLVM and model the movements using DMPs (see Section 8.2.1), enabling accurate pursuit of the desired trajectory. This approach is reasonable especially for the tracking data from the use case. However, this method does not provide for merging the data before the decision process. Besides, the quality of the latent representation is highly dependent on the initial guess. And as the GPs need to process all training data during evaluation large data sets including a multitude of motions will be prohibitively large.

*Gaussian process dynamical model (GPDM)* [Wang et al. 2008] extends GPLVM by mapping the data from observation space to a nonlinear dynamic system in latent space. In contrast to GPLVM applicable to any kind of data, GPDM is specifically designed to handle data with chronological structure. Besides mapping from latent space to observation space, GPDM maps the latent values at previous time steps to the latent values at the current time step, thereby ensuring a smooth transition between latent positions. Both mappings are based on GPs. GPDM generalizes well and can be trained from small data sets. Applied to human motion, it is able to model 50D data in 3D latent space and fill in missing frames of a motion. When applying GPDM to the use case, similar to GPLVM, it is highly suitable for tracking data. Analogously, the same limitations hold.

Quirion et al. [2008] provide a good overview of the GPLVM family and contrast the performance of the different implementations. The main drawbacks of these models are their sensitivity to the initial guess and the limitation to small data sets. Different from the GP-based models which store all training data, the following *deep-learning* models are parametric and able to train on large data sets.

Chen et al. [2015] extend deep neural networks in the form of an autoencoder (AE) (for more details refer to Section 8.3.1.1) with DMP. The resulting model is called AE-DMP. The DMPs feed into the latent space of the AE, which is then transferred to the state space by the decoder part of the AE. The architecture can be trained as one system. Sparse activation of the latent-layer neurons is encouraged to improve movement representation; therefore, various individual movements

can be observed in the latent space. AE-DMP can generate new movements which are not shown in the training data set by switching on/off or interpolating latent units. In addition, benefiting from the robustness of denoising AE with respect to corrupted input data, it can reconstruct movements from missing joints, e.g., related to sensor problems. In contrast to GP models, AEs can handle large data sets but require manual tuning of the hyper-parameters. Nevertheless, AE-DMP is also highly suitable for processing tracking data, and likely applicable to EMG and RGB recordings. The main benefit compared to the GP-based approaches is that AE-DMP is capable of handling very large data sets, as only the parameters are required after training, rather than the entire training data set.

*Convolutional neural network (CNN)* architecture is widely used for image recognition, and consequently can be used for modeling the movement from videos. It is robust to shifts, scales, and distortions of the input data and can be trained efficiently on large data sets. For instance, Oberweger et al. [2015] use a CNN with an additional bottleneck layer—a layer that has fewer dimensions than the input space and thus forces a latent space representation—to model finger joint locations from RGB-D data. The CNN architecture ensures suitable processing of the image data, while the *latent representation* ensures efficient calculations and improved generalization. CNN are well-suited for image data, like our RGB recording. However, they generate a translational-invariant representation of the data and thus are not suitable for the tracking data. Besides, CNNs in their basic form are discriminative rather than a generative model and prohibit the generation of new data samples.

This list of methods is by no means all-encompassing, and many extensions to the basic methods have been omitted for clarity. However, they provide a good introduction to typical approaches for motion modeling and help to illustrate the current drawbacks therein.

#### **8.2.4 Late Multimodal Fusion**

The reliability and robustness of motion models can be improved if combining several input modalities. In motion modeling, typically late fusion is used. For each modality a decision is learned. This decision can be discrete, like yes or no, a score, or a probability [Niaz and Merialdo 2013]. The decisions from the different modalities are combined using classifiers and rules. For instance, late fusion may average the confidence scores of different modalities. Since modalities are handled independently, new modalities can be incorporated by only updating the decision merging.

Late fusion has been performed to combine images, audio, or text [Snoek et al. 2005, Ye et al. 2012, Goutsu et al. 2015]. Kahou et al. [2013] train different deep-learning methods on audio and video frames to extract different features describing a subject's emotional state. Subsequently, they successfully integrate the modalities to predict facial expressions by averaging the prediction using different means of machine learning.

Any of the previously introduced methods is applicable for combination with late fusion in order to fuse multiple modalities. However, as exemplified using the use-case data, not each of these methods is appropriate for all data types. Thus, a careful selection of the respective motion modeling strategies, also considering the different model outputs, is imperative. For the use case, we might choose an AEDMP for the EMG and tracking data, and a modified CNN [Oberweger et al. 2015] for the video.

Up-to-date multimodal motion analysis mainly focuses on classification of motions and gestures, rather than continuous motion representation [Goutsu et al. 2015].

## 8.3

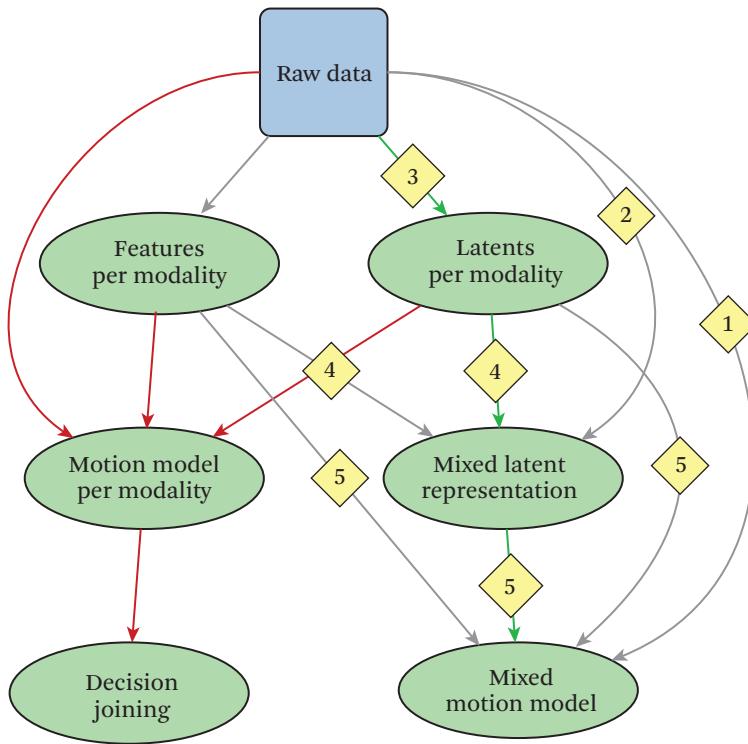
### Early Multimodal Integration for Motion Modeling

Ross and Govindarajan [2005] stated that fusion at the feature level is an under-studied problem, compared to fusion at the decision level. This imbalance is still in place today. Fusion at the sensor level is even less explored.

As depicted in Figure 8.3, the design of modality-fusing algorithms can be very diverse. The three paths building separate models for each modality (red) belong to late fusion approaches and are not further considered here. In the following we will discuss possibilities to employ the numbered paths, which can be considered early fusion. First, we specify implementation details for reducing the data dimensionality and generating the motion model in Section 8.3.1. Based on these algorithms, in Section 8.3.2 we describe the general approach to early integration. Section 8.3.3 concludes this section discussing the effects of recent advances in deep-learning on multimodal motion modeling. Finally, we apply the methodology to the use-case example introduced in Section 8.1.5 in Section 8.4.

#### 8.3.1 Applicable Algorithms for Early Fusion Approaches

In the following, we will describe an early-integration pipeline. For this, two types of algorithms are relevant: algorithms for calculating a latent representation and algorithms for motion modeling. We introduce our suggested realizations here, to make it easier to understand the processing setup.



**Figure 8.3** Different ways to fuse multimodal data for motion modeling. Red arrows indicate late fusion approaches. Numbered arrows refer to early fusion approaches introduced in Section 8.3. The green arrows are favorable.

### 8.3.1.1 Latent Representation for Early Fusion

Since the identification of latent models is an unsupervised task that is independent of the data at hand, we do not need to differentiate between representations of individual modalities or a mixture thereof. The main difference is the dimensionality of the input data and the level of correlation within the data. Rather than impacting the algorithmic setup, this affects the number of required layers and the number of units to represent the data in each layer adequately. These numbers can be specified as parameters for the algorithm and hardly influence the underlying structure.

Given data  $D$  with dimensionality  $d$  and a latent representation of this data  $L$  with dimensionality  $l \leq d$ , any data compressing algorithm can be used to calculate such a model. However, there are a few desirable attributes:

- ability to model nonlinearities;
- ability to reconstruct input;
- efficiency; and
- generative nature of the model.

As the relation between different spaces describing motion can be highly nonlinear, the ability to model nonlinearities is important. Accordingly, all linear compression methods can be ruled out.

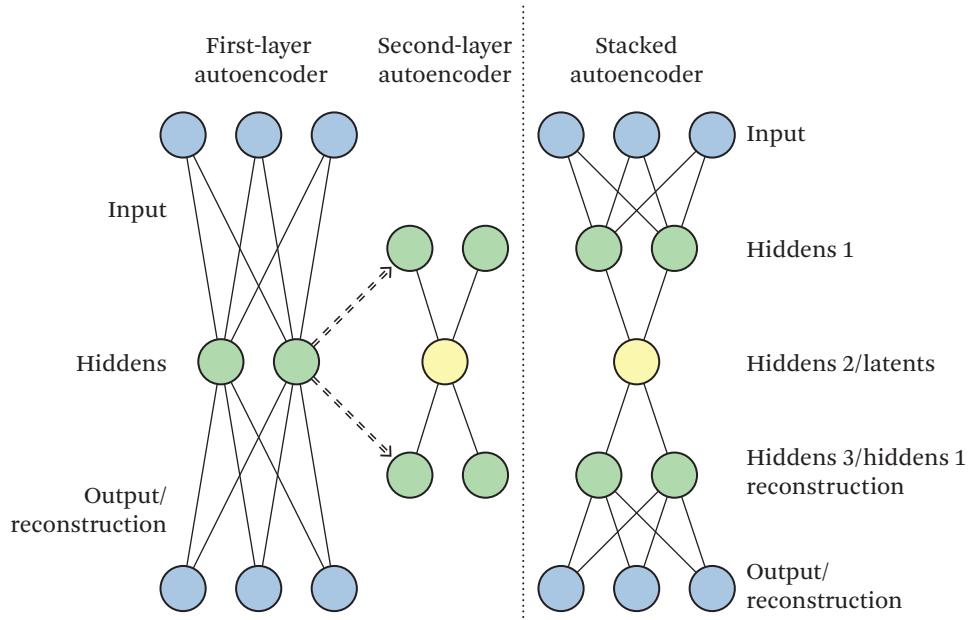
[Ngiam et al. \[2012\]](#) use restricted Boltzmann machines to model the latent space. A restricted Boltzmann machine has one layer of visible units, the inputs, and a set of hidden units, the model. Each hidden unit is connected to each visible unit, but there are no connections within a layer. Using contrastive divergence the model can efficiently learn a probability distribution over the input. While this is an appropriate method to generate a latent representation of the data, it requires extensive sampling from the data for training. Besides, these methods typically rely on binary latent variables. It is questionable whether they pose a reasonable prior for human or humanoid kinematics.

Another alternative would be the GPLVM [[Lawrence 2005](#)] or GPDM [[Wang et al. 2008](#)] already introduced in Section 8.2.3. The drawback of GP-based approaches is the increase in computational effort with high dimensions and large data sets.

Intending to reconstruct at least some of the modalities from the input data, a latent representation that can easily be remapped to the initial data space is helpful. AE learn compression and decompression of the data in a single step. An AE is basically a neural network with two parts, one that maps the input data to a lower-dimensional space and one that maps the latent representation back to the original space. Nonlinearity is ensured by applying nonlinear transfer functions. As shown in Figure 8.4, by stacking multiple AEs efficient training of highly nonlinear mappings can be achieved.

[Droniou et al. \[2015\]](#), aiming at good classification, add an unsupervised learning layer to the AE setup to identify different clusters in the data and to learn distinct manifold representations thereof. However, this approach is only reasonable if the data represents categorical information. Besides, this architecture inherently interferes with the ability to transit between different motions as it does not provide a shared latent representation.

A more powerful model than the AE, the variational autoencoder (VAE), was developed by applying variational inference to the AE [[Kingma and Welling 2013](#), [Rezende et al. 2014](#)]. Each latent dimension of the VAE tries to represent independent factors of the underlying data structure. If the predefined number of latent



**Figure 8.4** A stacked AE for highly nonlinear mapping of input to latent space. The hiddens of the initial single-layered AE can be used as input and output space of a second AE to compress the data further. The two AEs can be stacked to generate a deeper AE. Additional finetuning for the stacked setup is possible.

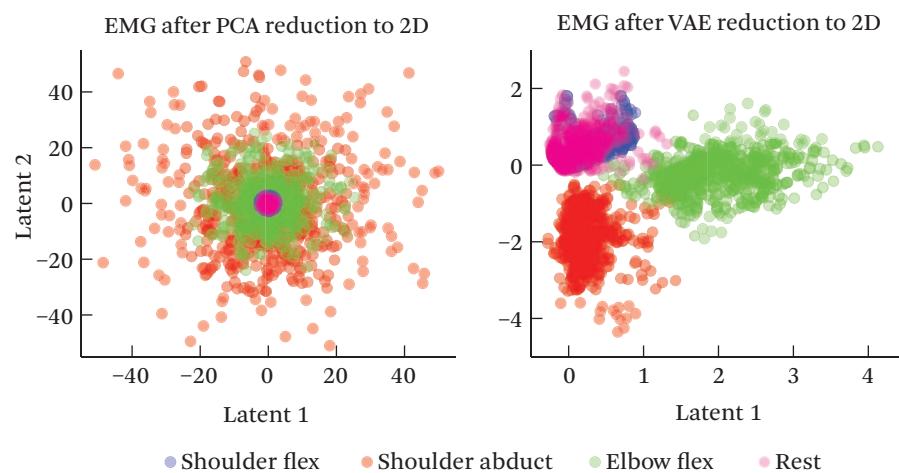
dimensions surpasses the number of required latents, the algorithm will disregard the superfluous dimensions and focus on the relevant ones. It is feasible to set the number of latent dimensions higher than the expectation of what is needed, learn the model, and then check which latent dimensions are really employed and omit the ones that are irrelevant to guarantee that all data are accounted for. Moreover, VAE are generative and thus facilitate the generation of new data that coheres to the recorded data. While Kingma and Welling [2013] describe the basic approach, there are numerous extensions to it, that should be considered to achieve optimal performance, e.g., enhancements for the training routine [Rezende and Mohamed 2015, Sønderby et al. 2016], and application to time series, such that smooth transitions between different states are ensured [Bayer and Osendorfer 2014, Chung et al. 2015, Karl et al. 2016].

Chen et al. [2016] embed DMPs into the latent space of a time-dependent variational autoencoder [Karl et al. 2016] framework (VAE-DMP). This method represents high-dimensional human/robot movements in a low-dimensional latent

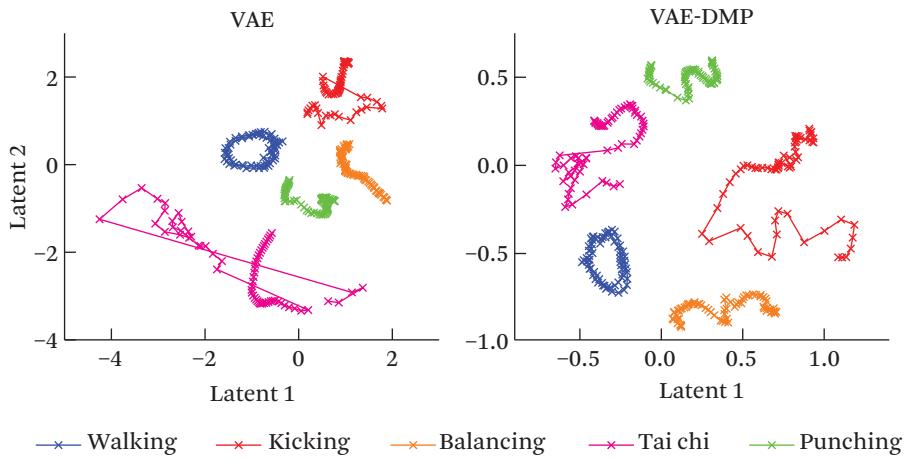
space. Experimental results demonstrate that the framework has excellent generalization in the latent space, e.g., switching between movements or changing goals. Also, the reconstructed movements outperform plain DMPs.

Using a latent representation based on variational inference, the latent space is a distribution rather than specific points. Accordingly, the system is self-aware of its confidence and if merging multiple modalities this additional information can be exploited to weight the modalities differently. This way the reliability of the model can further be improved and a failing modality hardly impairs the model.

Figure 8.5 contrasts dimensionality reduction of EMG data as described in Section 8.1.5 using PCA and VAE. The input space is 400D. For VAE, we use two layers of hidden units for reconstruction and compression, each. Each of these 4 layers has 300 units (dimensions). The latent dimensionality is two. Similarly, for PCA we only keep two of the principal components. To reconstruct at least 95% of data variance—a value typically chosen as a measure for an acceptable reconstruction—PCA needs 54 components. While PCA does not identify relevant attributes of the data in the latent space, VAE distributes the data according to the four different motion types within the recording.



**Figure 8.5** Comparison of dimensionality reduction using PCA and VAE. The input data is 400D EMG data. The latent data is 2D, each dimension is one of the plot axes. For easy interpretability, the data is color coded by one of four motion types: shoulder flexion, shoulder abduction, elbow flexion, and rest. Using PCA the different motion types overlap each other. Although VAE is not provided with any class information and only two latents are used, shoulder abduction and elbow flexion can be distinguished clearly.



**Figure 8.6** Comparison of five human movements in 2D latent space using VAE and VAE-DMP. The latent space of VAE-DMP is more meaningful than that of VAE. Since the latent space of VAE does not encode information about time, motion therein may spread far with different spacings between steps, especially for complex movements such as tai chi. Big gaps between two steps may cause difficulties for DMPs. In addition, kicking is a large range movement. In VAE-DMP latent space it covers a large area, while it only has a relative small range in VAE latent space.

Figure 8.6 compares the latent representations of a plain VAE and a VAE-DMP modeling five different full-body motions. This illustrates the better generalization of VAE-DMP with respect to the time domain.

Since VAE accomplish all of the desired attributes mentioned above, we highly recommend the use thereof for multimodal motion modeling. Additionally, there exist CNN-based implementations of VAE [Kulkarni et al. 2015], which may be of interest, especially for image-like data.

### 8.3.1.2 Motion Models for Early Fusion

Although the techniques for motion modeling introduced in Section 8.2.1 and Section 8.2.3 were not initially designed for early integration as described below, they are applicable. Instead of building the motion model from raw or featured data or from the latent representation of a single modality, the same methods are applied to the mixed latent representation.

Reusing well-known modeling strategies, experience gained over years of research on suitable motion models is exploited and the additional implementation expenses for an implementation of multimodal integration in motion modeling is

kept at a minimum. Further, as introduced by Chen et al. [2015], 2016, it is possible to incorporate these motion modeling techniques into the latent space identification and thereby improve the overall result.

### 8.3.2 Layout of Applicable Early Fusion Approaches

In the following we describe how to design an early fusion setup. Section 8.4 concretizes these ideas by providing an early fusion design for the use case introduced in Section 8.1.5.

#### 8.3.2.1 Merging Raw Data for Modeling

The ostensible solution to early fusion is to merge all sensor data at once and to subsequently build a motion model, as described in Section 8.2.1 from this large data set. In Figure 8.3 this is depicted by arrow 1. For the use case this would imply to concatenate the samples of RGB, EMG, and tracking data for each time step  $t$ , resulting in a vector with dimensionality 922,035, and to train a single motion model.

However, this is likely intractable. Different modalities will have different properties and will encode the motion parameters by different means [Srivastava and Salakhutdinov 2014, Ross and Govindarajan 2005]. In the example, the majority of dimensions will stem from the RGB data, thus it is likely that the information encoded in this modality will dominate the motion model, although not all of the motion will be observable from a single perspective, e.g., due to self-occlusion.

Moreover, data from different modalities typically has very different scales. For example, RGB data might range from 0 to 255, while EMG data might range from  $-10,000 \mu\text{V}$  to  $10,000 \mu\text{V}$ . Accordingly, EMG data variances are larger and thus may dominate. The differences can be alleviated by scaling to similar ranges, e.g., with norming to zero mean and unit variance. Beware however, that each sensor likely should be scaled independently. For example, it is reasonable to scale each EMG electrode individually, as the amplitudes of the signal are location dependent.

In summary, the model has to handle very large data with differently weighted modalities. And, the model will have to identify relevant information from raw and irregular data.

#### 8.3.2.2 Generating a Latent Representation of the Data

We conclude that it is important to generate a better representation of the contained information before calculating a motion model (arrow 2, Figure 8.3). This can be achieved by generating a latent representation (cf. Section 8.2.2) of the motion data having a much lower dimensionality. Thus, for the concatenated data

from above, we will identify a latent representation with, e.g.,  $l = 20$ , which in turn will be used to learn the motion model.

Using a latent representation, the modeling algorithm will only have to account for the variance of the reduced dimensions. On the one hand, noise contained in the raw data will be removed by compressing multiple dimensions. On the other hand, interdependencies between the different raw dimensions will already be identified by the latent representation and the motion model will only represent the motion at a more fundamental level. In order to find a reasonable latent representation, the model has to be sufficiently nonlinear and it has to be able to handle the different properties of the various modality types. Likely, the latent model will find correlations within single modalities more easily than across modalities [Ngiam et al. 2012, Srivastava and Salakhutdinov 2012, Srivastava and Salakhutdinov 2014]. Therefore, we expect a nonlinear deep-learning architecture to be a good choice for finding a suitable latent representation of the data. While it may be possible to approximate data of small motions with a linear latent model, exploitation of the entire workspace of a human arm will require greater flexibility, as the dependency between shoulder motion and hand position are highly nonlinear.

While we consider the calculation of a latent representation of the input data inevitable before learning the motion model, it does not imply that two different learning stages have to be applied (cf. Section 8.3.1.1).

### **8.3.2.3 Latent Representations for Each Modality**

Nevertheless, the model required to find a latent representation is large and differences in the properties of the modalities will not be considered in the latent model. Inspired by Ngiam et al. [2012], Srivastava and Salakhutdinov [2012], Srivastava and Salakhutdinov [2014], Sohn et al. [2014], and Neverova et al. [2015], we propose the prior calculation of latent representations of each modality (arrow 3 in Figure 8.3). So, in the use case we generate a latent representation for each of the different data streams separately.

Calculating the latent space for an individual modality has the advantage of similar statistical properties across dimensions. The number of required latent dimensions for a good representation may differ between modalities, as each modality describes different aspects of the motion. Even if two RGB streams of the same motion are used as input, the number of latent dimensions between the two streams may differ as neither of the two will represent the motion to its full extent. The stronger the limitation of representation, the fewer dimensions will be required.

Besides, a single modality can be processed at several resolutions. Exemplarily, Neverova et al. [2015] split video data into different areas of interest, leading to different spatial resolutions. While the suggested partitioning to separate geometric

areas seems superfluous due to the automatic detection of different change driving factors, separate areas for different spatial resolutions can be reasonable. For instance, if the smaller motion is in the order of magnitude of noise of the larger motion. Each resolution can be provided as an individual modality. Transferring this idea to the EMG data of our use case example, the data could be bundled in blocks of 10, 50, and 200 time points per sample. Including each of the different data stacks as single modalities may help to identify short-term and long-term effects.

Moreover, individual representations offer the advantage of being able to include unlabeled data for learning the latent model [Srivastava and Salakhutdinov 2014, Ross and Govindarajan 2005]. As unsupervised algorithms learning the latent representation do not need annotated data, any data representing the same modality under reasonable conditions can be used to improve its descriptive power. Motion capture (MoCap) data of different subjects can be used for learning a latent representation, given that the markers are attached to the same body parts. Additional unlabeled data may already be available online or can be retrieved more easily than labeled data.

### 8.3.2.4 A Two-layered Latent Representation

Although these representations will be much lower dimensional than the input data, we still expect that it is helpful to learn a mixed latent representation from these individual representations (Figure 8.3, arrows numbered 4). Referring to the example introduced above, we take the latent representations of RGB, EMG, and tracking data and concatenate these lower-dimensional representations in order to learn a combined latent representation.

As stated above, only using a mixed latent representation, correlations between modalities will be detected. Using unsupervised algorithms allows for autonomous identification of the relationships between the latent spaces, although these are likely unknown before.

Adding a mixed latent representation on top, information that might be useless when coming from a single modality might be decoded when combining different modalities. An example of this is stereo vision: Observing a motion from two different perspectives in an image sequence, the sensor readings are projections of the true trajectory; combining both sources it is possible to recover the true direction of motion.

The mixed latent representation is expected to have at least the same dimensionality as the largest of the modalities. Otherwise even information of this single representation has to be discarded. However, it is very probable that the combined

latent representation has fewer dimensions than the sum of latent dimensions, as corresponding information is encoded in different modalities.

Contrary to plain feature-level-fusion approaches, that require an additional measure for feature selection to avoid the curse of dimensionality [Bellman 1957, Bellman 2003] and to be able to handle unbalanced features sets [Ross and Govindarajan 2005, Planet and Iriondo 2012], using latent representations of the data at two levels, this can be automatized. In addition, if the same technique for retrieving the latent representation is used, the statistical properties of the latent spaces will be comparable across modalities, unlike raw data.

While it is possible to use the same algorithm for calculating the latent representation of each modality and the mixed representation, the representations can be learned in separate steps. Using this separation, the problem referred to as vanishing gradients is handled similarly to Schmidhuber [1992].

In theory, it is possible to use a deeper model to calculate the mixed latent representation directly from raw data. For example, Droniou et al. [2015] follow this approach by sharing all parameters across all modalities. However, in the lower levels the majority of identified correlations will be within single modalities [Srivastava and Salakhutdinov 2014]. So calculating the full structure is a waste of resources.

The applications introduced by Ngiam et al. [2012] and Srivastava and Salakhutdinov [2014] and related publications only use a single multimodal layer and one to two unimodal layers. It is reasonable to assume that a deeper multimodal model would provide better capabilities to model nonlinear relationships between the different modalities. Lin and Tegmark [2016] support this assumption by showing that deeper architectures can model complex relationships more efficiently.

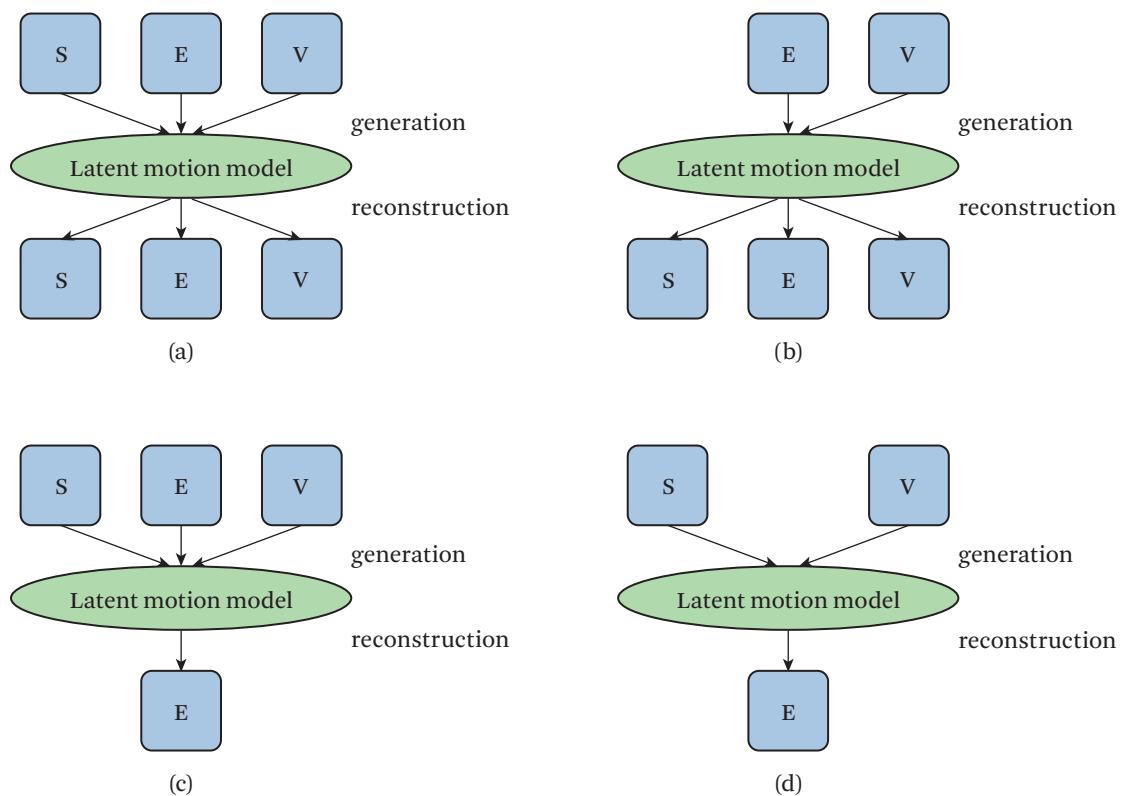
To finally learn the motion model (arrows numbered 5, Figure 8.3), the methods introduced in Sections 8.2.1 and 8.2.3 can be applied. The main requirement is to find a continuous representation of the motion that can be adjusted easily. For example, the motion model should allow parametrization for different time scales or start/end positions, like DMPs. For the mixed latent representation it is also possible, to use a model that directly incorporates the motion model (cf. Section 8.3.1.1).

### 8.3.2.5 Multimodal Synopsis

Besides being able to better identify correlations between modalities if using a mixed latent representation, this also facilitates the reconstruction of missing input modalities. Using a method referred to as ModDrop in Neverova et al. [2015], the model learns to handle corrupted input. Although not explicitly named, Ngiam et al. [2012] and Srivastava and Salakhutdinov [2014] also apply this approach. It is

based on *dropout* [Srivastava et al. 2014] and leaves out input from single modalities during training. This way, the model is forced to generalize sufficiently to be able to handle missing information. This approach can be extended to omitting multiple channels at the same time. Only due to the mixed latent representation the model is able to infer the lacking information. Going one step further, the described approach can also be used to reconstruct one or several of the input modalities, to, for example, predict the tracking marker data from the remaining sources. In this case the calculated motion model is mapped back to the dropped modalities.

Figure 8.7 displays the different setups profiting from multimodal fusion.



**Figure 8.7** Different setups benefiting from multimodal fusion in motion modeling. V, E, and S indicate different source types, e.g., video, EMG, and sound. (a) Multimodal fusion with reconstruction of input sources. (b) Multimodal synopsis with reconstruction of unavailable source. (c) Multimodal synopsis with additional input to improve model. (d) Multimodal synopsis with inference of unavailable source.

Even if a modality is only available during training, we suggest to include that modality as input to be able to learn a mixed representation and benefit from the correlation between inputs. To ensure generalization of the model, it is important that when dropping a modality we use data that has not been used to train the mixed latent model before. Otherwise, the results will be biased. For testing we have to use data that has not been used for training.

The same representation of motion can be used to achieve multiple tasks. In one setting it may be desired to identify a motion trajectory, in another task it may be more interesting to compare the effect of different gravity conditions on the prediction of EMG activity, and in a further task different types of motion like reaching, grasping, and retracting could be classified.

While each modality should be available at least part of the time and in different combinations with other modalities for training, for the final task execution it might suffice to know the EMG activations to predict the limb positions. Yet, switching to RGB and depth recordings to predict the motion type will neither require new training nor change any of the setup.

The accuracy, however, may differ when using different modalities as input, since each modality provides different amounts of motion information. Nevertheless, the overall model will likely be more accurate than when learning a model with currently needed modalities only, because multimodal training identifies intermodal correlations.

In addition, it is possible to combine the models of multiple individuals to create a general model of human motion. In that case each subject would be considered a modality. Comparing the individual models and the integrations thereof could provide an insight of what is normal. Further, automatic classification in different disease categories could be possible. If the subjects' data are recorded using multiple modalities, instead of a two-layered approach a three-layered approach might be helpful.

### 8.3.3 Multimodal Motion Modeling and Deep-learning

Sections 8.3.2 and 8.3.1 make clear that deep-learning is crucial for early fusion in multimodal motion modeling. Retrieving latent representations for individual modalities and the mixed model should be handled algorithmically, rather than by expert design. In order to be able to identify nonlinear correlations across features and modalities, the use of networks that span across several layers is helpful.

As described in Section 8.4, latent space calculation, feature extraction, and motion modeling blend into each other and distinguishing a feature extraction phase is not sensible. Similar to deep computer vision methods where several

convolutional layers are stacked with some fully connected layers for the final task (e.g., Lecun et al. [1998]), we also stack different neural networks on top of each other.

Lately, deep learning often is associated with the term *end-to-end learning*. End-to-end learning implies that we have a task and some data referring to the task. While we have means to assess the performance using the data, we do not use any expert knowledge to get from input data to the output. Instead, we only decide on a specific system architecture and let the system learn what will lead to success or failure.

We follow the same principles for early fusion. However, we have several inputs, and our objective is to find a task-independent motion model. To render the design task-specific, we would have to add an additional layer that applies the motion model to our task of choice. Focusing on a specific task may alleviate the burden on the feature and motion model as the system can focus on the factors that are relevant for a specific task. Nevertheless, an unrelated task may require other information. Thus, the task-specific model is less versatile.

In order to allow tasks that we learn about later to benefit from the available model yet being able to incorporate new information, Rusu et al. [2016] use progressive neural networks. In these, each task requires a new model. But, the new model has connections to the old ones and can learn how much of the available information it will incorporate into the new model. While this increases the model size tremendously, it prevents that new tasks will destroy what has been learned for other tasks if these are not going to be continuously trained.

Especially, if numerous tasks and modalities are available from the start, the early fusion design described here is beneficial. If too complex models should be avoided by any means, it is possible to train the entire setup with focus on all of the selected tasks. In case additional modalities become available much later or an entirely new task shall be included, mixing the existing early fusion model with progressive neural nets could be helpful.

Future advances in deep learning will directly affect early multimodal fusion. Reduced training times or less required training data will decrease the time required until we get satisfactory models of motion. Improved accuracies and better compression will enhance the capabilities of the models.

## 8.4

### Use Case Implementation

In Section 8.1.5 we introduce an example motion modeling problem to illustrate our arguments. In the following we will give some details on how to design an early

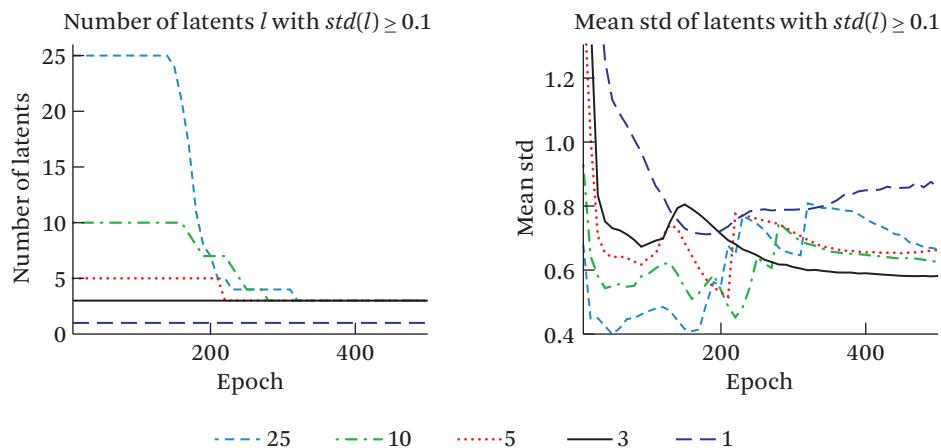
fusion system for this setup. The data describe motion of a human arm in space using an RGB stream, EMG recordings, and optical tracking. We want to find a model for this motion.

First of all, as the data stems from very different sources and has largely varying dimensionality, we choose a two-layered latent representation. The setup is depicted in Figure 8.10.

For each modality we first calculate the latent space individually using a VAE. The RGB video will be processed one frame at a time, since at each time step  $t$  we get one new sample. We treat each of the three color channels as an individual modality, thus processing them separately to keep the input dimensionality reasonable. The respective modalities are titled  $R$ ,  $G$ , and  $B$  in Figure 8.10. As translation invariance is important for image processing, we encourage the use of a convolutional VAE, although not required. In that case, it is possible to provide all color channels as a single input modality as the network size is reduced by architecture. A depth of at least one hidden layer for compression, followed by the latent representation and another hidden layer for reconstruction, is reasonable. The latent dimensionality should be much lower than the original dimensionality. The choice of dimensions should be validated by checking how much each dimension varies when the input changes after training. If all dimensions change significantly, it is possible, that we do not have enough units to represent the true latent dimensionality. In this case it is sensible to relearn the VAE with more latents and repeat the check. If only a part of the dimensions are relevant, we can reduce the network size to the respective size to ease the subsequent tasks. Figure 8.8 demonstrates that not all latent dimensions will exhibit notable variance if the latent dimensionality is higher than required. In contrast, in case too few latents are available—here only one—the variance is likely higher than when sufficiently many latents are available.

Also, for tracking data—called *tracking* in Figure 8.10—we have to validate the expected number of latent dimensions. It is sensible to use a plain VAE since the data are not translation invariant. Again, we suggest to use at least one hidden layer for each of the compression and reconstruction tasks, as well as another layer for the latent representation.

In contrast to per frame tracking and RGB data, we use windowed EMG data. While the former recordings are capable of conveying information about motion, respectively, current positions within a single frame, a single unfiltered EMG data point does not contain exploitable information. In Figure 8.10 these data windows are referred to as *EMG*. We suggest checking the expected latent dimensionality as before. Figure 8.8 displays how the latent space of EMG data develops during training. In contrast to pixels in an image where neighborhood implies that similar



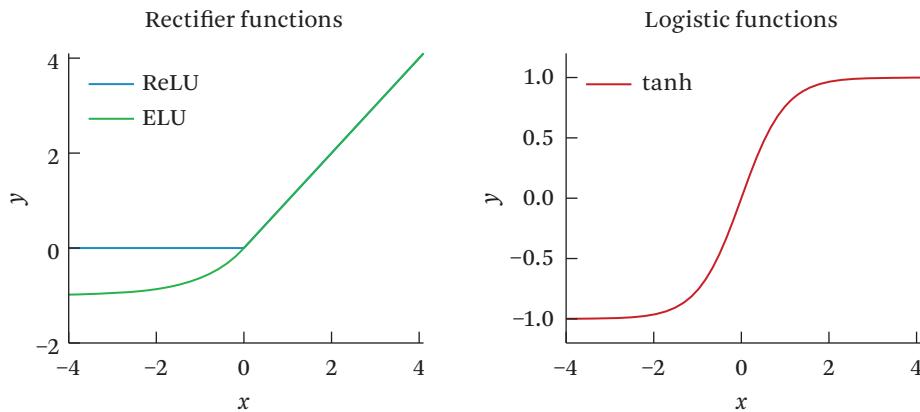
**Figure 8.8** Exemplary development of the latent space using VAE. The input data is 400D EMG data, as in Figure 8.5. Although the larger VAEs have more latent dimensions available, they spread the main variance across three dimensions.

information will be displayed, different electrodes may transmit very different information. Therefore, the use of 2D convolution is not adequate. The use of plain VAE or a 1D convolution are most reasonable. The VAE should have at least one hidden layer for compression and reconstruction, each, but more are possible.

After reducing the modalities with the previously calculated models, we obtain a much smaller number of total dimensions. We combine all the outputs and apply a VAE to this mixed data. Using a plain VAE is reasonable, since the translational invariant aspects of the motion should already be discovered within the modality specific latent representation. In order to allow the algorithm to identify the correlations between the different modalities, we suggest using at least two hidden layers on each side of the latent representation. As the motion of an arm with 7 DoF will likely not be driven by many more factors, we set the latent dimensionality of the final VAE to 30 dimensions. The size should at least account for all latents of the largest individual latent dimensionality and it should be validated as described before.

For all VAE we suggest using exponential linear units (ELU) or rectified linear units (ReLU) (Figure 8.9), as in practice they converge much faster than logistic functions, e.g., the tanh.

After training the six different VAE—five for the individual modalities, one mixed latent representation—separately, it is possible to fine-tune the parameters



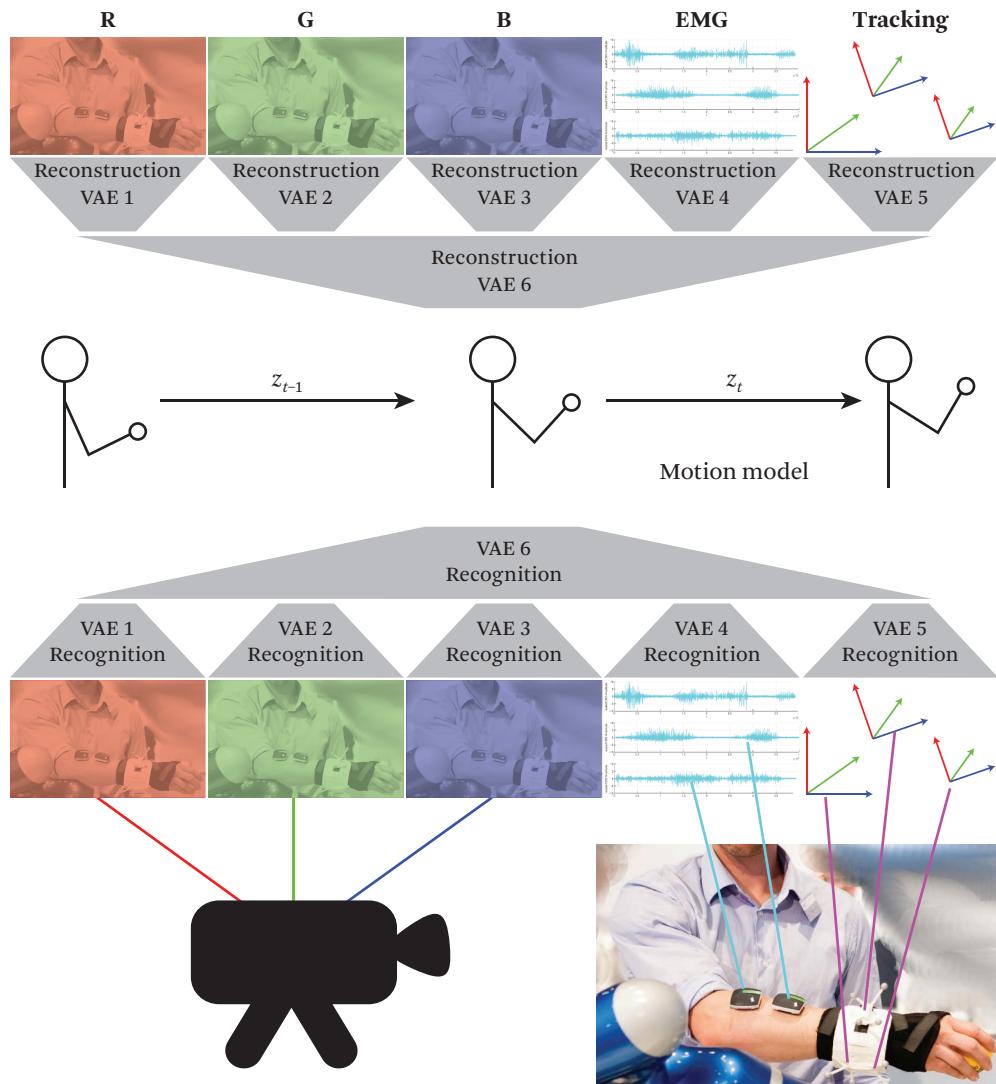
**Figure 8.9** Exemplary nonlinear transfer functions.

by combining the whole setup and training the stacked VAE as one. This way the reconstruction performance of the mixed model will also impact the individual representations and improve the coding and decoding strategies.

Finally, we are looking for a good representation of the motion between the latent representations  $\mathbf{z}_{t-1}$  and  $\mathbf{z}_t$ . For this purpose we rely on DMPs. We include the DMP calculation into the latent layer of the mixed model, as proposed by [Chen et al. \[2016\]](#). This way the objective of modeling smooth motions will also be taken into account while learning the reconstruction. Incorporating the motion modeling objective into learning of the latent dimensionality is beneficial, as it helps the system to discriminate relevant information. Otherwise, the model might waste factors on the representation of, for example, flickering light.

In detail, for the examples in Figures 8.5 and 8.8 with an input dimensionality of 400 we use a plain VAE. To compress the data to the latent space we use 2 hidden layers with 300 neurons each, and another 2 hidden layers with 300 neurons each for reconstructing the data. As transfer functions we use ReLU. The respective number of latents is indicated for each evaluation. For training the VAE we use the Adam optimizer [[Kingma and Ba 2014](#)].

The VAE architecture of the plain VAE and VAE-DMP in Figure 8.6 has 50 inputs, 200 hidden neurons for compressing, 2 latents, 200 hidden neurons for reconstruction, and 50 outputs. We use ReLU and identity activations for the hidden layers and the output layer, respectively. The DMPs are set to critical control. The hyperparameters of DMP and VAE are chosen based on the reconstruction error and the training time by grid search.



**Figure 8.10** An early fusion setup for learning arm motion from RGB video, EMG, and optical tracking at time step  $t$ . Each modality (R, G, B, EMG, and tracking data) is reduced by a separate VAE. The reduced modalities are concatenated and VAE 6 generates a mixed latent representation. In its latent space  $\mathbf{z}$  we model the motion using time series methods such as DMPs. Using the reconstruction layers of this setup the input data can be retrieved.

The full multimodal model including dimensionality reduction and motion modeling can now be used to solve very diverse tasks, for example:

- classify the motion into one of several different types;
- use the model to control a remotely located robot; and
- generate new motion patterns for a simulation.

## **8.5 Conclusion**

We present current techniques of motion modeling that are applicable to full multimodal data by reducing the dimensionality thereof and then learning a model of the motion. We show, that there are only few publications on multimodal fusion for motion modeling and that especially early fusion is an understudied problem.

The approach we introduce in Section 8.3.2 is flexible and can be applied to numerous multimodal settings. If deploying the full setup, each modality is reduced to its latent representation using an unsupervised learning algorithm. The individual latent representations are fused to a mixed latent representation and this is used to model the motion.

While training the mixed latent representation and motion model it is possible to dropout some modality data from the input. Nevertheless, the dropped modality can be used for gradient calculation in order to better learn correlations between modalities and get a more comprehensive understanding of the multimodal synopsis. Besides improving the accuracy of the motion model by using information of several sources, this later also allows for inference of missing modality information, and makes the model robust to modality failures. Using a generative latent model, it is possible to generate new data, for example for computer animation.

The ideas presented here are in part inspired by previous publications on multimodal deep-learning and fuse the setup with methods of variational inference and motion modeling. The most accurate results will be achieved if the different modalities observe the same behavior. It is possible to deploy the concepts to the observation of different behaviors that may or may not overlap in time, for example utterances and gestures of a person. However, the combinatorial space increases tremendously and before applying such an approach to multi-behavior observation it has to be evaluated whether this is beneficial.

Although this chapter focuses on motion modeling from available data, we have to consider how these models and the data we use can harm the subject's privacy. Friedland and Tschantz [2018] unveil how multimodal data can reveal information about our lives that we would rather not present to the world. Ensuring privacy

within the models is just as important and it has to be considered how the models can be prevented from malicious use.

### **Focus questions**

- 8.1.** What is motion modeling needed for and why is multimodal motion modeling beneficial?
- 8.2.** Why is early fusion desirable for motion modeling?
- 8.3.** How can the problem of the data with too large dimensionality in early fusion be overcome?
- 8.4.** Why is the suggested approach split into two levels?
- 8.5.** What are the advantages of learning a latent space representation compared to a traditional feature calculation?
- 8.6.** What are the benefits of using methods of variational inference for the identification of the latent space?
- 8.7.** Which methods are applicable to motion modeling, especially in the latent space?
- 8.8.** How does the introduced technique for multimodal motion modeling benefit from dropout?
- 8.9.** Why is the integration of the motion model into the identification of the latent space beneficial compared to identifying the latent space independently from the motion model?

### **References**

- F. C. Anderson and M. G. Pandy. 2001. Dynamic optimization of human walking. *Journal of Biomechanical Engineering*, (123): 381–390. DOI: [10.1115/1.1392310](https://doi.org/10.1115/1.1392310). 314
- J. Bayer and C. Osendorfer. 2014. Learning stochastic recurrent networks. In *Workshop on Advances in Variational Inference, Neural Information Processing Systems 2014*. <http://arxiv.org/abs/1411.7610>. 325
- R. E. Bellman. 2003. *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications. <https://books.google.de/books?id=fyVtp3EMxasC>. 310, 331
- R. Bellman. 1957. *Dynamic Programming*. Rand Corporation research study. Princeton University Press. <https://books.google.com/books?id=wdtoPwAACAAJ>. 310, 331
- S. Bitzer and S. Vijayakumar. 2009. Latent spaces for dynamic movement primitives. In *Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids'09)*, Paris, France. 315

- S. Bitzer, I. Havoutis, and S. Vijayakumar. 2008. Synthesising novel movements through latent space modulatoin of scalable control policies. In *Proceedings of the 10th International Conference on Simulation of Adaptive Behaviour (SAB'08)*, Osaka, Japan. 320
- N. Chen, J. Bayer, S. Urban, and P. van der Smagt. 2015. Efficient movement representation by embedding dynamic movement primitives in deep autoencoders. In *15th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2015*, Seoul, South Korea, November 3–5, pp. 434–440. DOI: [10.1109/HUMANOIDS.2015.7363570](https://doi.org/10.1109/HUMANOIDS.2015.7363570). 320, 328
- N. Chen, M. Karl, and P. van der Smagt. 2016. Dynamic movement primitives in latent space of time-dependent variational autoencoders. *EEE-RAS Conference on humanoid robots*. 325, 328, 337
- J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. 2015. A recurrent latent variable model for sequential data. *CoRR*, abs/1506.02216. <http://arxiv.org/abs/1506.02216>. 325
- S. K. D'Mello, N. Bosch, and H. Chen. 2018. Multimodal-multisensor affect detection. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA.
- A. Droniou, S. Ivaldi, and O. Sigaud. 2015. Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems*, 71: 83–98. <http://www.sciencedirect.com/science/article/pii/S0921889014002474>. DOI: [10.1016/j.robot.2014.11.005](https://doi.org/10.1016/j.robot.2014.11.005). 324, 331
- G. Friedland and M. Tschantz. 2018. Privacy concerns of multimodal sensor systems. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüeger, editors, *xThe Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 339
- Y. Goutsu, T. Kobayashi, J. Obara, I. Kusajima, K. Takeichi, W. Takano, and Y. Nakamura. 2015. Multimodal gesture recognition using integrated model of motion, audio and video. *Chinese Journal of Mechanical Engineering*, 28(4): 657–665. DOI: [10.3901/CJME.2015.0202.053](https://doi.org/10.3901/CJME.2015.0202.053). 322
- I. Havoutis. 2007. Scalable movement representation in low dimensional latent space. Master Thesis, University of Edinburgh. 314, 315, 316
- A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. February 2013. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural Computation*, 25(2): 328–373. ISSN 0899-7667. <http://europepmc.org/abstract/MED/23148415>. [10.1162/neco\\_a\\_00393](https://doi.org/10.1162/neco_a_00393). 315
- S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülc̄ehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin,

- N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 543–550, ACM New York. DOI: [10.1145/2522848.2531745](https://doi.org/10.1145/2522848.2531745). 322
- J. Kang, B. Badi, Y. Zhao, and D. K. Wright. 2006. Human motion modeling and simulation by anatomical approach. *WSEAS Transactions on Computers* 5(6): 1325–1332. <http://bura.brunel.ac.uk/handle/2438/1824>. DOI: [10.1.1.426.5826](https://doi.org/10.1.1.426.5826). 306, 314
- M. Karl, M. Soelch, J. Bayer, and P. van der Smagt. 2016. Deep variational Bayes filters: Unsupervised learning of state space models from raw data. *arxiv*. <http://arxiv.org/abs/1605.06432>. 316, 325
- E. Keogh and A. Mueen. 2010. Curse of Dimensionality. *Encyclopedia of Machine Learning*, pp. 257–258. Springer US, Boston, MA. DOI: [10.1007/978-0-387-30164-8\\_192](https://doi.org/10.1007/978-0-387-30164-8_192). 310
- D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. <http://arxiv.org/abs/1412.6980>. 337
- D. P. Kingma and M. Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114. <http://arxiv.org/abs/1312.6114>. 324, 325
- E. A. Kirchner, S. H. Fairclough, and F. Kirchner. 2018. Embedded Multimodal Interfaces in Robotics: Applications, Future Trends, and Societal Implications. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 312
- L. Kovar, M. Gleicher, and F. Pighin. 2002. Motion graphs. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pp. 473–482, ACM New York. DOI: [10.1145/566570.566605](https://doi.org/10.1145/566570.566605). 314
- T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. 2015. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167. <http://arxiv.org/abs/1503.03167>. 327
- N. Lawrence. December 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6: 1783–1816. <http://dl.acm.org/citation.cfm?id=1046920.1194904>. 319, 324
- N. D. Lawrence. 2003. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, pp. 329–336. MIT Press. 319
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791). 334
- H. W. Lin and M. Tegmark. 2016. Why does deep and cheap learning work so well? *arXiv preprint arXiv:1608.08225v1*. DOI: [10.1007/s10955-017-1836-5](https://doi.org/10.1007/s10955-017-1836-5). 331

- N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. 2015. Moddrop: adaptive multimodal gesture recognition. *CoRR*, abs/1501.00102. <http://arxiv.org/abs/1501.00102>. 329, 331
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2012. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 689–696, Bellvue, WA. 324, 329, 331
- U. Niaz and B. Merialdo. July 2013. Fusion methods for multimodal indexing of web data. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pp. 1–4. DOI: [10.1109/WIAMIS.2013.6616129](https://doi.org/10.1109/WIAMIS.2013.6616129). 321
- M. Oberweger, P. Wohlhart, and V. Lepetit. 2015. Hands deep in deep learning for hand pose estimation. *CoRR*, abs/1502.06807. <http://arxiv.org/abs/1502.06807>. 321, 322
- S. Oviatt and P. R. Cohen. 2015. The paradigm shift to multimodality in contemporary computer interfaces. *Synthesis Lectures on Human-Centered Informatics*, 8(3): 1–243. DOI: [10.2200/S00636ED1V01Y201503HCI030](https://doi.org/10.2200/S00636ED1V01Y201503HCI030). 310, 311, 312, 317
- D.-H. Park, H. Hoffmann, P. Pastor, and S. Schaal. 2008. Movement reproduction and obstacle avoidance with dynamic movement primitives and potential fields. In *IEEE International Conference on Humanoid Robots*. DOI: [10.1109/ICHR.2008.4755937](https://doi.org/10.1109/ICHR.2008.4755937). 315
- S. Planet and I. Iriondo. June 2012. Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on*, pp. 1–6. 331
- S. Quirion, C. Duchesne, D. Laurendeau, and M. Marchand. 2008. Comparing GPLVM approaches for dimensionality reduction in character animation. *Journal of WSCG*, 16(1–3): 41–48. <http://dblp.uni-trier.de/db/journals/jwscg/jwscg16.html#QuirionDLM08>. 320
- C. E. Rasmussen and C. K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. 319
- D. J. Rezende and S. Mohamed. July 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, Lille, France, pp. 1530–1538. <http://jmlr.org/proceedings/papers/v37/rezende15.html>. 325
- D. J. Rezende, S. Mohamed, and D. Wierstra. 2014. Stochastic backpropagation and variational inference in deep latent Gaussian models. In *International Conference on Machine Learning*. 324
- A. Ross and R. Govindarajan. 2005. Feature level fusion using hand and face biometrics. In *Proceedings of SPIE Conference on Biometric Technology for Human Identification II*, vol. 5779, pp. 196–204, Orlando, USA. DOI: [10.1117/12.606093](https://doi.org/10.1117/12.606093). 308, 310, 311, 322, 328, 330, 331
- A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. 2016. Progressive neural networks. *CoRR*, abs/1606.04671. <http://arxiv.org/abs/1606.04671>. 334

- A. Safonova, J. K. Hodgins, and N. S. Pollard. 2004. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pp. 514–521. ACM New York. DOI: [10.1145/1186562.1015754](https://doi.org/10.1145/1186562.1015754). 314, 316, 319
- S. Schaal. 1997. Learning from demonstration. In *Advances in Neural Information Processing Systems 9*, pp. 1040–1046. MIT Press. 314
- S. Schaal. 2003. Dynamic movement primitives—a framework for motor control in humans and humanoid robots. In *The International Symposium on Adaptive Motion of Animals and Machines*. Tokyo. <http://www-clmc.usc.edu/publications/S/schaal-AMAM2003.pdf>. 315
- J Schmidhuber. March 1992. Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2): 234–242. ISSN 0899-7667. 10.1162/neco.1992.4.2.234. 331
- B. Schuller. 2018. Multimodal user state & trait recognition: An overview. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA.
- C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pp. 399–402. ACM New York. DOI: [10.1145/1101149.1101236](https://doi.org/10.1145/1101149.1101236). 308, 310, 322
- K. Sohn, W. Shang, and H. Lee. 2014. Improved multimodal deep learning with variation of information. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2141–2149. Curran Associates, Inc. <http://papers.nips.cc/paper/5279-improved-multimodal-deep-learning-with-variation-of-information.pdf>. 329
- D. Sonntag. 2018. Multimodal interaction for medical and health systems. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Publishers, San Rafael, CA. Morgan & Claypool Publishers, San Rafael, CA.
- N. Srivastava and R. Salakhutdinov. 2014. Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research*, 15: 2949–2980. <http://jmlr.org/papers/v15/srivastava14b.html>. 312, 328, 329, 330, 331
- N. Srivastava and R. R. Salakhutdinov. 2012. Multimodal learning with deep Boltzmann machines. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 2222–2230. Curran Associates, Inc. <http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf>. 329

- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15: 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>. 332
- C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. 2016. How to train deep variational autoencoders and probabilistic ladder networks. arXiv preprint. <http://arxiv.org/abs/1602.02282>. 325
- G. W. Taylor, G. E. Hinton, and S. T. Roweis. 2007. Modeling human motion using binary latent variables. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pp. 1345–1352. MIT Press. <http://papers.nips.cc/paper/3078-modeling-human-motion-using-binary-latent-variables.pdf>. 306, 310, 314, 317
- M. K. Titsias and N. D. Lawrence. May 2010. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010*, Chia Laguna Resort, Sardinia, Italy, pp. 844–851. <http://www.jmlr.org/proceedings/papers/v9/titsias10a.html>. 319
- A. Vinciarelli and A. Esposito. 2018. Multimodal analysis of social signals. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 310
- J. M. Wang, D. J. Fleet, and A. Hertzman. 2008. Gaussian process dynamical models for human motion. *Transactions on Pattern Recognition and Machine Intelligence*, 30(2): 283–298. DOI: [10.1109/TPAMI.2007.1167](https://doi.org/10.1109/TPAMI.2007.1167). 320, 324
- Z. Wang, M. Deisenroth, H. B. Amor, D. Vogt, B. Scholkopf, and J. Peters. July 2012. Probabilistic modeling of human movements for intention inference. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia. 316
- A. Witkin and M. Kass. Junw 1988. Spacetime constraints. *SIGGRAPH Computer Graphics*, 22(4): 159–168. ISSN 0097-8930. DOI: [10.1145/378456.378507](https://doi.org/10.1145/378456.378507). 314
- G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. 2012. Robust late fusion with rank minimization. In *CVPR*, pp. 3021–3028. IEEE Computer Society. 322





# **Standardized Representations and Markup Languages for Multimodal Interaction**

**Raj Tumuluri, Deborah Dahl, Fabio Paternò,  
Massimo Zancanaro**

## **9.1**

### **Introduction**

This chapter discusses some standard languages that can be used to specify multimodal systems and describes their benefits and their relationships to each other. These languages provide ways for multimodal components to communicate with each other, to control the behavior of components through scripting, and to represent multimodal inputs and outputs. The chapter focuses on standards that have been developed for voice and multimodal components by Working Groups at the World Wide Web Consortium (W3C), in particular the Voice Browser and Multimodal Interaction Working Groups.

In this chapter we first discuss the need of languages for describing multimodal interaction. We then introduce model-based languages for interactive applications and discuss how they can be exploited in multimodal user interfaces. We then continue on to introduce some languages developed to address specific aspects of various interaction modalities, followed by a description of how the interaction through multiple modalities can be specified and managed. Finally, we will discuss some challenges that are still open.

### Glossary

**Abstract User Interface (AUI).** User interface description in terms of elements that are independent of the possible interaction modalities.

**Application Programming Interface (API).** Set of procedures made available by a software application to provide services to external programs.

**Behavior Markup Language (BML).** An XML-based language for describing behaviors that should be realized by animated agents.

**Concrete User Interface (CUI).** User interface description in terms of elements that are modality-dependent but implementation language independent.

**Emotion Markup Language (EmotionML).** An XML markup language for describing emotion, standardized by the W3C Multimodal Interaction Working Group.

**Extensible Multimodal Annotation (EMMA).** An XML markup language for describing the results of multimodal processors such as speech recognition, image recognition, and natural language understanding, standardized by the W3C Multimodal Interaction Working Group.

**Hypertext Markup Language (HTML).** A graphical markup language for defining web pages.

**Hypertext Transfer Protocol (HTTP).** An application protocol commonly used in the World Wide Web. A protocol is a set of rules to exchange data among different applications. HTTP dictates as web browsers and other similar applications can access web content.

**Ink Markup Language (InkML).** An XML markup language for describing digital ink traces and their properties, standardized by the W3C Multimodal Interaction Working Group.

**Interaction Manager (IM).** A component in the W3C Multimodal Architecture that coordinates operations among modality components.

**JavaScript.** A programming language used in conjunction with HTML for defining the behavior of web pages.

**Markup /Markup language.** An approach for annotating text in which the annotations are merged within the text (in a way that they are syntactically distinguishable from the text itself); one of the most common markup languages is HTML which is used to describe web pages and (by embedding programming code) as the basis to develop web-based applications.

**Modality Component (MC).** A component in the W3C Multimodal Architecture that processes a certain type of input, for example, a speech recognizer, handwriting recognizer, or natural language understanding system.

**Multimodal Architecture (MMI).** A software architecture for a multimodal system, for example, the W3C Multimodal Architecture, the Open Agent Architecture, or DARPA Communicator.

**Glossary** (*continued*)

**Multimodal fission.** The process of splitting a generic meaning into two or more modalities for presentation to a user.

**Multimodal fusion.** The process of combining results from two or more modalities into a single meaning.

**Representational State Transfer (REST).** An architecture based on the HTTP protocol commonly used in web-based applications.

**Speech Synthesis Markup Language (SSML).** An XML markup language for defining how a text should be pronounced, standardized by the W3C Voice Browser Working Group.

**State Chart XML (SCXML).** An XML language based on Harel State Charts, used for describing reactive processes, standardized by the W3C Voice Browser Working Group.

**Synchronized Multimedia Integration Language (SMIL).** An XML-based language for describing interactive multimedia presentations.

**Tag.** An XML annotation that defines metadata for the content that it surrounds; for example, “<sentence>I would like to go from Philadelphia to San Francisco</sentence>” expresses the fact that “I would like to go from Philadelphia to San Francisco” is a sentence.

**VoiceXML.** An XML markup language for defining form-filling spoken dialog applications, standardized by the W3C Voice Browser Working Group.

**World Wide Web Consortium (W3C).** An international organization whose mission is to support interoperability, security and accessibility of web pages by defining standard languages.

**eXtensible Markup Language (XML).** A markup language standardized by the World Wide Web Consortium and used in numerous applications.

**XML container element.** An XML annotation which surrounds other content with begin and end tokens, for example in “<author>William Shakespeare</author>”, “author” is the container element.

**XML child content.** Material included in an XML container element. “William Shakespeare” in the XML container element example.

In recent years, the basic underlying technologies for multimodal interaction, such as speech recognition, natural language understanding, and gesture recognition, have made striking technical advances. Applications using these technologies, including Apple Siri, Google Now, Microsoft Cortana, and Amazon Echo have also

become dramatically more capable within the space of only a few years. Nevertheless, compelling multimodal applications combining speech with other modalities are far less common. One common approach to integrating speech and graphics in applications is to support spoken inputs but show the user graphical results, possibly reinforced by text to speech, but this is only a limited use of multimodality. Very few commercial applications support fully integrated graphical and spoken input, although Openstream's EVA [[Openstream 2015](#)] is one example.

Arguably, one reason for the limited number of truly multimodal applications is the fact that access to the component technologies through *Application Programming Interfaces (APIs)* is frequently proprietary and vendor-specific. Consequently, developing multimodal applications requires not only mastering the conventions of multiple APIs, but also fusing results that may be represented in very different formats. This unnecessarily complicates the development process, increases the chances of integration errors, and makes debugging more difficult. In addition, it makes it harder for third parties to develop components that can be used with multiple systems, because they must develop several alternative versions of their software to fit into all of the proprietary systems. Finally, it reduces the incentive for third parties to develop tooling, because, again, any tooling will have to support multiple proprietary approaches, instead of a single approach that the community agrees to use.

The goal of the W3C in proposing multimodal standards is twofold: from one side, the aim is avoiding a proliferation of different vendor-specific approaches and on the other side, by leveraging the widespread knowledge of the *Hypertext Markup Language (HTML)*, the goal is to propose a simpler declarative approach based on markup languages for the specifications of multimodal systems. Unlike traditional desktop applications, applications authored in HTML can run on a wide variety of platforms, including desktop platforms (Windows, Apple, Linux) as well as many different mobile devices (iOS, Android) with widely varying screen sizes, keyboard capabilities, and pointing devices. By using standard protocols such as the *Hypertext Transfer Protocol (HTTP)* [[Fielding et al. 1999](#)], for communicating with servers, HTML applications can interoperate with remote servers running many different types of operating systems.

One of the early approaches towards an open architecture for multimodal interaction was the Open Agent Architecture [[Cheyer and Martin 2001](#), [Cohen et al. 1994](#)]. Open Agent Architecture (OAA) for short is a *framework* for integrating a community of heterogeneous *software agents* in a distributed environment developed at *SRI International Artificial Intelligence Center*. It is a distributed *agent* framework with a dynamic community of agents, where multiple agents contribute services to the community. When external services or information are required by a given

agent, instead of calling a known agent to perform a task, the agent submits a high-level expression describing the needs and attributes of the request to a specialized *Facilitator agent*. The Facilitator agent (analogous to the *Interaction Manager* of the W3C Multimodal Interaction Architecture, described later in this chapter) makes decisions about which agents are available and capable of handling sub-parts of the request, and manages all agent interactions required to handle the complex query.

The real challenge has been the lack of a standard/uniform interfaces for developing multimodal applications. There are a number of ways that technical communities can agree that they will build their technologies in an interoperable way, including informal agreements among labs, common protocols required by multi-site research programs, through industry consortia, or through standards organizations.

This chapter will discuss proposed standards that have been developed under the auspices of a formal standards organization, the W3C. Since 1994, the W3C has guided the definition of open standards and guidelines for the Web. The overall goal of W3C standards is to promote the use of the World Wide Web as an open infrastructure. Through a well-defined consensus and review process, W3C standards are agreed upon among industry partners.

In general, the W3C provides a standards development infrastructure, a publication process that insures broad review of any proposed standards and proper attention to general concerns such as security, internationalization, and accessibility. W3C standards are based on a consensus process that aims to ensure that even small organizations have input into the standards and have their issues addressed. Initial work on voice standards in the W3C began in 1999, with the formation of the Voice Browser Working Group. Work on multimodal standards began in 2002, with the formation of the Multimodal Interaction Working Group. This chapter focuses mainly on the multimodal standards. It describes the benefits of standards for multimodal applications and discusses the W3C standards that support multimodal interaction. It also discusses the general value of declarative representations such as those in the W3C standards. Voice standards are briefly discussed as appropriate in conjunction with multimodal standards.

### **9.1.1 Benefits of W3C MMI Standards-based Implementation**

In addition to the general benefits of standards described above, the specific approach to multimodal standards which will be described in this chapter adds some additional benefits.

Some of the key benefits which can be achieved by implementing multimodal applications using the W3C multimodal standards are outlined below.

- **Encapsulation.** Application authors need not make any assumptions about the internal implementation of components; thus, components can be treated as black boxes.
- **Distribution.** The architecture allows for implementing applications that could have both local (co-hosted) and distributed components. For example, local speech recognition could be used for navigation and remote recognition for complex and dynamic grammars and/or language-models.
- **Extensibility.** Authors can extend individual components without modifying the rest of the system. For example, a camera component could be extended to become a QR-Code scan or biometric authentication component by adding code to interpret images to the basic image capture code.
- **Nesting.** Components can be nested to create complex modality components, such as face-recognition and voice-authentication combined into one biometric verification. In this way the nested components can be used as a single component, thereby supporting encapsulation of more complex functionality.
- **Modularity.** One of the chief advantages of the MMI architecture is that it provides for the separation of data, control, and presentation, in line with the Model View Controller (MVC) design pattern. The separation of control from presentation makes code easy to read and maintain. In addition, it allows for maximum flexibility in configuring the system's graphical, speech modality (speech recognition and text to speech) and other modality components from various vendors, without loss of inter-operability.

A standard, uniform, API to different modalities would greatly simplify the development of multimodal applications. This situation has certainly improved recently for web-based modalities, because HTTP (Hypertext transfer protocol) and the REST paradigm (representational state transfer) [Fielding et al. 1999] have greatly contributed to the simplification of APIs. However, even given an HTTP REST interface there can be significant vendor-specific variations in the messages and vocabularies used in communication between components. Not only is a standard API important, but it is necessary for such a standard API to be as clear and comprehensive as possible. For this reason, the use of declarative language for describing modality control and user interaction is much more developer-friendly than approaches based on procedural paradigms such as JavaScript APIs.

The annotations in a declarative language may be semantic information or even executable instructions, and in this sense a markup language can be an alterna-

tive for an API. An improvement in usability results from the fact that declarative markup abstracts from the processing details that are required for execution, thereby allowing the developer to focus on the implementation of the application at a higher level.

A second factor limiting the number of multimodal applications involves the conceptual complexity of the multimodal interface. Coordinating multiple input and output modalities to present a coherent and easy-to-use interface can become very complex as the number of modalities increases. In addition, it is also highly desirable to be able to adapt the interface dynamically to take into account user abilities and preferences as well as the context of the interaction, all of which increase the complexity of the development process. Layers of abstraction in multimodal fusion and fission can reduce this complexity by isolating the core interaction logic from the modality-specific presentation. However, as in the case of APIs, further valuable simplification can be obtained by describing the multimodal user interface in terms of declarative markup.

Finally, multimodal systems also lend themselves very naturally to dynamic architectures, where multimodal components periodically join and leave the system as users move through their environments. This kind of dynamic configuration means that components must have agreed-upon ways of announcing their presence, informing other components of their capabilities, and reporting their status. A system could theoretically encounter dozens of these components within the course of executing an application. It is difficult to imagine this discovery and registration process taking place in a purely proprietary way, given the large numbers of potential system components.

## 9.2

### How the Standards Fit Together

This chapter discusses a number of standards that can be used separately or together in multimodal applications, so it is useful to begin with a quick summary of what they are and how they complement each other.

Multimodal systems based on the W3C standards are described in the W3C Multimodal Architecture standard [Barnett et al. 2012]. This standard can be conceptualized as a system consisting of a central coordinator, called the Interaction Manager (IM), which integrates the activities of one or more independent components (Modality Components or MCs), by means of standard events (Life Cycle Events). Examples of MCs include both input and output processors. Input processors could include speech recognizers, natural language understanding components, handwriting recognizers, and emotion recognizers. Output processors

include natural language generation components and text-to-speech systems. Several formal declarative markup languages have been defined which are used to represent the results of processing inputs and outputs. These will be discussed in detail in this chapter. Briefly, they include:

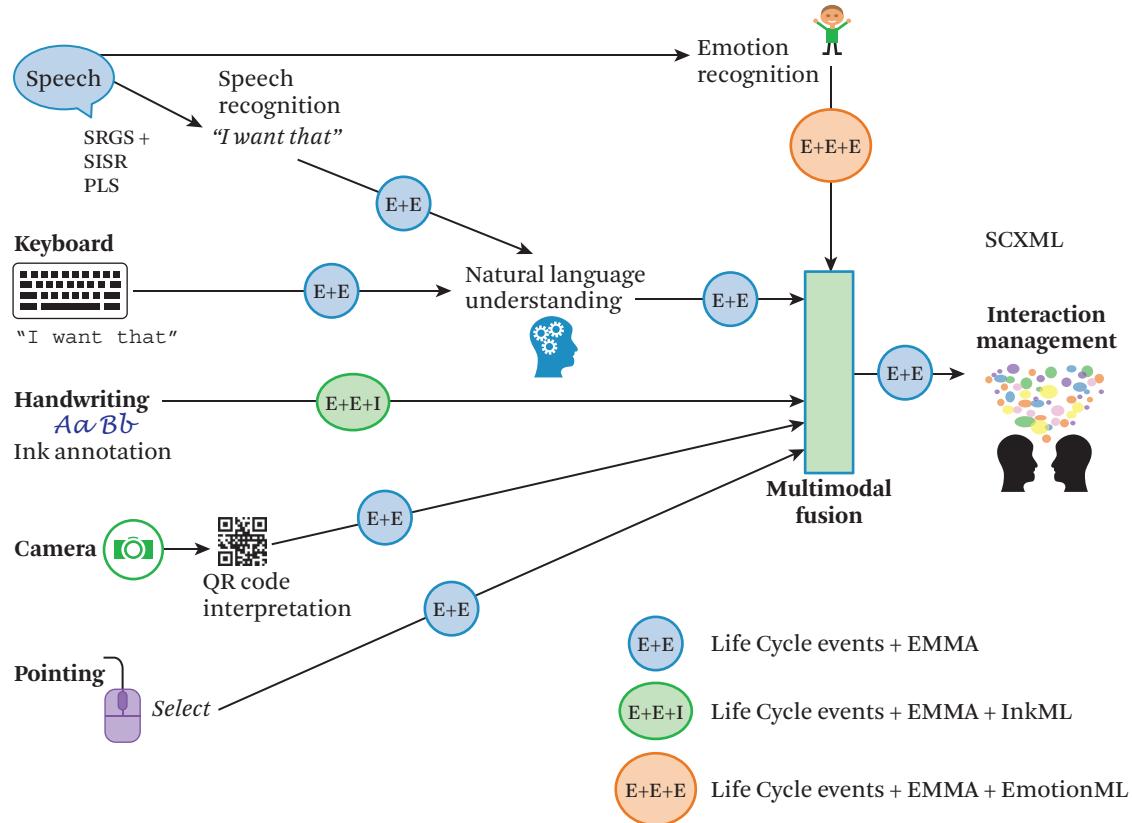
1. **Extensible Multimodal Annotation (EMMA)** (“AVIOS,; [Johnston 2009] [Johnston et al. 2015], which is used to represent the semantics of user inputs and system outputs as well as metadata (for example, confidence, input tokens, timestamps and alternative results);
2. **InkML** [Watt et al. 2011] which represents digital ink traces; and
3. **EmotionML** [Schröder et al. 2014], which represents emotions.

In addition to these data representation languages, a scripting language that defines the behavior of an IM has been standardized, State Chart XML (SCXML) [Barnett et al. 2015].

In addition to these standards, there is also a family of related voice standards which will not be discussed in detail here due to space limitations. These standards include VoiceXML [Oshry et al. 2007], a scripting language for voice systems, *Speech Synthesis Markup Language (SSML)* [Burnett et al. 2004] for text-to-speech systems, Speech Recognition Grammar Specification (SRGS) [Hunt and McGlashan 2004] for defining speech grammars, Semantic Interpretation for Speech Recognition (SISR) [Van Tichelen and Burke 2007] for defining the semantics of speech grammars, and the Pronunciation Lexicon Specification (PLS) [Baggia et al. 2008] for defining the phonetic pronunciations of words.

The interaction of the standards discussed in this chapter can be seen in Figure 9.1 and Figure 9.2.

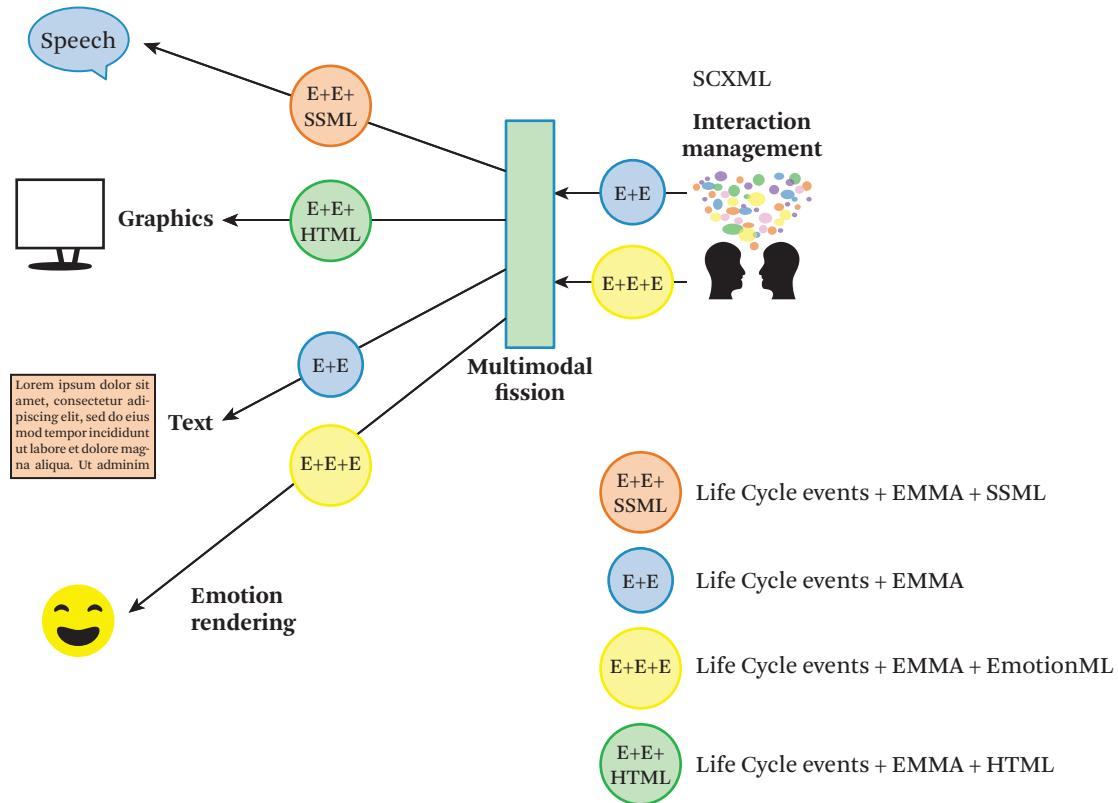
Figure 9.1 shows an example of a multimodal system supporting various modalities and shows how the standards discussed in this chapter apply to the multimodal inputs. The results of processing all the input modalities are represented in EMMA and are sent to later processing stages as data within multimodal Life Cycle events. These are the “E+E” labels for the arrows in Figure 9.1. In some cases, for example emotion recognition from speech, the result is represented in EmotionML contained in EMMA, which in turn is contained in a Life Cycle event (E+E+E). Similarly, an ink annotation can be represented in InkML contained in EMMA within a Life Cycle event (E+E+I). The only differences in results from different modalities that



**Figure 9.1** Examples of multimodal inputs

represent the same user intent will be the annotation of the **medium**<sup>1</sup> (acoustic, visual, or tactile) and mode (for example, speech, video, or ink) of the inputs. When multimodal inputs are fused, the result can again be represented in EMMA and sent via a Life Cycle event to an IM, which orchestrates the interaction. SCXML is a standard that can be used for specifying the behavior of an Interaction Management. Specifically, for the speech recognition modality, recognizers can use the SRGS, SISR, and PLS standards (which we will not discuss in detail here) to define

1. The medium refers to how information is stored or transferred while the mode (or modality) concerns the format (syntax, semantics and pragmatics) with which the information is realized [Cohen 1992]. Specifically, modality is a particular way in which information is to be encoded for presentation to humans.



**Figure 9.2** Examples of multimodal outputs.

speech grammars, semantic interpretation, and pronunciations, respectively. The modalities in Figure 9.1 are just examples; many other modalities are possible.

On the output side, as shown in Figure 9.2, we can see a similar pattern. Amodal (that is, generic modality-independent) system intents created by the IM undergo fission into separate multimodal outputs, which are sent to the appropriate output Modality Components for rendering. For speech output, a Life Cycle event containing an EMMA 2.0 “output” document is sent. The EMMA document contains the text to be rendered with a speech synthesizer, marked up with SSML that defines the intended detailed pronunciation of the text, including prosody. This is the “E+E+SSML” message shown in Figure 9.2. For expressing emotions, the IM can generate a modal intent for the system to express a particular emotion, again sent to the fission engine in a Life Cycle Event containing EMMA and EmotionML (E+E+E). Figure 9.2 shows the case of an emotion rendered with an emoji, but, de-

pending on the application, context, and user preferences, the emotion could be rendered in any number of alternative ways, such as tone of voice, text, or images.

The uniform formats and events provided by the standards mean that any software (for example, fusion and fission engines) that interprets or analyzes results can be greatly simplified by reducing the need to convert between different proprietary formats and syntaxes. Note that the standards are used either for communication (Life Cycle events, EMMA, EmotionML, and InkML) or for defining the application-specific behavior of components (SRGS, SISR, PLS, SSML, SCXML). The internal operation of components is outside the scope of the standards; for example, the standards do not say anything about how speech recognition, handwriting recognition, natural language understanding, fusion, or fission are actually done. The communication standards simply provide a format for representing and communicating the processing results to other components.

## 9.3

### **The Importance of Declarative Languages for Describing Multimodal Interaction**

Multimodal interaction languages are useful for representing design choices such as relations among parts and higher-level coordination in abstract terms. Declarative approaches allow designers and developers of interactive applications to address this increasing complexity in a more manageable way than traditional procedural paradigms. The main point in such approaches is to describe what features should be supported in order to facilitate the main design decisions and then leave the method to obtain them to a next phase (declarative vs. procedural knowledge). Such declarative approaches have often used markup languages, which are based on text markers that are used to annotate documents.

XML [Bray et al. 2004], defined by the W3C, is the most common of such languages in which the textual information is divided into markup (annotations or instructions) and content, which may be distinguished by means of simple syntactic rules. The most commonly used markups are tags which are sequences of characters surrounded by brackets. For example, in the following example, XML is used to annotate (or mark) the sentences in a document using the tag `<sentence>` (and the corresponding end tag `</sentence>`):

```
<sentence>Lorem ipsum dolor sit amet, consectetur adipiscing elit,  
sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.  
</sentence>
```

```
<sentence>Ut enim ad minim veniam, quis nostrud exercitation ullamco  
laboris nisi ut aliquip ex ea commodo consequat. </sentence>
```

XML can also be used to define instructions rather than to just annotate content. The structure of the tags allows the definition of actions and parameters such as in the example below which models a (simplified) financial transaction between two actors:

```
<transaction>
<payment currency='euro'>100</payment>
<to>John Smith</to>
<from>Bill Joe</from>
</transaction>
```

XML itself does not dictate which tags should be used but just how they should be used. Therefore, XML is a tool to define markup languages, and many such languages have been defined to represent rich content and web-based transactions (<https://www.w3.org/Consortium/mission>).

Several languages have been proposed for describing various aspects of multimodal interactions (some of them will be discussed in detail in the next sections), each of them targeting specific types of modality or specific types of design choices. In general, an effective representational language should:

1. be easy for a designer to represent and for a reader to understand multimodal design decisions even without programming skills
2. provide an appropriate degree of abstraction
3. be able to represent both content and control aspects of multimodal interpretation

## 9.4

### Model-based Specifications for Multimodal Interaction

In declarative approaches, the specific languages and methods adopted are often called model-based [Cantera et al. 2010] to highlight that they are based on languages describing the main application capabilities in conceptual terms. The purpose is to allow designers and developers to concentrate on such main semantic aspects and avoid learning a plethora of implementation languages. In addition, by linking semantic information and implementation elements it is then possible to obtain device interoperability through several possible implementation languages, and to facilitate support of assistive technology.

Various possible abstraction levels have been identified in order to describe interactive systems. We discuss these in the next few sections, going from the most abstract to the most concrete level.

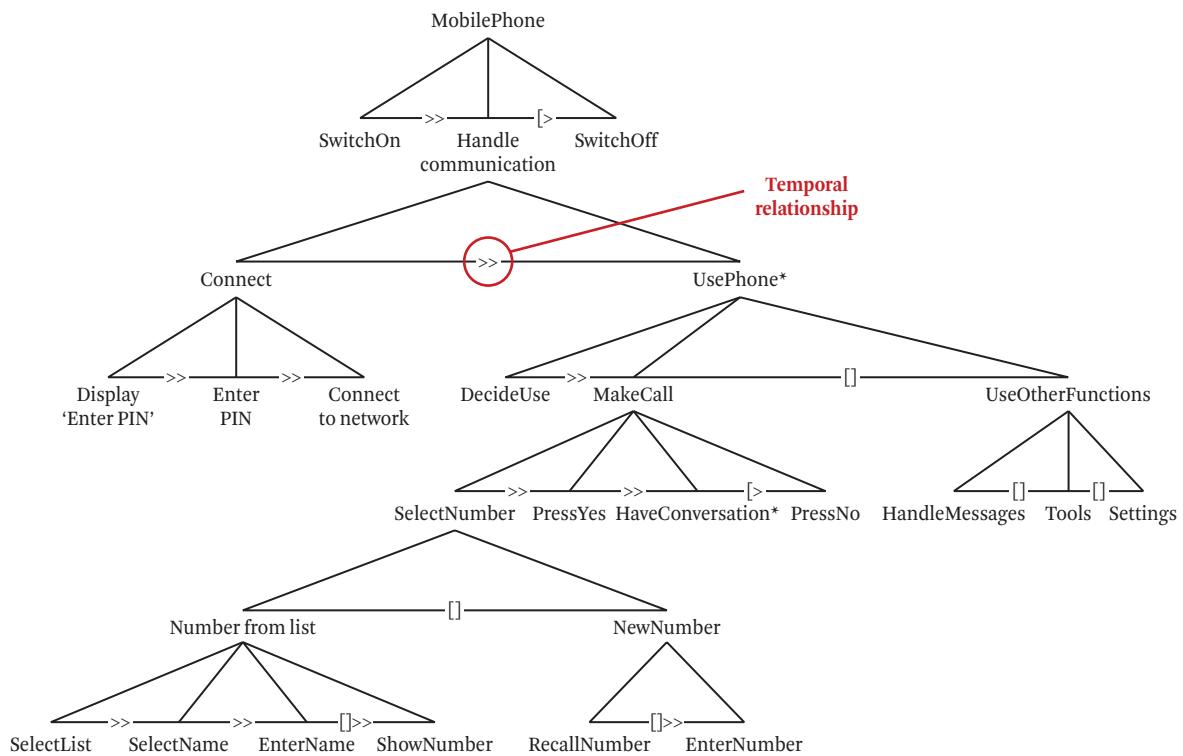
### 9.4.1 Logical Tasks

The most abstract level is the level of logical tasks and associated domain object descriptions. At this level the purpose is to indicate the possible activities that should be carried out through the considered application (for example, “I want to select an artwork description” or “I want to turn the room light on”). A possible way to describe interactive systems at this abstraction level is indicated at <https://www.w3.org/TR/task-models/>, which is based on the long experience of the wide community that has used the ConcurTaskTrees [Paterno 2003] language. Figure 9.3 shows an example of graphical task model specification (it is possible to obtain a corresponding XML-based description). It is characterized by three main aspects:

1. hierarchical description. High-level tasks can be decomposed into more concrete ones (going from top to bottom in the description);
2. temporal relationships, which facilitate the possibility of describing flexible behaviors in which activities can be performed concurrently or optionally or interrupt others; and
3. task allocation, indicating what actors should perform the tasks (the user, the system or an interaction between them).

Figure 9.3 is an example describing some tasks that can be performed with a smartphone. We can see that the root task (Using MobilePhone) is first decomposed into an interactive task (Switch On) followed ( $>>$  is the sequential operator) by an abstract task concerning Handling communication, which can be interrupted ( $[>$  is the disabling operator) by The SwitchOff task. Then, going down in the hierarchical description we can see that the Handling Communication task is decomposed into a Connect to Network task followed by the Use Phone task, which are detailed in the next levels.

This level of description can be used in various ways. When designing a new interactive application, after an initial task analysis aimed at identifying the main tasks and associated information, it can be useful to precisely indicate the relationships among the various tasks in order to better understand what the overall resulting behavior can be. In addition, since the task model should be the result of an interdisciplinary discussion amongst the various stakeholders (domain experts, designers, developers, end users) it can provide an overall view of the application’s interactive and functional aspects able to satisfy their requirements, and thus provide a useful starting point for the actual implementation work. Since the task model contains indications of the temporal relationships among the various tasks, it can be useful to obtain an interaction manager, which provides a more detailed



**Figure 9.3** Example of Task model specification.

description of such relationships at the implementation level and can also be specified by using SCXML.

#### 9.4.2 Abstract User Interface

Another possible description level is the Abstract User Interface (AUI) level. In this case the purpose is to describe the user interface in terms of elements that are independent of the possible interaction modalities. Thus, for example it is possible to indicate that in one point there is need for a single selection object with high cardinality, but without indicating whether such selection is realized through a graphical element or a vocal command or a gesture. The main criteria used to specify the meta-model of abstract user interfaces are completeness, extensibility, and conciseness.

### 9.4.3 Concrete User Interface

A more refined description can be provided at the Concrete User Interface (CUI) element. In this case the description is modality-dependent but implementation language independent. This means that it assumes the use of a specific modality but not the use of a specific implementation language. Thus, for example it is possible to indicate that a radio-button is necessary but then the radio button can still be implemented using different implementation languages (e.g., Java and its libraries or HTML or Windows toolkits).

### 9.4.4 Relationships Between Abstraction Levels

It is important to note that it is possible for systems to move through such abstraction levels both through forward engineering techniques (from more abstract to more concrete) and through reverse engineering techniques (from more concrete to more abstract). Moving across different abstraction levels requires the use of mappings between elements at such levels or the application of specific transformations. It is important to note that the same task model may give rise to one or many abstract user interfaces, one abstract user interface may give rise to one or many concrete user interfaces or to one or more final user interfaces subsequently.

In model-based languages and frameworks there is usually one abstract description and various concrete refinements according to the modalities addressed. Such refinements can also address multiple modalities at the same time (for example able to support graphical and vocal interaction in a combined manner). In this case they have to indicate how the various parts of the user interface exploit the available modalities. For this purpose, often the CARE [Coutaz et al. 1995] properties have been used. They are: Complementarity—the elements of the user interface considered are partly supported by one modality and partly by another one; Assignment—the user interface elements considered are associated with one given modality; Redundancy—the user interface elements considered are associated with multiple modalities at the same time; and Equivalence—the user interface elements considered can be supported by one modality or another. It is interesting to note that such properties can be applied to various granularity levels of the user interface: they can involve the entire user interface, or groups of elements within it, or single elements, or even parts of single elements.

### 9.4.5 The Role of Models in Contextual Adaptation

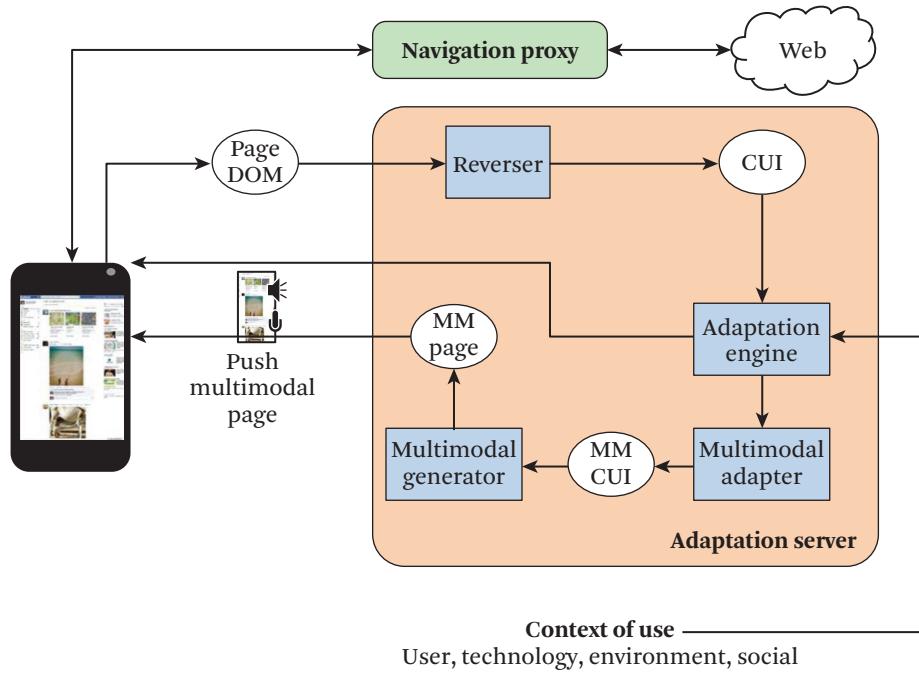
One interesting application of such model-based approaches in context-dependent multimodal augmentation is the ability to transform a web application that was

originally graphical into a multimodal application in order to better support a change in the context of use [Cantera et al. 2010].

One example is accessing a Web application while starting to drive a car, thus the visual channel is busy and can no longer be used for the interaction while vocal access is still possible. In this case the contextual event that triggers the adaptation is that the user gets in the car. At this point, the original graphical Web application is transformed in such a way to provide also vocal output and receive vocal commands equivalent to those of the original graphical version. This can be obtained through first a reverse engineering process that builds the logical description corresponding to the initial graphical user interface, which is then transformed into a multimodal concrete description that still supports the same interaction semantics but through different modalities. This is then used to obtain the corresponding multimodal implementation. Figure 9.4 shows an architecture able to support such dynamic multimodal augmentation. The user needs to access the web application through a proxy that includes in it scripts that enable the possibility of multimodal augmentation. When some contextual change occurs (e.g. the user enters the car) this is communicated to an adaptation engine, which includes the specification of the adaptation rule. In this case the rule indicates that the application should become multimodal in order to allow users to access it even when the visual channel is busy for driving. Then, the Document Object Model (DOM) of the web page (which is the description of the elements composing the current web page) of the considered application is retrieved in order to build its logical description (CUI) through a reverse engineering transformation. This is used to obtain a corresponding multimodal version (MM CUI), which maintains the same semantics, but it is able to provide the possibility of interacting vocally as well. Its corresponding implementation is generated and uploaded in the users' browser so that they can interact with the new possibilities. In this transformation it is possible to add rules that consider the specific characteristics of each modality. For example, for the vocal part for a driver it may be important not to present and read long lists of items, and to present only the top 3, and the user would navigate by saying things like "next."

Please note that this transformation among different versions of web pages was made possible because the pages are written in a standardized language.

Having discussed the importance of standards, declarative representations, and model-based specifications for multimodal interaction, we now turn to standards that address specific modalities. We will discuss existing standards for voice, ink, behavior, and emotion. Following the modality-specific discussion, we will address how standards can address the generic multimodal interaction problems of fusion,



**Figure 9.4** Architecture for model-based multimodal augmentation.

fission, and interaction management. We conclude with a discussion of future directions and lessons learned in the implementation of commercial, standards-based, applications.

Each modality has specific features. Thus, there are languages that have been developed for the purpose of capturing such features and to allow developers to manipulate them. In this section we introduce some of them, and then we move on to discuss how it is possible to specify multimodal user interfaces able to use multiple modalities at the same time.

#### 9.4.6 Dealing with Voice: Voice Standards

One of the earliest and most important voice standards was VoiceXML, an XML language which allowed developers to define form-filling dialogs with declarative markup. However, because VoiceXML assumes that the entire dialog is voice-based, without a GUI component, it is difficult to use in multimodal applications, and for that reason it will not be discussed in detail in this chapter.

For multimodal interaction, a better approach is to have a higher-level controller which is not modality-specific, and which responds to inputs from multiple modalities. This chapter discusses a candidate language for defining the behavior of such a higher-level controller (SCXML) below. However, there are specific standards for speech recognition and text-to-speech which were originally developed to support VoiceXML, and which are useful in multimodal systems. Speech Recognition Grammar Specification (SRGS) [Hunt and McGlashan 2004] is a standard format for context-free grammars for speech recognizers. SRGS grammars can also include notations for describing how a particular speech recognition result is to be semantically interpreted. These notations are defined in the Semantic Interpretation for Speech Recognition format (SISR) [Van Tichelen and Burke 2007]. SRGS can also be used in reverse to generate the text of system responses [Dahl et al. 2011] which would be rendered by a text to speech (TTS) engine. Speech Synthesis Markup Language (SSML) [Burnett et al. 2004] is a standard format for marking up text to indicate features such as volume, speech rate and prosody. Finally, the Pronunciation Lexicon Specification (PLS) [Baggia et al. 2008] can be used to define the pronunciations of words for both speech recognition and text to speech engines.

#### 9.4.7 Dealing with Digital Ink: InkML

Writing with a digitizing pen/stylus/finger has proven to be a natural and effective way to provide input since most people learn to write at school and tend to prefer this input method as more electronic devices with touch interfaces continue to become available. [Johnston 2019] provides additional details on the advantages of pen interfaces.

Typically, a touch-sensitive device allows movements of the digitizing pen to be captured as digital ink using radio frequency, optical tracking, physical pressure, or other technologies. It can then be either stored as documents or notes for later retrieval or passed on to recognition software for conversion into appropriate computer actions. Since ink documents capture information as the user composed it, including text in any mix of languages and drawings such as equations and graphs, they are often found to be very useful.

In order to provide a public and comprehensive digital ink format for the capture, transmission, processing, and presentation across heterogeneous devices developed by multiple vendors, the Ink Markup Language [Watt et al. 2011] was standardized by the W3C Multimodal Interaction Working Group in 2011. It provides a simple and platform-neutral data format to promote the interchange of digital ink between software applications. The markup allows for the input and processing of handwriting, gestures, sketches, music, and other notational languages

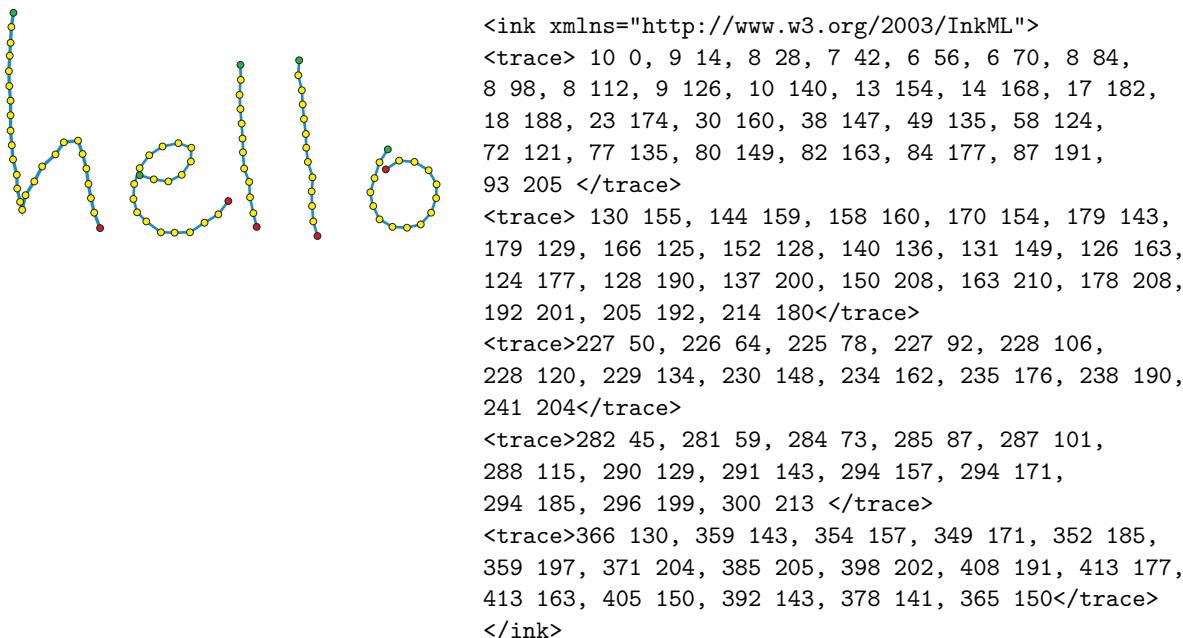
in applications. It provides a common format for the exchange of ink data between components such as handwriting and gesture recognizers, signature verifiers, and other ink-aware modules.

In addition to the pen position over time, InkML allows recording of information about device characteristics and detailed dynamic behavior to support applications such as handwriting recognition and authentication. For example, there is support to record additional information such as pen tilt and pen tip force (often referred to as “pressure”) and information about the recording device such as accuracy and dynamic distortion. InkML also provides features to support rendering of digital ink captured optically to approximate the original appearance. For example, stroke width and color information can be recorded.

InkML does not describe and store semantic information, such as the plain text of ink recognized as handwriting, nor does InkML store the contextual information about the ink, such as what kind of field in a form where ink was written. However, InkML provides means for extension and can include XML from other schemas at specific locations in a file or stream. For example, the InkML implementation in Microsoft Office includes embedded EMMA [Johnston 2009] in order to represent the results of handwriting recognition. Additionally, InkML could be embedded within other XML documents, for example, it can be embedded within EMMA documents to represent handwriting or ink gesture input as discussed in the next section. EMMA documents containing InkML can include contextual information about the containing form field and the results of handwriting recognition.

With the establishment of a non-proprietary ink standard, a number of applications, old and new, are expanded where the pen can be used as a very convenient and natural form of input. Here are a few examples.

- Ink Synchronized Multimedia Integration Language Messaging. Two-way transmission of digital ink. Users can draw or write with a pen on the device’s screen to compose a note in their own handwriting and the recipient views the message as the sender composed it, including text in any mix of languages and drawings.
- Ink and SMIL [Bulterman et al. 2008] Synchronized Multimedia Integration Language (SMIL) is a W3C standard for synchronizing multimedia outputs, thus it provides support for coordinated outputs including ink, images, and audio. For example, a photo taken with a digital camera can be annotated with a pen; the digital ink can be coordinated with a spoken commentary. The ink annotation could be used for indexing the photo (for example, one



**Figure 9.5** InkML representation for “hello”.

could assign different handwritten glyphs to different categories of pictures); see Figure 9.6

- **Ink Archiving and Retrieval.** A software application may allow users to archive handwritten notes and later retrieve them by a variety of mechanisms. This could include archiving signatures from signed documents, for example, signatures acknowledging receipt of deliveries.
- **Electronic Form-Filling.** In support of natural and robust data entry for electronic forms on a wide spectrum of keyboard-less devices.
- **Pen Input and Multimodal Systems.** Robust and flexible user interfaces can be created that integrate the pen with other input modalities such as speech [Cohen and Oviatt 2017, Johnston 2019, Sonntag 2019] Multimodal applications may share context information across modalities, leading to better recognition in each modality individually. In this setting, pen input may be used as input to a component for disambiguating voice recognition and vice versa.



**Figure 9.6** Example of ink annotation on image of damaged wall. (Source: Openstream.com)

Most ink-related applications fall into two broad categories: “Streaming” and “Archival”. Archival ink applications capture and store digital ink for later processing, such as document storage/retrieval applications and batch forms processing. In these applications, an entire `<ink>` element is written prior to processing.

Streaming ink applications, on the other hand, transmit digital ink as it is captured, such as in the electronic whiteboard. In order to support a streaming style of ink markup generation, the InkML language supports the notion of a “current” state (e.g., the current brush) and allows for incremental changes to this state.

InkML, with its rich capabilities for describing the detailed properties of digital ink, provides a foundation for applications that do further processing on the ink traces. While handwriting recognition is one clear example of a possibility of further processing of InkML data, recognition of mathematical expressions has also been done using InkML as a representation format [[Mouchere et al. 2016](#)]. Other applications based on InkML include collaboration between remote researchers [[Quiniou et al. 2009](#)].

```

<sentence id='sent1'> Thanks for the present!</sentence>
<emotion xmlns="http://www.w3.org/2009/10/emotionml"
          category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
    <category name="happy" value="0.4"/>
    <reference role="expressedBy" uri="#sent1"/>
</emotion>

```

**Figure 9.7** An example of content an annotation in EmotionML.

### 9.4.8 Representing Emotions: EmotionML

Expressing emotional content or recognizing emotion in a message is often part of multimodal interaction. In 2009, an exploratory work was formalized in the “Recommendation Track” [Oviatt et al. 2017] at W3C and the First Public Working Draft (FPWD) of a representational language specifically meant to describe emotional content, EmotionML 1.0, was released in 2010 and the final Recommendation delivered in 2014 [Schröder et al. 2014].

EmotionML is meant to provide a framework and a vocabulary to mark up emotion-related content to be used as a plug-in language to enrich the expressiveness of other markup languages. In this way, for example, a text message may be reproduced with the intended emotional connotation. The approach taken by EmotionML is to be independent from the actual way of expressing the emotion (with voice tone and facial expression by a synthetic character or with a simple visual annotation by a cartoon, for example).

Figure 9.7 shows how the description of a sentence (not part of the EmotionML language) may be annotated with emotional content by providing a linked element that enriches the information without requiring modifications to the original representational language.

Since EmotionML is meant to be embedded in other markup languages, emotions are not annotated using specific tag names but rather as attribute values connected to the elements by means of unique identifiers (called URI’s). In this way, with reference to Figure 9.6, there is no need to change the specification of the `<sentence>` tag to allow embedding new tags to describe emotions but rather the element that needs to be tagged is referenced through the unique identifier `"#sent1"/>`.

This approach also supports accommodating the different ways commonly used to represent emotions: as categories, dimensions, appraisals, and action-tendency. Each perspective corresponds to a tag with relevant attributes. Therefore, in the example of Figure 9.6, the perspective taken is a categorical one, which assumes a

finite set of emotion types. The specific label used comes from Cowie's "everyday emotion vocabulary" [Cowie et al. 2001]. In this respect, the language is relatively agnostic with respect to the theoretical approach chosen to model emotions in a given application (and the different perspective can also be merged in the same annotation). Of course, it is important to share a common vocabulary for the category names and the other features. Indeed, the W3C proposed such a vocabulary as an addendum and it is integral part of the standard [Burkhardt et al. 2014].

EmotionML also includes a specific attribute to mandate how a specific emotion should be conveyed. It is the attribute "expressed-through" and this attribute may specify a list of modalities such as face, gaze, or voice.

Since emotions may have time constraints, EmotionML borrows from EMMA (see Section 9.5) an embedded mechanism to specify time intervals providing attributes for both absolute and relative time specifications.

## 9.5

### Modality Fusion and Media Synchronization

Multimodal fusion is a central task in the interpretation of multimodal user input. As inputs arrive from the different modalities, their interpretations must be integrated in order to derive core user intents. If the modalities produce interpretation results in very different formats, this integration process can be quite complex. In addition, it is not enough to simply fuse the original modal inputs into a single semantic intent, it is also important to be able to trace back the interpretation to the original input modalities. Preserving the origin of the separate inputs can be useful for several purposes: determining what output modalities to use, allowing an interaction manager to judge the relative reliability of the modalities in case of a conflict, and logging and debugging. Furthermore, since the relative timing of modal inputs is necessary for understanding which modality results are part of a single user input, input timing must also be tracked. Finally, it is possible for fusion to occur in more than one stage, and the history of these stages should be retained. For example, fusion for audio-visual speech recognition [Potamianos et al. 2017] might precede fusion of speech recognition results with the results of gesture interpretation. Knowing the outcome of the audio-visual speech recognition process separately would also be very useful for debugging and tuning.

Thus, there are five important requirements for an approach to representing fused results. The fusion approach must be able to:

1. represent the semantics of the original modal inputs in a uniform format;
2. represent the semantics of the fused input;

```
<emma:interpretation id="r1" emma:medium="acoustic" emma:mode="voice">
  <departure_city>philadelphia</departure_city>
</emma:interpretation>
```

**Figure 9.8** An example of `<emma:interpretation>`.

3. preserve the modality information from the modal inputs;
4. preserve the relative timing of the original inputs;
5. support multiple stages of fusion.

The Extensible Multimodal Annotation (EMMA) specification [Johnston 2016, Johnston 2009, Johnston et al. 2015] has been designed to meet these requirements. It is an XML (W3C 2000) language standardized by the W3C Multimodal Interaction Working Group for the purpose of providing a standard representation for semantically complex multimodal user input. EMMA 1.0 [Johnston 2009] addressed only user input; however, EMMA 2.0 [Johnston et al. 2015] addresses the representation of both user input and system output. In this section we will focus on support for fusing user input. We discuss these requirements in detail in the following sections and explain how EMMA meets them. Note that EMMA is agnostic to the specific fusion algorithm used in a system. It assists fusion by providing a uniform set of metadata and syntax across modalities, reducing the complexity of parsing widely varying modality output formats.

### 9.5.1 Representing the Semantics of a User Input

The element `<emma:interpretation>` is the container for the interpretation of an input. The XML child content under `<emma:interpretation>` represents the interpretation of the input using an application-specific notation. Figure 9.8 shows an example of a simple XML interpretation for an utterance like “*I'd like to leave from Philadelphia.*”

Although the interpretation itself is modality-independent, note that the EMMA representation includes references to the medium and mode of the input. This makes it possible to distinguish between a spoken input and other forms of input. For example, if the user had selected the departure city from a dropdown list, the medium would be “tactile,” and the mode would be “gui,” but the semantics would be the same.

```

<emma:interpretation id="multimodal1"
    emma:medium="acoustic tactile"
    emma:mode="voice ink"
    emma:derived-from resource="#voice1" composite="true"
    emma:derived-from resource="#ink1" composite="true"
    <destination>Boston</destination>
/>
```

**Figure 9.9** Composite multimodal input.

### 9.5.2 Representing the Semantics of a Fused Input

A composite multimodal input derived from multiple unimodal inputs uses the element `<emma:derived-from>` to point to the unimodal inputs, and includes an attribute `<emma:composite="true">` to indicate that the interpretation was the result of fusion. The values of the `<emma:medium>` and `<emma:mode>` attributes are the union of the values of the medium and mode values of the unimodal inputs.

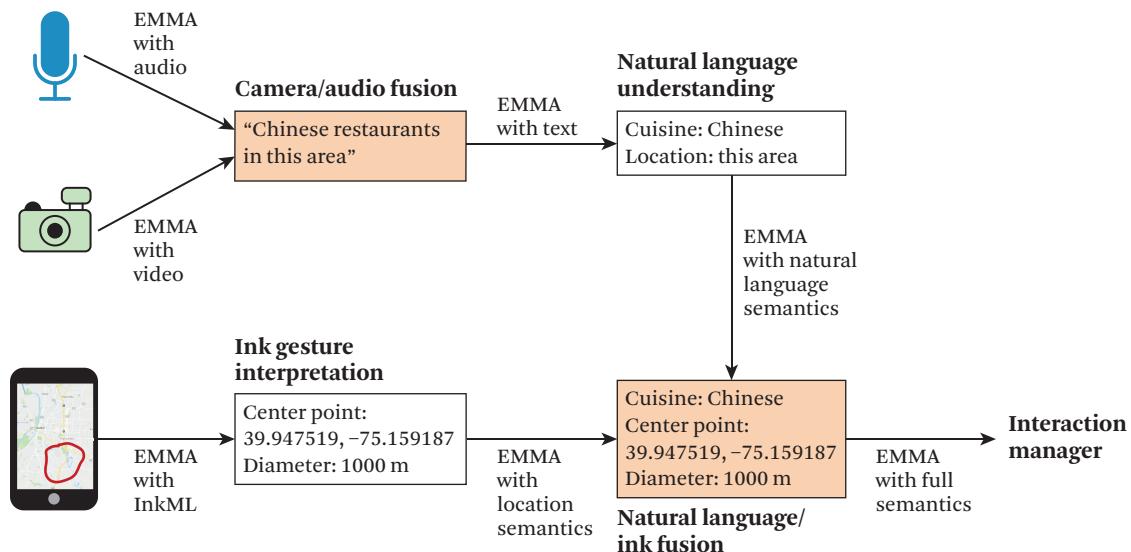
Figure 9.9 shows a composite multimodal interpretation where the user said “destination” while circling Boston on a map in ink mode. The “derived-from” links point to the unimodal voice and ink inputs (not shown) from which the composite input was derived.

### 9.5.3 Preserve the Modality Information From the Modal Inputs.

The *medium* and *mode* attributes of a composite or fused interpretation contain the union of the *medium* and *mode* values of the original inputs. This information may be sufficient for some applications; however, it does not preserve the information about which input came from which medium and mode. If this information is necessary, the overall interaction coordinator (the Interaction Manager as shown in Figures 9.1 and 9.2) must follow the *derived-from* links and inspect the original unimodal `<emma:interpretation>` elements.

### 9.5.4 Preserve the Relative Timing of the Original Inputs

EMMA supports a comprehensive set of timestamping options, including both absolute and relative timestamps. The timestamps for a fused input include the entire interval in which any one of the original inputs occurred; thus, this interval is likely to be longer than any one of the original inputs. As in the case of *medium* and *mode*, the timestamps of the original inputs are available, if needed, by inspecting the timestamp attributes in their respective `<emma:interpretation>` containers, available by following the *derived-from* links. Although timestamps are optional in



**Figure 9.10** Multiple stages of fusion with EMMA.

EMMA in general, if the application involves fusion, timestamps will be necessary for that application.

### 9.5.5 Support Multiple Stages of Fusion

The derivation of an interpretation can include multiple steps and combinations of input from different modalities. Figure 9.10 shows a schematic example of local business search including audiovisual speech input from a camera and microphone and ink input from a touchscreen. The user says, “Chinese restaurants in this area.” The camera and microphone inputs are fused to produce the speech recognition result, which is then fused with the touchscreen input to identify the area referred to. The results of each stage of processing can be represented in the `<emma:interpretation>` elements in the derivation chain.

## 9.6 Multimodal Fission and Media Synchronization

EMMA 2.0 [Johnston et al. 2015] is a major extension to the EMMA language that enables it to support system output as well as user input. With this feature, both sides of a dialog can be included in the same EMMA document, using the same syntax and the same metadata (where applicable). The derivation for an output is

basically the inverse of the derivation of a semantic interpretation for an input. It begins with an abstract system intention and proceeds through stages that become progressively more concrete until we arrive at the final concrete presentation to the user. These stages could include fission into multiple modalities, if the goal is to produce a multimodal presentation, or they could be simply stages of increasingly concrete representations if the ultimate presentation will be unimodal.

Multimodal output of dynamic formats (audio, video, animation) must often be synchronized with other outputs. EMMA 2.0 does not define its own method of synchronizing outputs, because the W3C has produced a number of other standards that can be used to support media synchronization. SMIL [Bulterman et al. 2008] allows an author to control the timing of a media presentation with an XML format. Timed Text Markup Language (TTML) [Adams et al. 2015] is another W3C specification aimed specifically at synchronizing video and text. TTML is used, for example, by broadcasters in applications such as closed captioning and subtitling.

A third approach to synchronizing media is to make use of the HTML5 `<audio>` and `<video>` tags, whose properties (such as `currentTime`) can be changed in JavaScript. However, if the functionality required by the application can be supported in a declarative markup language (Timed Text or SMIL), it would be preferable for ease of development and maintainability to use these declarative approaches rather than JavaScript.

EMMA 2.0 supports incorporating existing synchronized media markup languages such as TimedText and SMIL in the `<emma:output>` element. It also indirectly supports the JavaScript approach because in general, HTML can be included in `<emma:output>`.

Figure 9.10 is an example (modified from an example in the EMMA 2.0 Working Draft) of how EMMA can be used to represent the results of multimodal fission. In this case, the original amodal semantics (represented in the `<emma:derivation>` section in Figure 9.11) is an application-specific semantic frame for an air travel application containing slots for origin, destination, airline and time. The fission process produces a spoken output, “I found two flights from Boston to Denver.” as well as a graphical table that could be displayed in a web browser.

EMMA can also include other languages for describing system output. For example, it could include Behavior Markup Language BML [Kopp et al. 2006] describing the behavior of an avatar, as shown in Figure 9.12. The EMMA document in Figure 9.12 instructs a text to speech system to say “I found three flights from Boston to Denver” while also instructing the avatar to widen its eyes.

```
<emma:emma version="2.0">
  <emma:group
    emma:process="http://example.com/multimodal_presentation_planner">
      <emma:derived-from resource="oooo" composite="false"/>
      <emma:output id="ooo1"
        emma:medium="acoustic"
        emma:mode="voice"
        emma:result-format="application/ssml+xml">
        <speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
          xml:lang="en-US">
          I found two flights from Boston to Denver.
        </speak>
      </emma:output>
      <emma:output id="gui1"
        emma:medium="visual"
        emma:mode="gui"
        emma:result-format="text/html">
        <html xmlns="http://www.w3.org/1999/xhtml">
          <body>
            <table>
              <tr>
                <td>United</td>
                <td>5:30pm</td>
              </tr>
              <tr>
                <td>American</td>
                <td>6:10pm</td>
              </tr>
            </table>
          </body>
        </html>
      </emma:output>
    </emma:group>
```

**Figure 9.11** Representing multimodal fission in EMMA.

```

<emma:output id="oooo">
  <flights>
    <flight>
      <origin>Boston</origin>
      <destination>Denver</destination>
      <airline>United</airline>
      <time>5:30 P.M.</time>
    </flight>
    <flight>
      <origin>Boston</origin>
      <destination>Denver</destination>
      <airline>Delta</airline>
      <time>7:00 P.M.</time>
    </flight>
  </flights>
</emma:output>
</emma:derivation>
</emma:emma>

```

**Figure 9.11** (continued)

### 9.6.1 Interaction Management: State Chart XML (SCXML)

Multimodal applications are unpredictable, in that it is not always possible to know what the user will do next. Hence, procedural languages would be a bad choice for implementing multimodal systems. One needs a structured way to represent an application and its behavior through various events. While there have been several graphical specification languages such as Unified Modeling Language (UML) [James et al. 1999] that can describe the semantics for sophisticated constructs such as parallel states, they do not have declarative XML representations. An efficient and conceptually clear way of representing the behavior of many types of systems, including user interfaces, is by means of state machines [Harel 1987]. A system may be modeled as being in one of a set of possible states and for each state, a behavior is defined for each possible input or event. The behavior may be performing some operations, but it also includes the system entering- a different state or remaining in the same state. State Charts, as defined in Harel [1987], are a visual formalism for describing state machines as applied to reactive or event-driven systems, such as user interfaces.

In 2005, the W3C Voice Browser Working Group, recognizing the value of a formal XML definition of state charts, began specifying SCXML [Barnett 2016, Barnett et al. 2015] as an XML dialect and defining corresponding semantics for state charts

```

<emma:emma version="2.0">
<emma:group
    emma:process="http://example.com/multimodal_presentation_planner">
    <emma:output emma:confidence="0.8" id=" tts1">
        emma:medium="acoustic"
        emma:mode="voice"
        emma:result-format="application/ssml+xml">
            <speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis">
                I found three flights from Boston to Denver.
            </speak>
        </emma:output>
        <emma:output id="gui1">
            emma:medium="visual"
            emma:mode="gui">
                <bml
                    character="Alice"
                    id="bml1">
                    <face id="behavior1" amount="0.8" start="0" end="4">
                        <ext:facs au="1" side="BOTH"/>
                        <lexeme lexeme="WIDEN_EYES"/>
                    </face>
                </bml>
            </emma:output>
        </emma:group>
    </emma:emma>

```

**Figure 9.12** EMMA with embedded BML.

and their interpretation. SCXML achieved W3C Recommendation status in 2015. In the context of multimodal interaction, SCXML is the markup language proposed for describing multimodal dialogs in the W3C Multimodal Architecture. SCXML is a general-purpose event-based state machine language that combines Harel State Tables semantics [Harel 1987] with an XML syntax.

The basic state machine concepts are states, transitions, and events. A state contains a set of transitions that define how it reacts to events, generated by the state machine itself or by external entities.

In a traditional state machine, the machine is always in a single state, called the active state. When an event occurs, the state machine checks the transitions that are defined in the active state. If it finds one that matches the event, it moves from the active state to the state specified by the transition (called the “target” of the transition.) Thus, the target state becomes the new active state.

```

<x-scxml initialstate="Login">
  <state id='Login'> . . . </state>
  <state id='Order'> . . .
    <transition target='Payment' />
  </state>
  <state id='Payment'> . . . </state>
</x-scxml>

```

**Figure 9.13** An example of part of a state machine encoded in SCXML.

The example in Figure 9.13 shows a part of a state chart for an ecommerce application with the three states: login, order and payment. The figure shows how the login state is the initial one that means that all the transitions will start from here and the transition between order and payment.

Beyond the traditional model, SCXML also defines the notions of compound state and parallel elements.

A compound state is an element that contains nested state elements. When a compound state is active, one and only one of its child states is active. Conversely, when a child state is active, its parent state is active too.

The parallel element represents a state whose children execute in parallel. Specifically, when the state machine enters the parallel state, it also enters each child state. The child states execute in parallel in the sense that any event that is processed is processed in each child state independently, and each child state may take a different transition in response to the event.

SCXML can potentially serve as a description language for the logical tasks in model-based user interfaces: we can see how these pieces fit together by going back to Figures 9.1 and 9.2. EMMA contains the user inputs and system outputs, conveying them to the overall Interaction Manager, which is written in SCXML. In the specific cases of ink or emotion input and output, InkML and EmotionML can be used within EMMA to represent specific information about emotion or ink inputs or outputs.

## 9.7 Lessons Learned From the Implementation of Multimodal Standards

While there have been several attempts [Cross et al. 2003, IBM 2001, 2003, SALT 2002] at standardizing structure and tooling for multimodal application development, the effort by the W3C Multimodal Interaction Working Group that started

in 2002 and ratified in 2012 has been the most comprehensive one that combined the work of academia and industry. It is derived primarily from the Galaxy Communicator, a distributed hub-and-spokes architecture [Bayer 2005, Bayer et al. 2001] and the MVC architectural pattern.

W3C Multimodal Architecture supports distributed-modality components and defines a component life-cycle API for messaging between modality components (MCs) and IM. It specified a runtime framework (RTF) that provides runtime support for the IM, MCs, and data component.

There have been several commercial implementations, at least in part, of the W3C MMI Architecture by AT&T, Google, HP, IBM, Loquendo (now Nuance), Openstream, and Microsoft. Examples of applications based on the Multimodal Architecture include accessible smart homes [Ketsmur et al. 2018] and a multi-agent communication framework [López Herrera and Ríos-Figueroa 2018].

Some of the lessons learned from these implementations are given below.

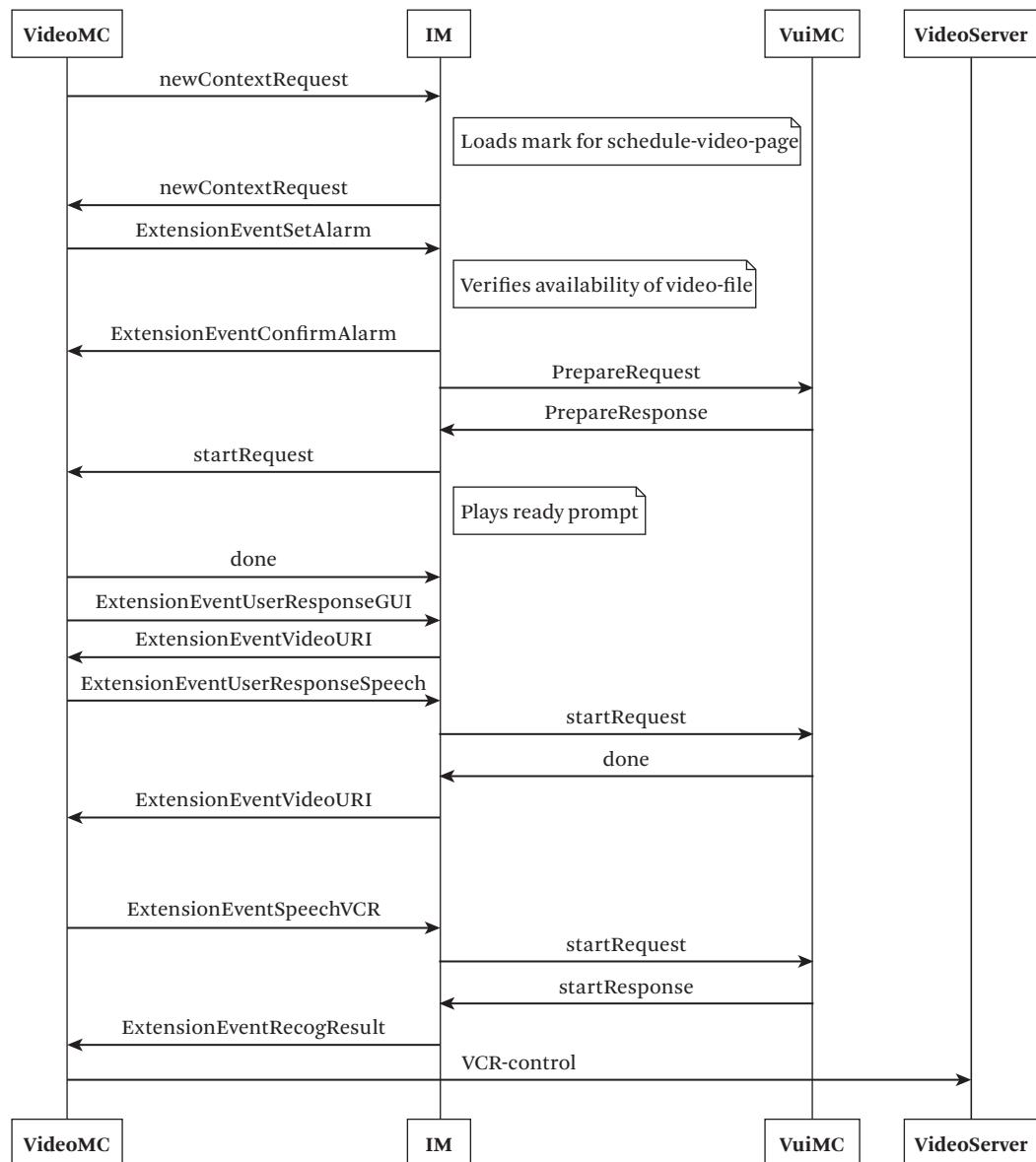
### 9.7.1 Communication and Event Processing

The W3C MMI Architecture provides support for interaction management in a declarative way; that is, describing the desired results of processing as opposed to describing the processing itself. It is designed to be used for dialog management, synchronization in a distributed environment, context management, and managing application data. It also supports managing history for dialog management, as well as accessing user preferences and static device profiles for output generation.

Because all communication between each modality component happens through the interaction manager, all data received by one modality component has to be routed to the interaction manager and back to the client—including any data and other information. This makes it particularly inefficient when both modality components are running on the same device and in the same process.

As a user navigates through the multimodal application, it results in lot of messaging between MCs, IM, and RTF as can be seen through the ladder diagram given below in Figure 9.14. The diagram shows the flow of events over time (from top to bottom) between the video modality component, the interaction manager, and the voice modality components. where:

- the user selects the schedule for video play through Device-GUI;
- at the scheduled time, a voice-prompt is played informing user about the video about to be played; and
- The user can accept or decline play through speech or GUI.



**Figure 9.14** An example of the interaction between a video component, a voice component, the interaction manager, and a video server (the vertical lines represent the components of a system while the horizontal lines the messages they exchange).

The user has the ability to use VCR-like controls through speech or GUI.

Note that the VideoServer is outside of the multimodal architecture system and is directly controlled by the VideoMCs media-player sub-component. Consequently, it does not use MMI-life-cycle events (e.g., the gui-VCR control does not result in an MMI event to the IM, instead is directly handled by VideoMC).

### 9.7.2 Modality Fusion

Currently, there is no standardization of how multimodal fusion is performed, except through the standard input/output format EMMA, on how modality fusion across distributed MCs and IMs occurs. This is intentional, because it was believed that attempting to standardize the fusion step at this point would be premature. By not attempting to define standard fusion processing, the fusion step is left open to future experimentation, for example, as discussed in [Schnelle-Walka et al. 2017].

### 9.7.3 Runtime Framework (RTF) and Extension Notification

Given the objective of the specification to maintain an open architecture that facilitates seamless integration of evolving technology components from groups of vendors, the ability to discover the presence and capabilities of these components at runtime becomes necessary, as not all components are authored and started at the same time.

In addition, the order in which the components and IM are started cannot be always guaranteed and the IM/RTF should have a way to start the MCs, should they become unavailable during the execution of an MMI application. The “StartIt” application, described in Martin et al. [1996] is an example of an application addressing this requirement. Similarly, the MCs should have a way to announce their availability to the IM/RTF. The current architecture specification does not address the events between the RTF and the MCs/IM.

In order to achieve control over discovery, starting and stopping of components, MCs and the IM/RTF need information about the connectivity and profile of a given execution environment for startup and communication of events. The details on Connectivity and Profile are discussed in the Discovery and Registration specification [Rodríguez et al. 2015].

The registration of MCs ensures that the IM and RTF know of components’ availability, capability, connection and invocation information such as protocols and ports. In addition, dynamic properties and capabilities of the modality components can be updated through status messages.

Components can register their profiles and capabilities with the RTF and the IM through the *newContextRequest* event. Currently, the generic ExtensionNotification

event is used by each implementer to achieve this communication, because it has not been defined by the W3C MMI architecture specification.

**No Standard way to Synchronize User Preferences** Users of multimodal applications should be able to select their preferences for use with various modalities of interaction such as:

- language, Visual: Font, style, color, and size;
- voice: Font, speed, and volume;
- application-preferences: automatically read output on this page, show or hide menu-options a particular way;
- location-related: for example, home location, screen orientation: lock or adapt to current orientation.

In order to collect user preferences, users can be prompted at the time of first use of the application. These initial user preferences should be stored by the application (with provision for subsequent modification), and then queried at runtime.

The Openstream CueMe platform implements a number of these standards, including InkML, EMMA, SMIL, and SCXML. This platform has been successfully used for many different customers. The standards have worked well for most applications; however, the most important current challenge we see is extending the standards to dynamic systems—more specifically, creating a standard way to discover and register new components, as discussed below.

## **9.8 The Future and Open Challenges**

### **9.8.1 The Future of Standardized Representations**

The technical complexity of many input modalities, such as the speech and ink examples discussed above, and the need for interoperability between components mean that standard APIs to modalities will become even more important as multimodal user interfaces increase in popularity. In fact, it seems clear that additional standards will develop, and existing standards will be refined as, through experience, the industry converges on best practices in multimodality.

Standardization can occur at many levels, and it is not always necessary to completely standardize all levels in order to gain the benefits of at least some standardization. As an example, we have seen in this chapter that EMMA builds on

an existing well-defined syntactic standard, XML, by adding specific elements and attributes to standardize the meta-data that has been shown to be of value in representing multimodal inputs and outputs. However, the representation of the exact semantic content expressed by the user or system is left open in EMMA. As multimodal systems become more widespread, and technologies like natural language understanding become more widely used, we believe that a natural consensus will tend to emerge on the best formats and vocabularies for semantic content. This is clearer for output than it is for input, since standard system output is already very widely used, HTML being an obvious example. This natural consensus will introduce new candidates for standardization. In commercial natural language systems such as Facebook wit.ai [wit.ai 2015], Microsoft LUIS [Microsoft 2015a], Google DialogFlow (originally api.ai) [DialogFlow 2018], Nuance Mix [Nuance 2016] and the Amazon Alexa Skills Kit [Amazon 2016], we can already see some movement toward a common vocabulary for describing the semantics of user input. We are starting to see terminology like “domain” to refer to an overall topic, “intent” for a user utterance goal and “entity” for something that a user is talking about, used in SDK’s from these different vendors. It is easy to imagine this kind of terminology first becoming an industry-wide common vocabulary and then a formal standard as different products begin to use the same vocabulary for the same concepts.

Similarly, as commercial natural language understanding systems increase in complexity, capabilities that are currently more common in academic research systems will start to be added to commercial systems. We will begin to see linguistic features like coreference, named entity recognition, quantifier scope, temporal reference, hypothetical situations, and relationships between propositions in commercial systems. Consequently, representations originating in academic systems will start to become included in standards for commercial systems. For example, some of the ISO standards and proposed standards for dialog acts [Bunt et al. 2012], semantic roles [Bonial et al. 2011] and events [Pustejovsky et al. 2010] could become integrated with more application-oriented standards such as EMMA.

New challenges to standards are also likely to be introduced by the diverse form factors and capabilities of newer devices. These include, for example, connected things in the Internet of Things, smart speakers, wearable technology, and robots. While there has been some discussion of using EMMA to represent sensor input as well as user input [Johnston 2009], this use case has not been extensively explored. Similarly, using EMMA to represent output commands to control a robot may introduce new requirements for metadata appropriate for that use case.

Not only are new types of devices constantly being introduced, but multi-device applications such as second screens and multi-user collaborations will create re-

uirements for coordinating presentations and inputs from more than one location. SCXML and the MMI Architecture should provide starting points for applications that require coordination across devices, but again, fully-realized applications may introduce new requirements for the standards.

### **9.8.2 Open Challenges with Standardization of Registration and Discovery of Multimodal Modality Components**

As the number and variety of connected things grows, the ability to discover, register, and integrate available modality services into dynamic systems will become increasingly important. An initial outline for state management of devices in dynamic systems has been proposed in [Rodríguez et al. 2015], which adds a Resource Manager to the W3C Multimodal Architecture [Barnett et al. 2012]. The Resource Manager handles the states of new components as they enter the system, become available for processing, and leave the system. Two new events are also proposed for communication between Resource Managers and components. These events communicate the availability and status of components. Since this problem resembles the problem of discovering and using Web Services, some of the concepts in the Web Services Description Language [Chinnici et al. 2007] may be of value in addressing the discovery of modality services.

Another aspect of registration and discovery that must be addressed is the fact that Resource Managers need to understand the capabilities of discovered connected devices in order to determine whether the component is a candidate for inclusion in a dynamic system. Vocabularies for modality capabilities are needed for this purpose. Ontology languages such as the Web Ontology Language (OWL) (W3C OWL Working Group 2012) can provide framework languages for representing vocabularies. There is also some work being done on vocabularies specifically for connected devices (a good survey is provided in [Lefort et al. 2011]). Using common vocabularies, Resource Managers can query potential modality components that it discovers to determine if the components meet the application's requirements, and so should be included in the system.

Not only is a common vocabulary necessary, but there needs to be an extensibility mechanism for capabilities, since the set of possible capabilities is not static, but is continually developing. At least two approaches to extensibility are possible. One is to add an open-ended extensibility point, that is, a place in a standard message or document where the content is not defined by the specification, but can be application-specific. An extensibility point can be added to a modality description in a place where new vocabularies can be specified. This is similar to the extensibility points in EMMA (`<emma:info>`) and the MMI Architecture (`<mmi>Data>`),

which can contain arbitrary, application-specific data. The other approach is to allow the standard framework to refer to a registry of vocabularies, as described in EmotionML [Burkhardt et al. 2014], specifying the vocabulary to be used. Although both approaches will work, the second approach is probably preferable, where possible, because the extensibility options are constrained to values in the registry rather than being completely open-ended.

The final issue with discovery and registration involves the need to be able to interoperate with systems that include devices using different low-level protocols for transport. There are currently a large number of existing protocols, including UPnP, Zigbee, Bluetooth Low Energy, and ECHONet Lite, for example. Any solution to discovery and registration must involve a stacked approach that supports abstraction over these low-level protocols from the developer perspective in order to have any hope of interoperability. Discovery and registration standards will be supplementary to user interaction standards such as EMMA because discovery and registration happens before the system's interaction with the user has begun. While still at an early stage, the MMI Discovery and Registration Working Draft [Rodríguez et al. 2015] supports this kind of a stacked architecture, and seems like a promising approach to addressing these requirements for dynamic multimodal systems.

### 9.8.3 Implementations

There are many systems which have been built using the W3C multimodal standards, by both participants in the standards efforts as well as by other developers. [Johnston et al. 2008] describes the 11 implementations that participated in the W3C implementation test for EMMA, including implementations from AT&T, Avaya, Conversational Technologies, Deutsche Telekom, Microsoft, Nuance, and DFKI. EMMA has also been used in conjunction with InkML to represent the results of handwriting recognition in Microsoft Office [Microsoft 2015b]. Several multimodal applications using EMMA developed at AT&T are reported in Johnston [2009]. EMMA has also been used for audio alignment in language learning applications [Dahl 2017].

In addition to the five implementations documented in the implementation report [Wiechno et al. 2012], The Multimodal Architecture has been used in a variety of applications, including personal assistants for the elderly [Teixeira et al. 2014], accessible TV [Ashimura et al. 2014], smart cities [Pous and Ceccaroni 2010] and a natural language understanding web service [Dahl 2017]. Its applicability to multi-device scenarios is discussed in Almeida et al. [2017]. An interoperability test between implementations from Openstream, France Telecom, and Deutsche Telekom is reported in Kliche et al. [2012].

EmotionML was implemented by nine companies and universities as part of the implementation report process [Burkhardt et al. 2012]. EmotionML is used for sentiment analysis in text [Empathic Products 2016]. Open-source implementations include the Mary TTS system implemented in Java and developed at DFKI [Schroeder et al. 2010] and a Python implementation [Begoli 2013].

In addition to the Microsoft Office InkML implementation for pen input and the Openstream InkML implementations mentioned above, InkML has also been used for mathematical collaboration [Hu and Watt 2013].

## 9.9

### Conclusions

As multimodal technology improves, systems will start to include more and more different modalities, which will have been developed by many different vendors. Integrating multiple modality components into systems will be very difficult if vendors do not agree at a minimum on their output and input formats. In addition, declarative scripting languages such as SCXML, as well as model-based interfaces, will help reduce the conceptual complexity of defining dialogs and building multimodal applications. The standards discussed in this chapter include EMMA, InkML, and EmotionML for representing inputs and outputs, the MMI Architecture Life Cycle Events for communication, and SCXML for defining dialogs. The standards have been implemented in a number of systems, and suggestions for additional work have included a standard way to represent user preferences and mechanisms for discovery and registration of the components of dynamic systems.

### Focus Questions

- 9.1. What are the benefits of standards for multimodal applications? What is the role of W3C in this respect? What are the disadvantages of standard languages for multimodal applications?
- 9.2. What is the role of the Facilitator agent in OAA and of the Interaction Manager in the W3C Multimodal Interaction Architecture? Why is it crucial for a multimodal system?
- 9.3. What are the benefits of encapsulation of modalities in multimodal architectures? Are there any disadvantages?
- 9.4. Describe how EmotionML represents emotions and how it allows embedding of emotion annotations in other markup languages.

**9.5.** What is multimodal fusion and how does EMMA support the five important requirements for representing fused multimodal input? Why is it important to know the original modalities of each input in fusion?

**9.6.** What is multimodal fission and how does EMMA support it?

## References

- G. Adams, M. Dolan, S. Hayes, F. de Jong, P.-A. Lemieux, N. Megitt, D. Singer, J. Smith, and A. Tai. February 2015. Timed Text Markup Language 2 (TTML2). <https://www.w3.org/TR/ttml2/>. 373
- N. Almeida, S. Silva, A. Teixeira, and D. Viera. 2017. Multi-device applications using the multimodal architecture. In D. A. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*. Springer, Berlin. DOI: [10.1007/978-3-319-42816-1\\_17](https://doi.org/10.1007/978-3-319-42816-1_17). 384
- Amazon. 2016. Alexa Skills Kit. Retrieved from <https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit>. 382
- Applied Voice Input Output Society. 2015. <http://avios.org>
- K. Ashimura, O. Nakamura, and M. Isshiki. October 2014. Accessible TV based on the W3C MMI Architecture. Paper presented at the *Consumer Electronics (GCCE), 2014 IEEE 3rd Global Conference on*. DOI: [10.1109/GCCE.2014.7031260](https://doi.org/10.1109/GCCE.2014.7031260). 384
- P. Baggio, P. Bagshaw, D. Burnett, J. Carter, and F. Seahill. 2008. Pronunciation Lexicon Specification (PLS) Version 1.0. <http://www.w3.org/TR/pronunciation-lexicon/> 354, 364
- J. Barnett. 2016. Introduction to SCXML. In D. Dahl editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*. New York: Springer. DOI: [10.1007/978-3-319-42816-1\\_5](https://doi.org/10.1007/978-3-319-42816-1_5). 375
- J. Barnett, R. Akolkar, R. J. Auburn, M. Bodell, D. C. Burnett, J. Carter, S. McGlashan, T. Lager, M. Helbing, R. Hosn, T. V. Raman, K. Reifenrath, and N. Rosenthal. 2015. State Chart XML (SCXML): State machine notation for control abstraction. <http://www.w3.org/TR/scxml/> 354, 375
- J. Barnett, M. Bodell, D. A. Dahl, I. Klische, J. Larson, B. Porter, D. Raggett, T. V. Raman, B. H. Rodriguez, M. Selvaraj, R. Tumuluri, A. Wahbe, P. Wiechno, and M. Yudkowsky. October 2012. Multimodal architecture and interfaces. <http://www.w3.org/TR/mmi-arch/> 353, 383
- S. Bayer. 2005. Building a standards and research community with the Galaxy Communicator software infrastructure. In D. A. Dahl, editor, *Practical Spoken Dialog Systems*. vol. 26, pp. 166–196). Kluwer Academic Publishers, Dordrecht. 378
- S. Bayer, C. Doran, and B. George. 2001. *Dialog interaction with the DARPA Communicator: The development of useful software*. Proceedings of HLT 2001, San Diego, CA. 378
- E. Begoli. 2013. EMLPy. <https://github.com/ebegoli/EMLPy> 385

- C. Bonial, W. Corvey, M. Palmer, V. Petukhova, and H. Bunt. 2011. *A hierarchical unification of LIRICS and VerbNet semantic roles*. IEEE-ICSC 2011 Workshop on Semantic Annotation for Computational Linguistic Resources, Stanford, CA. 382
- T. Bray, J. Jean Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau. 2004. Extensible Markup Language (XML) 1.0 (Third Edition). <http://www.w3.org/TR/2004/REC-xml-20040204/> 357
- D. Bulterman, J. Jansen, P. Cesar, S. Mullender, E. Hyche, M. DeMeglio, J. Quint, H. Kawamura, D. Weck, X. G. Pañeda, D. Melendi, S. Cruz-Lara, M. Hanclik, D. F. Zucker, and T. Michel. 2008. Synchronized Multimedia Integration Language (SMIL 3.0). <http://www.w3.org/TR/2008/REC-SMIL3-20081201/> 365, 373
- H. Bunt, J. Alexandersson, J.-W. Chae, A. C. Fang, K. Hasida, O. Petukhova, A. Popescu-Belis, and D. Traum. 2012. *ISO 24617-2: A semantically-based standard for dialogue annotation*. Paper presented at the Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. 382
- F. Burkhardt, M. Schröder, and P. Baggio. 2012. EmotionML 1.0: Implementation Report. <http://www.w3.org/2002/mmi/2013/emotionml-ir/> 385
- F. Burkhardt, M. Schröder, and C. Pelachaud. 2014. Vocabularies for EmotionML. <http://www.w3.org/TR/emotion-voc/> 369, 384
- D. C. Burnett, M. R. Walker, and A. Hunt. 2004. W3C Speech Synthesis Markup Language (SSML). <http://www.w3.org/TR/speech-synthesis/> 354, 364
- J. M. Cantera, J. González, G. Meixner, F. Paternò, J. Pullmann, D. Raggett, D., Schwabe, and J. Vanderdonckt. 2010. Model-Based UI XG Final Report. <http://www.w3.org/2005/Incubator/model-based-ui/XGR-mbui-20100504/> 358, 362
- A. Cheyer and D. Martin. 2001. The Open Agent Architecture. *Autonomous Agents and multi-agent Systems*, 4(1–2): 143–148. DOI: [10.1023/a:1010091302035](https://doi.org/10.1023/a:1010091302035). 350
- R. Chinnici, J.-J. Moreau, A. Ryman, and S. Weerawarana. 2007. Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language. <https://www.w3.org/TR/wsdl20/> 383
- P. R. Cohen. 1992. *The role of natural language in a multimodal interface*. Proceedings of the 5th Annual ACM Symposium on User Interface Software and Technology (UIST '92), New York, NY. 355
- P. R. Cohen and S. L. Oviatt. 2017. Multimodal speech and pen interfaces. In S. Oviatt, B. O. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3015783.3015795](https://doi.org/10.1145/3015783.3015795). 366
- P. R. Cohen, A. Cheyer, M. Wang, and S. C. Baeg. March 1994. *An Open Agent Architecture*. AAAI Spring Symposium. 350

- J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. Young. 1995. *Four easy pieces for assessing the usability of multimodal interaction: the CARE properties*. Proc. of the INTERACT '95 - IFIP TC13 Fifth International Conference on Human-Computer Interaction, Lillehammer, Norway. [http://iigm.imag.fr/publis/1995/Interact95\\_CARE.pdf](http://iigm.imag.fr/publis/1995/Interact95_CARE.pdf) 361
- R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1): 32–80. DOI: [10.1109/79.911197](https://doi.org/10.1109/79.911197). 369
- C. Cross, J. Axelsson, G. McCobb, T. V. Raman, and L. Wilson. 2003. XHTML+Voice Profile 1.1. <http://www-306.ibm.com/software/pervasive/multimodal/x+v/11/spec.htm> 377
- D. A. Dahl. 2017. A standard portal for intelligent services. In D. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*. Berlin: Springer. DOI: [10.1007/978-3-319-42816-1\\_11](https://doi.org/10.1007/978-3-319-42816-1_11). 384
- D. A. Dahl, E. Coin, M. Greene, and P. Mandelbaum. 2011. A conversational personal assistant for senior users. In D. Perez-Marin & I. Pascual-Nieto (Eds.), *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. pp. 282–301. Hershey, PA, IGI Global. DOI: [10.4018/978-1-60960-617-6.ch012](https://doi.org/10.4018/978-1-60960-617-6.ch012). 364
- D. A. Dahl and B. Dooner. 2017. Audio alignment for multimedia language learning: Applications of SRGS and EMMA in Colibro Publishing. In D. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*. Berlin: Springer. DOI: [10.1007/978-3-319-42816-1\\_14](https://doi.org/10.1007/978-3-319-42816-1_14).
- DialogFlow. 2018. <http://dialogflow.com/> 382
- Empathic Products. 2016. SATI API. <http://portal.empathic.eu/?q=products/sati-api> 385
- R. T. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. 1999. RFC 2616 Hypertext Transfer Protocol—HTTP/1.1. <http://tools.ietf.org/html/rfc2616> 350, 352
- D. Harel. 1987. Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, 8: 231–274. DOI: [10.1016/0167-6423\(87\)90035-9](https://doi.org/10.1016/0167-6423(87)90035-9). 375, 376
- R. Hu and S. M. Watt. 2013. *InkChat: A collaboration tool for mathematics*. Mathematical User Interfaces Workshop 2013, (MATHUI 2013), Bath, U.K. 385
- A. Hunt and S. McGlashan. 2004. W3C Speech Recognition Grammar Specification (SRGS). <http://www.w3.org/TR/speech-grammar/> 354, 364
- IBM. 2001. XHTML+Voice. <http://www.w3.org/Submission/2001/13/> 377
- IBM. 2003. X+V 1.1. <http://www-3.ibm.com/software/pervasive/multimodal/x+v/11/spec.htm> 377
- R. James, J. Ivar, and B. Grady. 1999. *The Unified Modeling Language Reference Manual*. Addison-Wesley Longman Ltd. 375
- M. Johnston. 2009. *Building multimodal applications with EMMA*. International Conference on Multimodal Interfaces, Cambridge, MA. 354, 365, 370, 382, 384

- M. Johnston. 2016. Extensible multimodal annotation for intelligent interactive systems. In D. Dahl, editor, *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything*. New York: Springer. DOI: [10.1007/978-3-319-42816-1\\_3](https://doi.org/10.1007/978-3-319-42816-1_3). 370
- M. Johnston. 2019. Multimodal integration for interactive conversational systems. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 364, 366
- M. Johnston, M. Johnston, P. Baggio, D. C. Burnett, J. Carter, D. Dahl, and K. Ashimura. 2008. EMMA: Extensible MultiModal Annotation 1.0: Implementation Report. <http://www.w3.org/2002/mmi/2008/emma-ir/> 384
- M. Johnston, P. Baggio, D. Burnett, J. Carter, D. A. Dahl, G. McCobb, and D. Raggett. 2009a. EMMA: Extensible MultiModal Annotation markup language. <http://www.w3.org/TR/emma/>
- M. Johnston, D. A. Dahl, I. Kliche, P. Baggio, D. C. Burnett, F. Burkhardt, and K. Ashimura. 2009b. Use Cases for Possible Future EMMA Features. <http://www.w3.org/TR/emma-usecases/>
- M. Johnston, D. A. Dahl, T. Denny, and N. Kharidi. 2015. EMMA: Extensible MultiModal Annotation markup language Version 2.0. <http://www.w3.org/TR/emma20/> 354, 370, 372
- M. Ketsmur, A. Teixeira, N. Almeida, S. Silva, and M. Rodrigues. 2018. Conversational Assistant for an Accessible Smart Home: Proof-of-Concept for Portuguese. *DSAI 2018*. Thessaloniki, Greece. [https://www.researchgate.net/publication/325155533\\_Conversational\\_Assistant\\_for\\_an\\_Accessible\\_Smart\\_Home\\_Proof-of-Concept\\_for\\_Portuguese](https://www.researchgate.net/publication/325155533_Conversational_Assistant_for_an_Accessible_Smart_Home_Proof-of-Concept_for_Portuguese). 378
- I. Kliche, N. Kharidi, and P. Wiechno. January 2012. MMI interoperability test report. <http://www.w3.org/TR/2012/NOTE-mmi-interop-20120124/> 384
- S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. 2006. *Towards a Common Framework for Multimodal Generation: The Behavior Markup Language*. International Conference on Intelligent Virtual Agents, Marina del Rey, CA. 373
- L. Lefort, C. Henson, and K. Taylor. 2011. Semantic Sensor Network XG Final Report. <https://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/> 383
- J. López Herrera and H. Ríos-Figueroa. 2018. JaCa-MM: A user-centric BDI multiagent communication framework applied for negotiating and scheduling multi-participant events A Jason/Cartago extension framework for diary scheduling events permitting a hybrid combination of multimodal devices based on a microservices architecture. *ICAART 2018: 10th International Conference on Agents and Artificial Intelligence*, Funchal, Madeira, Portugal. 378

- D. L. Martin, A. J. Cheyer, and G.-L. Lee. 199. *Agent development tools for the Open Agent Architecture*. Proc. of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, London. **380**
- Microsoft. 2015a. Language Understanding Intelligent Service (LUIS). <http://www.projectoxford.ai/luis> **382**
- Microsoft. 2015b. Office Drawing Extensions to Office Open XML Structure. [http://download.microsoft.com/download/D/3/3/D334A189-E51B-47FF-B0E8-C0479AFB0E3C/\[MS-ODRAWXML\].pdf](http://download.microsoft.com/download/D/3/3/D334A189-E51B-47FF-B0E8-C0479AFB0E3C/[MS-ODRAWXML].pdf) **384**
- H. Mouchere, C. Viard-Gaudin, R. Zanibbi, and U. Garain. January 2016. ICDAR 2016 CROHME: Third International Competition on Recognition of Online Handwritten Mathematical Expressions. *ICDAR 2016: 18th International Conference on Document Analysis and Recognition*, Johannesburg, South Africa. **367**
- Nuance Communications. 2016. Nuance Mix. <https://developer.nuance.com/public/index.php?task=home> **382**
- Openstream, Inc. 2015. Eva, Face of the Digital Workplace. Retrieved from <http://www.openstream.com/eva.html> **350**
- M. Oshry, R. J. Auburn, P. Baggia, M. Bodell, D. Burke, D. C. Burnett, E. Candell, J. Carter, S. McGlashan, A. Lee, B. Porter, and K. Rehor. 2007. Voice Extensible Markup Language (VoiceXML) 2.1. <http://www.w3.org/TR/voicexml21/> **354**
- S. Oviatt. 2017. In S. Oviatt, B. W. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces: Volume 2, Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan and Claypool Publishers, San Rafael, CA. **368**
- F. Paternò. 2003. ConcurTaskTrees: An Engineered Notation for Task Models. In D. S. S. Diaper, N. editor, *The Handbook of Task Analysis for Human-Computer Interaction*. pp. 483–503. Lawrence Erlbaum Associates, Mahwah, NJ. **359**
- G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Hermos. 2017. Audio and Visual Modality Combination in Speech Processing Applications In S. Oviatt, B. O. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*, pp. 489-543. Morgan & Claypool Publishers, San Rafael, CA. **369**
- M. Pous and L. Ceccaroni. 2010. Multimodal Interaction in Distributed and Ubiquitous Computing. *Internet and Web Applications and Services (ICIW)*, 2010 Fifth International Conference on. **384**
- J. D. Pustejovsky, H. C. Bunt, K. Lee, and L. Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. **382**

- S. Quiniou, M. Cheriet, and E. Anquetil. 2009. Design of a framework using InkML for pen-based interaction in a collaborative environment. *Proceedings of the International Conference on Human Computer Interaction (HCI International)*. 367
- B. H. Rodríguez, J. Barnett, D. Dahl, R. Tumuluri, N. Kharidi, and K. Ashimura. 2015. Discovery and Registration of Multimodal Modality Components: State Handling. <https://www.w3.org/TR/mmi-mc-discovery/> 380, 383, 384
- SALT Forum. 2002. Speech Application Language Tags (SALT). <http://www.saltforum.org> 377
- D. Schnelle-Walka, C. Duarte, and S. Radomski 2017. Multimodal fusion and fission within the W3C MMI architectural pattern. In D. A. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural Interaction with Everything*. pp. 393–415. Switzerland: Springer. DOI: [10.1007/978-3-319-42816-1\\_19](https://doi.org/10.1007/978-3-319-42816-1_19). 380
- M. Schröder, P. Baggio, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato. May 2014. Emotion Markup Language (EmotionML) 1.0 <http://www.w3.org/TR/emotionml/> 354, 368
- M. Schröeder, S. Pammi, R. Cowie, G. McKeown, H. Gunes, M. Pantic, M. Valstar, D. Heylen, M. ter Maat, F. Eyben, B. Schuller, M. Wöllmer, E. Bevacqua, C. Pelachaud, and E. de Sevin. 2010. Demo: Have a Chat with Sensitive Artificial Listeners. *AISB 2010 Symposium “Towards a Comprehensive Intelligence Test,”* Leicester, UK. DOI: [10.1.1.364.3795](https://doi.org/10.1.1.364.3795). 385
- T. Schwartz and M. Feldt. 2018. Software Platforms and Toolkits for Building Multimodal Systems and Applications. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, & A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA.
- D. Sonntag. 2019. Multimodal interaction for medical and health systems. In S. Oviatt, B. W. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors , *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 366
- A. Teixeira, A. Hämäläinen, J. Avelar, N. Almeida, G. Németh, T. Fegyó, C. Zainkó, T. Csapó, B. Tóth, A. Oliveira, and M. S. Dias. 2014. Speech-centric multimodal interaction for easy-to-access online services—A personal life assistant for the elderly. *Procedia Computer Science*, 27, 389–397. DOI: [10.1016/j.procs.2014.02.043](https://doi.org/10.1016/j.procs.2014.02.043). 384
- L. Van Tichelen and D. Burke. April 2007. Semantic Interpretation for Speech Recognition. [http://www.w3.org/TR/semantic-interpretation/](https://www.w3.org/TR/semantic-interpretation/) 354, 364
- W3C. 2000. Extensible Markup Language (XML) 1.0 (Second Edition). <http://www.w3.org/TR/REC-xml>
- W3C OWL Working Group. 2012. OWL 2 Web Ontology Language Document Overview (Second Edition). <http://www.w3.org/TR/owl2-overview/>

- S. M. Watt, T. Underhill, Y.-M. Chee, K. Franke, M. Froumentin, S. Madhvaniath, J.-A. Magaña, G. Pakosz, G. Russell, M. Selvaraj, G. Seni, C. Tremblay, and L. Yaeger. 2011. Ink Markup Language (InkML). [http://www.w3.org/TR/InkML\\_354](http://www.w3.org/TR/InkML_354), 364
- P. Wiechno, N. Kharidi, I. Kliche, B. H. Rodriguez, D. Schnelle-Walka, D. A. Dahl, and K. Ashimura. 2012. Multimodal Architecture and Interfaces 1.0 Implementation Report. <http://www.w3.org/2002/mmi/2012/mmi-arch-ir/> 384
- wit.ai. 2015. wit.ai. <https://wit.ai/> 382

# Multimodal Databases

Michel Valstar

## 10.1

### Introduction

In the preceding chapters, we have seen many examples of Multimodal, Multisensor Interfaces (MMIs). Almost all of these interfaces are implemented as computer systems, and many of these have “intelligent” components that are increasingly created using Machine Learning techniques. Many of these systems are developed and/or evaluated using databases of interactions with such MMI systems, or databases of scenarios where humans or Wizard of Oz systems replace what could be an MMI. Often, the developed MMI systems are only as good as the data that was used to train or otherwise develop them. Or, seen from another perspective, the quality of data is a limiting factor of system performance, but if one has good data, good systems can arise naturally from them. It is thus of utmost importance to have access to the highest-quality data that is most relevant to your study or system. Creating good datasets requires expertise, space, time, and money. It is therefore always preferable to use existing data, and one should check if there are any existing databases that fulfil your requirements, before embarking on the painstaking effort of creating your own.

This chapter will help you to refine your data requirements and make them explicit, by first elaborating on the motivation for collecting sets of multimodal interactions (Section 10.2), and then providing a structured overview of existing publicly available Multimodal-Multisensor Interface databases (Section 10.3). Finally, in case you find that your existing need for data has not been met by existing databases, the chapter will pass on some of the valuable lessons learned in creating databases in Section 10.4.

## 10.2

### Need for Data

In the domain of Multimodal Interfaces, databases are generally acquired for three reasons: (i) to obtain the data necessary for developing systems with novel or improved capabilities, often achieved by training new machine learning hypotheses on hand-labeled data or by transferring existing systems to a new domain, again often achieved by (partial) labeling of the data. In this chapter we will call these **development databases**; (ii) to evaluate the performance of systems on a new or extended domain in a repeatable and objective manner, thus allowing comparison with other systems. We will call these **benchmark databases**; and (iii) to create new knowledge by evaluating theoretical hypotheses (as opposed to Machine Learning hypotheses; see [Mitchell \[1997\]](#)). We will call these **hypothesis testing databases**. The motivation underpinning data collection should have a significant impact on how a good MMI database is collected. In particular, when a database is designed properly, each of the three motivations listed above will result in a different set of criteria and will thus ultimately result in very different types of databases.

*Development databases* created to train or develop Multimodal Interface (sub)-systems clearly need to be comprised of large amounts of data, sufficient to train modern machine learning systems. The exact number depends on the task, but we are talking at least thousands of examples per phenomenon of interest, often many more. Note that this does not necessarily mean you need to collect thousands of people—the same person may elicit the same phenomenon of interest many times during a single interaction. Second, development databases need to be generic enough to be useful in multiple scenarios. The capabilities learned from such databases should be as generic as possible, and the data should reflect that. When building a smile detector, we want it to perform in any condition, not just the lab. This does not mean you can't record in the lab, but it does mean that you need to ask yourself clearly what variables the system should be able to generalize to. Good examples are generalizability to different people (meaning you need significant numbers of participants), generalizability across demographic variables such as ethnicity, gender, or age, and generalizability to environmental conditions such as illumination, noise, or social setting. Third, development databases need to be realistic enough so that the systems developed with them can be used in real use cases.

Note that many researchers will here call for databases that are ready to be used *in the wild* [[Crabtree et al. 2013](#), [Rogers 2011](#)]. While this is a useful term in what could be thought of as a ‘call to arms’ for making datasets more widely applicable, this has sometimes mistakenly been construed as meaning that datasets should be created in such a way that they can be used to train systems to do absolutely

## Glossary

**Affective Computing** is a scientific interdisciplinary field involving computer and cognitive sciences, psychology and philosophy, concerned with the conceptualization and implementation of systems that recognize and simulate human emotions and empathy.

**AIBO Robot** is a robotic dog targeted at end customers that has been developed and sold by Sony since 2006.

**Arousal** is used in psychology to describe the degree of stimulation of the nervous related to emotions. It is usually seen as an integral part of the experience of emotions.

**EEG** is the acronym for electroencephalography. It is a monitoring method to record the electrical activity of the brain. It usually requires electrodes to be placed on the subjects scalp or head. It provides good temporal but poor spatial resolution of the electrical activities.

**6DOF** stands for six degrees of freedom and usually describes the freedom of position through translation (3 dimensions) and orientation through rotation (3 angles) of an object.

**Microsoft Kinect** is a device that has been developed and sold by Microsoft from 2009–2017 as a controller for the Xbox computer console series. It uses a depth- and RGB-camera and a microphone array to process human gestures and body postures, which are used as input. It has been widely adopted in human-computer-interaction research due to its cheap availability and easy programmability.

**Network Time Protocol** is a networking protocol to synchronize clocks over data networks. It is based on a client-server-model and uses standard internet protocol layers.

**Raspberry PI** is a series of small computers developed and sold by the Raspberry PI foundation since 2012 usually used in teaching of basic computer science skills. Due to its small form factor it is also often used in robotics and Ubiquitous Computing research.

**RFID** stands for Radio Frequency Identification, which is a technology to track and identify objects, which are usually equipped with respective tags, read by an RFID-reader. RFID can provide an inexpensive mean for tracking and identification, in particular if passive RFID tags are used that do not rely on an internal battery, since they are powered by the electromagnetic field of the reader.

**Valence** is one dimension used in psychology to describe emotions. It is related to the intrinsic attractiveness or averseness of an emotion. Often those are referred to as positive and negative valence respectively.

**Virtual Human** is a digital anthropomorphic representation of a system with which users can interact with. It combines realistic rendering techniques with intelligent behavior to make the virtual human as believable as possible.

**Wizard of Oz** is an experimental methodology in the field of human-computer interaction. It is used to evaluate functionality of a system that has yet not been implemented, but is controlled by a wizard. Experimental subjects usually do not know that the functionality of the system is provided by a non-visible human and thus behave as if the system would function autonomously.

anything. This is not what *in the wild* research means though: deploying a facial expression recognition system in a single room in a hospital for routine use may sound like a very constrained and well-defined environment, but this is actually a prime example of *in the wild* deployment of MMI systems. Clearly one cannot collect databases that are optimal for all use-cases, yet on the other hand, the previously stated need for generizability does mean it should not be limited to one single use case. The result is that those who construct databases need to carefully consider the scope of use cases they wish to address; clearly, there is a balance to be struck in this regard.

Development databases are often widely shared between MMI practitioners, although researchers often delay making the data public for some time so they can have a head-start in developing novel methods based on these hardly won data.

*Benchmark databases*, i.e., databases created to evaluate the performance of MMI systems first need to be large enough to easily result in statistically significant results. Second, they need to clearly focus on a particular (set of) interaction scenario(s), which should be replicated/represented as realistically as possible in order to be able to proof that a system will work in the intended situation. Third, they should be collected with ample meta-data of the recordings so that it is possible to evaluate different aspects of a system (e.g., gender specificity). Often it is not entirely known beforehand what aspects of a system you want to evaluate, so it is wise to make a record of as much meta-data as possible.

Benchmark databases are by their very definition intended to be shared as widely as possible. Benchmark databases increase their value as a particular interaction or interface becomes more mainstream. By definition, they facilitate comparison between state of the art methods, and as such they can only truly be created once a community has agreed on the problem definition and corresponding measures of success.

*Hypothesis testing databases* on the other hand are used for knowledge generation rather than system creation or algorithm evaluation, e.g., studying cognitive load of humans when using novel types of interfaces. Therefore, there is no need to replicate a real use-case scenario. It is entirely justified to design elaborate but unrealistic experimental scenarios, as long as they facilitate providing a definitive test of the hypothesis. The amount of data can often be fairly small, as the hypotheses are well defined beforehand, which allows researchers to limit the experimental variables and thereby maximize the statistical power. For example, if gender effects are not the focus of a particular study, one can choose to have only male-male and female-female interactions, thus removing a complex factor of social interactions in human-human studies.

In the field of MMI it is much rarer, although not entirely without precedent, for hypothesis testing databases to be collected and subsequently shared publicly. This is because it is generally unnecessary to share this data—once the hypothesis is tested and the new knowledge has been determined, the data has served its purpose. One could argue that sharing the data would allow other researchers to replicate the findings. However, performing the replication study on the same data risks missing potential flaws in the data collection process itself. You could thus argue that it might even be better not to share such databases, as long as the process of data collection is explained in great detail in any manuscripts reporting on studies performed on the data. In this chapter we will focus on publicly available datasets, and as such we will not discuss databases that were created for knowledge generation.

The area of Affective Computing [[Picard 1997](#)] has a brief but active tradition of collecting and sharing databases. In a 2005 review paper, Pantic et al. mentioned that none such databases were available [[Pantic et al. 2005](#)], yet now 12 out of 20 of the databases listed in this chapter can be called Affective Computing databases. In fact, there is a strong overlap between affective datasets and MMI datasets. Affect is per definition a multimodal concept, and indeed most affective computing datasets are multimodal (bar a few very old ones, e.g., [[Lucey et al. 2010](#), [Valstar and Pantic 2010](#)]). It is true that some databases were recorded with a focus on studying human behaviour rather than creating novel interfaces. However, endowing interfaces with a sense of affect has been an MMI goal for a long time, and the majority of these datasets were either directly recorded with the intention to create more human-like interfaces or could be used to that aim without any modification.

It is thus valuable to take some inspiration from the literature on Affective Computing datasets. In particular, [Picard et al. \[2001\]](#) outlined five factors that influence affective data collection.

- Spontaneous vs. posed: Is the emotion elicited by a situation or stimulus outside the subject's control or is the subject asked to elicit the emotion?
- Lab setting vs. real-world: Is the data recording taking place in a lab or is the emotion recorded in the usual environment of the subject?
- Expression vs. feeling: Is the emphasis on external expression or on internal feeling?
- Open recording vs. hidden recording: Is the subject aware that they are being recorded?

- Emotion-purpose vs. other-purpose: Does the subject know that they are part of an experiment and the experiment is about emotion?

While these considerations were written down to describe important properties of datasets of *affective* data, they can be altered and adopted to hold for any Multimodal Multisensor Interface interaction databases. In particular, the following categorization of MMI databases is useful.

- Real vs. abstract interfaces: is the interface a concrete, functional, and realistic implementation or is it an abstract portrayal of an interface that is possibly mocked-up, wizarded, or even just sketched?
- Lab setting vs. real-world: Is the data recording taking place in a lab or is the interaction recorded in the usual environment of the subject?
- Open recording vs. hidden recording: Is the subject aware that they are being recorded?
- Interaction-purpose vs. other-purpose recording: Does the subject know that they are part of an experiment and the experiment is about interfaces?

For example, the AIBO robot [Batliner et al. 2004] database of human-robot interactions was first recorded with a researcher controlling the robot, even though the interaction was presented to the users as being with a fully autonomous robot. Hence, it was to an extent an abstraction of the studied interface. It was also recorded in a dedicated university lab, far from the targeted scenario where children would interact with the robots, for example at home or at school. The AIBO database was open, in the sense that the children were told they were being recorded, but they were not really aware that one of the purposes of the data collection was to study the way they interacted with the robot.

We will use these properties to categorize existing databases in Section 10.3, and they should prove useful in determining what database you need for your research as well.

## 10.3

### Existing Databases

In order to provide a comprehensive yet manageable list of existing MMI databases, a structured review of the literature on this topic was conducted. The Google Scholar, ACM, arXiv, DBLP, and IEEEXplore search engines were given four queries varying the terms separated by a “slash” in the term: “Multimodal Interaction/Interface Database/Dataset.” The resulting set of papers were filtered based on the following inclusion criteria:

- presents a publicly available database;
- published after 2000 AD;
- has at least two modalities;
- has at least 25 participants/recording (whichever is larger); and
- focuses on MMI.

Some notes on the inclusion criteria: The 2000 AD cut-off is fairly arbitrary, but coincides with the publication of a seminal paper on multimodal speech and pen-based interaction [Oviatt et al. 2000], and in general this year saw a number of important surveys and position papers on MMIs, which in turn led to the creation of MMI databases. In terms of sensors, those that were clearly only added to generate automatic annotation rather than to be used as inference sources weren't taken into account (e.g., motion capture suits in combination with video cameras, where the suits provide body pose annotations for the video data). In terms of focusing on MMI: datasets that feature human-machine interaction were considered to be MMI focused. Human-human interaction databases were included only if the motivation of the data collection was to be able to replace (aspects of) one interlocutor with a machine interface in the future. This would include many Wizard of Oz studies, but excludes databases that study human behavior from a purely social sciences perspective and also databases that only collect biometric data and don't focus on interaction with MM interfaces. A fairly large number of multimodal, multisensor databases are of an affective nature. However, databases that only induce particular affective states without a readily identifiable use of an interface are again excluded. The definition of "publicly available" would include datasets restricted for academic research after signing a non-commercial exploitation EULA, as is common in the field—it would not need to be completely open to the general public.

There are a number of ways in which databases can be categorized. In this chapter, we'll distinguish between groups of databases based on (i) the mode of interaction that is being recorded: human-computer, human-robot, human-human, single human, or a group of people. Note that human-computer interaction is in itself a category that you could split further; there's a big difference between a human interacting with a power-point presentation on the one hand, and a virtual human on the other. Nevertheless, for simplicity we will both treat them as human-computer interaction, where we take "computer" to mean a personal computer. Human robot interaction is treated as a different class though, given

that the physical embodied presence of the machine drives a very different type of interaction.

Another way of categorizing databases is based on the types of sensors used. Traditionally, databases would thus be divided into audio, visual, and audio-visual databases, depending on the use of microphones, cameras, or both, respectively. We will call both microphones and cameras non-invasive sensors. However, as of late other sensors have been used as well. In particular, wearable sensors are increasingly being used, measuring in particular biological signals such as heart-rate or skin conductance. It is sometimes argued that facial expression and tone of voice are also biological signals, but for the sake of clarity we do not group them as such in this chapter. The emergence of consumer grade depth sensors (e.g., Microsoft Kinect) has meant that a large number of databases now also come with depth data. Finally, many of the interfaces collect (meta-)data that could be used to enhance the interaction, for example logging of a touch-based interface or the writing made by a digital pen.

Last, we can categorize databases based on the experimental setting used to collect the data. This kind of taxonomy can clearly be structured in many different ways. Here, we will use the following types: (i) experimental psychology, (ii) realistic scenario replication, and (iii) in the wild.

Table 10.1 gives an overview of the general experimental setting of all databases considered in this chapter, along with a reference to the authoritative paper on each particular database. Table 10.2 gives a further categorization of these databases according to the scheme set out in Section 10.2. Table 10.3 gives an overview of the size of the databases according to the number of participants and recording sessions. Finally, Table 10.4 provides some more practical details on the size and resource location of the various databases.

Publicly available MMI databases can in many ways be considered to be living, evolving entities. By their very nature, they are supposed to be used, and in their use they are often extended, either with more recordings, or with more annotations. While the properties of the databases reviewed here have been listed as precisely as possible, small deviations from the descriptions should be expected when actually working with the data.

For the interface realism, we distinguish between (from high to low realism) *real product*, *fully functional prototype*, *partially functional prototype*, and *mockup*. A real product is exactly what it says—an interface that is already on the market and can thus be deemed to be fully functional and highly polished. Examples of this are the AIBO robot [Batliner et al. 2004] or recordings of people interacting with their mobile phones. Fully functional prototypes are interfaces that are perhaps

**Table 10.1** Scenario overview of existing multimodal multisensor interface databases

Database	Reference	Interface	Modalities <sup>a</sup>	Sensors <sup>b</sup>
AIBO robot	[Batliner et al. 2004]	Human-Robot	A, V	1c, 3m
AMI	[Carletta et al. 2005]	Human-Human Group	A, V, O	≈7c, ≈12m, o
BP4D	[Zhang et al. 2016]	Mixed	A, V, D, P	1c, 1m, 2o
Canal 9	[Vinciarelli et al. 2009]	Human-Human Group	A, V	1c, 1m
CMU MMAC	[De la Torre et al. 2008]	Human-Instrument/Object	A, V, O	6c, 6m, 5o
DEAP	[Koelstra et al. 2012]	Human-Computer	P, V	1o, 1c
EmoTV Corpus	[Abrilian et al. 2005]	Human-Human	A, V	N.A.
GEMEP	[Bänziger and Scherer 2010]	Human	A, V	4c, 4m
HUMAINE	[Douglas-Cowie et al. 2007]	Human-Human	A, V, P	varies
MAHNOB-HCI	[Soleymani et al. 2012]	Human-Computer	A, V, P, G	6c, 1m, 6o
MANHOB-Laughter	[Petridis et al. 2013]	Human-Computer	A, V, T	1c, 1m, 1o
Math Data Corpus	[Oviatt et al. 2013]	Human-Human-Computer	A, V, O	5c, 4m, 3o
MHAD	[Ofli et al. 2013]	Human	A, V, D	12c, 4m, 13o
MHI-Mimicry	[Sun et al. 2011]	Human-Human	A,V	15c, 3m
NoXi	[Cafaro et al. 2017]	Human-Computer	A, V, D	2kinect, 2m
RECOLA	[Ringeval et al. 2013]	Human-Human	A, V, O	1m, 2c, 1o
SEMAINE	[McKeown et al. 2012]	Mixed	A, V	5c,4m
SMARTKOM	[Wahlster 2006]	Human-(Human aided) Computer	A, V	3m, 2c, 2o
TUM Kitchen	[Tenorth et al. 2009]	Human	V, O	4c, 2o
VAM	[Grimm et al. 2008]	Human-Human, Human - Group	A,V	1c, 1m

a. Modalities: A(udio), V(ideo), P(hysiological), D(epth), G(aze), T(hermal), O(ther).

b. Sensors: c(amera), m(icrophone), o(ther).

not as well polished as the real product, may still have some bugs, and may need expert knowledge to set up and operate. They do feature full functionality, however, or something very close to full functionality. Partially functional prototypes are interfaces that either need manual interference by, e.g., a wizard to allow it to be studied, or the study focuses on the non-functional aspects of it (e.g., the design). Studies on partially functional prototypes will usually focus on particular aspects of the interface rather than the interface in its entirety. Finally, mockup interfaces do not have any implemented functionality and are often simply made of cardboard or even just drawn examples of what an interface may look like.

For the recording setting, we will distinguish between *lab*, *office*, *home*, and *outdoor*, terms which are self-explanatory. Below we describe each listed database in some detail.

The AIBO database [Batliner et al. 2004] recorded children in English and German while interacting with small AIBO robots, which were operated by a researcher.

**Table 10.2** Categorization of Databases according to the criteria set out

Database	Interface Realism	Recording Setting	Subject Awareness	Purpose
AIBO robot	Partially functional	Lab	Open	Other-focus
AMI	Fully functional	Office	Open	Interface focus
BP4D	Partially functional	Lab	Open	Other-focus
Canal 9	Real product	Office	Hidden	Other-focus
CMU MMAC	Fully functional	Home	Open	Other-focus
DEAP	Partially functional	Lab	Open	Other-focus
EmoTV Corpus	Partially functional	Outdoor	Hidden	Other-focus
GEMEP	Partially functional	Office	Open	Other-focus
HUMAINE	Partially functional	Mixed	Mixed	Other-focus
MAHNOB-HCI	Partially functional	Lab	Open	Other-focus
MAHNOB-Laughter	Mockup	Lab	Open	Other-focus
Math Data Corpus	Fully functional	Lab	Open	Interface focus
MHAD	Partially functional	Lab	Open	Other-focus
MHI-Mimicry	Partially functional	Lab	Open	Other-focus
NoXi	Fully functional	Lab	Open	Interface focus
RECOLA	Partially functional	Lab	Partially hidden	Other-focus
SEMAINE	Partially functional	Office/Lab	Open	Interface focus
SMARTKOM	Fully functional	Outdoor	Hidden	Interface-focus
TUM Kitchen	Fully functional	Home	Open	Other-focus
VAM	Partially functional	Outdoor	Hidden	Other-focus

The children thought that the experiment was about their ability to command the robots to perform certain tasks, and it is unlikely that they would be aware that the true study was to collect the audio-visual language they used to interact with the robots. While recorded in English and German, and while both audio and video data was collected, unfortunately only the German audio data is available. The data is orthographically transcribed, meaning that the sound of the voice is recorded in the target language's script. Both verbal and non-verbal events other than spoken words were annotated as well (e.g., sighs, laughter, breathing in microphone, etc.).

The AMI corpus [Carletta et al. 2005] is a large corpus of group interactions, containing over 100 h of recorded meetings in a standard meeting room. Audio-visual sensors range from close-talk microphones and personal cameras to wide-angle overview cameras. In addition the rooms are instrumented to allow capture of what is presented during meetings, both presentation slides and what is written on an electronic whiteboard. Finally, individual note-taking was recorded using digital pens. Due to the high quality of the data collection process and its professional

dissemination, the AMI corpus has been used extensively. As a result, it is probably one of the most densely annotated databases in existence.

The BP4D database [Zhang et al. 2016] is a truly multimodal collection of emotional facial expressions. Sensor streams focused on the face include high-resolution 3D dynamic imaging, high-resolution 2D video, thermal video, and contact physiological sensors that included electrical conductivity of the skin, respiration, blood pressure, and heart rate. The data was taken under extremely controlled laboratory conditions, with the ultimate aim to record natural and spontaneous expressions of emotion. To do so, a protocol with four approaches was used: a social interview, watching film clips, physical experience, and participating in a set of controlled activities. While the data is of very high quality, the strict lab conditions and expensive sensors used limit its use to primarily benchmarking and hypothesis evaluation.

The Canal-9 corpus of political debates [Vinciarelli et al. 2009] is a densely annotated collection of Swiss televised political debates in French. Typically, 2–4 people participate in each debate, and the debates are moderated by a host. While the studio used multiple cameras and microphones to record the participants in the debate as well as the host, only the final, edited video and audio streams are available. This means that while originally close-ups of multiple people would have been recorded, at any one time only the data from a single camera will be available to the researcher. This in turn means that during a close-up of one person, you cannot determine the reaction of the other debaters.

The CMU MMAC database [De la Torre et al. 2008] records a number of people preparing meals in an almost fully operational kitchen. Multiple sensors were used to record the real cooking sessions, including audio, video, wearable motion sensors, and a Vicon motion capture system. Multiple video streams were captured, including first-person view. Synchronization between sensors is done at individual system clock level, synchronizing the clocks of the various computers used to store the data. As such, one can expect synchronization errors to be in the order of dozens of milliseconds [Lichtenauer et al. 2011]. The database comes with timed textual descriptions of the actions recorded, e.g., “6845 6927 put-pan-into-cupboard\_bottom\_right.”

The DEAP database [Koelstra et al. 2012] was created to record the physiological signals of participants when watching a wide range of one-minute long music videos. The database consists of 32 subjects, who rated each video in terms of arousal, valence, like/dislike, dominance, and familiarity. For each participant, the 32-channel 512 Hz EEG signal was recorded, along with peripheral physiological signals, and (for 22 participants) the face video. The stimuli selection (i.e., the

**Table 10.3** Size overview of existing multimodal multisensor interface databases

Database	Participants	Duration	Demographics	Annotations
AIBO robot	51	9.5 hr	59% female, age 10-13	Orthographical transcription, 11 discrete affective states
AMI	—	100 hr	—	Orthographical transcription, emotion, discourse properties, gestures
BP4D	140	7.8 hr	59% female, 4 ethnic groups	Discrete emotions, FACS
Canal 9	190	43 hr	13% female	Turn taking, (dis)agreement, social role
CMU MMAC	46	57 hr	Unknown	Action descriptions
DEAP	32	—	50% female, age 19-37	Arousal, valence, like/dislike, dominance, and familiarity
EmoTV Corpus	48	12 min	Unknown	14 labels: primary and secondary emotions
GEMEP	10	60 hr	50% female	12 emotions ranked by arousal/valence + 6 additional states
HUMAINE	48	1 hr	Unknown (different languages)	Perceived emotion, signs that convey emotion, contextual factors
MAHNOB-HCI	27	—	—	Arousal, valence, dominance, predictability, emotional keywords, agreement/disagreement
MAHNOB-Laughter	22	—	45% female, avg age 24 (std. 4)	smiles, laughs, speech, other vocalizations
Math Data Corpus	18	29 hr 50%	Transcriptions, problem difficulty & performance, written symbol meaning	
MHAD	12	82 min	42% female, 23-30 yr + 1 elderly	3D positions and performed actions
MHI-Mimicry	40	11 hr	30% female	Dialogue act, turn-taking, affect, head and hand gestures, body movements, and facial expressions
NoXi	89	22.5 hr	Recordings in English, French, German, Spanish, Indonesian, Italian and Arabic	Not given yet?

**Table 10.3** (*continued*)

	Database	Participants	Duration	Demographics	Annotations
RECOLA	46	4 hr	27(out of 46) female, 22 yr avg: (3 yr std), French speakers, although with different native languages		
SEMAINE	150	80 hr	62% female, age 22–60 (32.8 avg), most Caucasian		Valence, activation, power, anticipation, intensity/7 emotions/Epistemic states, interaction process Analysis, Validity/Laughs, Transcripts, Nods and Shakes, FACS (partially)
SMARTKOM	45	unknown	55% female, 80% German, Age 16–45 + 6 subjects > 45		Transliteration, turn segmentation, 2D gestures, user states
TUM Kitchen	4	Unknown	Unknown		Motion segmentation Left hand, right hand, trunk, Set of motion or actions
VAM	104 (video), 47 (audio)	12 hr	Unknown		Valence, activation, and dominance

**Table 10.4** Resource location of the database, user licence information, and approximate total database size

Database	Resource Location	User License	Disk Space
AIBO robot	<a href="https://www5.cs.fau.de/de/mitarbeiter/streidl-stefan/fau-aibo-emotion-corpus/">https://www5.cs.fau.de/de/mitarbeiter/streidl-stefan/fau-aibo-emotion-corpus/</a>	T.B.C.	T.B.C.
AMI	<a href="http://groups.inf.ed.ac.uk/ami/download/">http://groups.inf.ed.ac.uk/ami/download/</a>	CC-BY 4.0	96 GB
BP4D	<a href="http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html">http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html</a>	T.B.C.	10 TB
Canal 9	<a href="http://canal9-db.sspnet.eu/">http://canal9-db.sspnet.eu/</a>	Academic	—
CMU MMAC	<a href="http://kitchen.eecs.qmul.ac.uk/mmv/datasets/deap/">http://kitchen.eecs.qmul.ac.uk/mmv/datasets/deap/</a>	Academic	—
DEAP	<a href="http://www.eecs.qmul.ac.uk/mmv/datasets/deap/">http://www.eecs.qmul.ac.uk/mmv/datasets/deap/</a>	— <sup>a</sup>	—
EmoTV Corpus	—	Academic	—
GEMEP	<a href="http://www.affective-sciences.org/en/gemep/">http://www.affective-sciences.org/en/gemep/</a>	Academic	—
HUMAINE	<a href="http://humaine-db.ssprnet.eu/">http://humaine-db.ssprnet.eu/</a> or <a href="http://emotion-research.net/download/pilot-db/">http://emotion-research.net/download/pilot-db/</a>	Academic	150 MB
MAHNOB-HCI	<a href="http://mahnob-db.eu/hci-tagging/">http://mahnob-db.eu/hci-tagging/</a>	Academic	—
MAHNOB-Laughter	<a href="http://mahnob-db.eu/laughter/">http://mahnob-db.eu/laughter/</a>	Academic	—
Math Data Corpus	<a href="http://mla.ucsd.edu/agreement.pdf">http://mla.ucsd.edu/agreement.pdf</a>	Academic	SIZE
MHAD	<a href="http://tele-immersion.citris-uc.org/berkeley_mhad">http://tele-immersion.citris-uc.org/berkeley_mhad</a>	Open-license and copyright?	822 GB
MHI-Mimicry	<a href="http://mahnob-db.eu/mimicry/">http://mahnob-db.eu/mimicry/</a>	Academic	—
NoXi	<a href="https://noxi.aria-agent.eu/">https://noxi.aria-agent.eu/</a>	Academic	0.5 TB
RECOLA	> <a href="https://diuf.uniffr.ch/diva/recola/news.html">https://diuf.uniffr.ch/diva/recola/news.html</a>	Academic/ Commercial	4 TB
SEMAINE	<a href="http://semaine-db.eu/">http://semaine-db.eu/</a>	Academic	1 TB
SMARTKOM	<a href="http://www.smartkom.org/start_en.html">http://www.smartkom.org/start_en.html</a>	Public?	400 GB
TUM Kitchen	<a href="http://ias.in.tum.de/software/kitchen-activity-data">http://ias.in.tum.de/software/kitchen-activity-data</a>	Academic	360 MB
VAM	<a href="http://emotion-research.net/download/vam">http://emotion-research.net/download/vam</a>	Academic	20 GB

a. Copyright restrictions prevent release.

videos from which certain emotions are meant to be provoked in the users) was done in a semi-automated manner: 60 videos were manually selected, and 60 were chosen using the Last.fm music enthusiast website. The 120 videos were finally reduced to 40 based on the maximum emotional content. Participants then watched the videos and performed a self-assessment of their levels of arousal, valence, liking and dominance. The physiological signals recorded were the EEG, using 32 active AgCl electrodes, along with 13 peripheral physiological signals, placed in the facial skin, hands, neck, and chest.

The EmoTV corpus [Abrilian et al. 2005] is composed of 51 recordings from TV interviews, each lasting between 4 and 43 s, corresponding to 48 subjects. The kind of interviews ranges from environmentally controlled, i.e. “studio” recordings, from recordings “in-the-wild,” i.e., in the street. The annotations were made by two coders, using the Anvil tool, at different levels: (1) audio without video, (2) video without audio, and (3) video with audio. This way, annotations are focused on how people perceive emotional expressions in audio channels (vocal cues and linguistic issues), and in visual channels (facial expressions, body movements, etc.). The annotations include the primary emotions (anger, disgust, fear, joy, neutral, sad, and surprise), and the secondary emotions (despair, doubt, exaltation, irritation, pain, serenity, and worry), thus making 14 labels per clip.

The GENEva Multimodal Emotion Portrayals (GEMEP, [Bänziger and Scherer 2010]) includes a collection of 10 actors portraying 18 affective states, with different verbal contents and modes of expression. The data includes digital high-quality audiovisual recordings, and includes different modalities, using both pseudo linguistic utterances and affect burst. The data also contains stimuli with systematically varied intensity levels, as well as instances of marked expressions. The data was labeled separately in three groups: audio only, video only, and video and audio together. Labelers were recruited from different university departments, to annotate each of the videos and modalities. Each annotation task was performed by several people per media clip, none of whom were an expert.

The Humaine dataset [Douglas-Cowie et al. 2007] includes a mixed collection of recordings mostly drawn from existing databases (some of them included in this Chapter). The database has been collected as part of the HUMAINE project <http://emotion-research.net/download/pilot-db/>, and labeled accordingly. The dataset consists of 48 videos fully annotated using the ANVIL Video Annotation Research Tool [Kipp et al. 2008]. The process of collecting and annotating the data was part of a major work package within HUMAINE project, done in several stages. The first stage consisted on collecting the data, which was categorized as either

“naturalistic” or “induced.” The naturalistic data comes from different existing datasets, the Belfast Naturalistic [Douglas-Cowie et al. 2003], the EmoTV [Abrilian et al. 2005], and the Castaway Reality TV Database. The induced data comes from either self-produced data and databases created as whole packages. The induced data includes Human-Computer SAL data, an Activity dataset of people performing specific actions, the Belfast Driving simulator data [McMahon et al. 2003], in which participants used a driving simulator under induced emotions, the EmoTABOO, where participants were performing games in pairs, the Green Persuasive Dataset, where participants were trying to convince partners about multiple emotional overtones, the DRIVAWORK, where a simulated driving task was used to collect data of participants relaxing, driving normally, or driving while performing a mental task, and the GEMEP dataset, described above.

Although each of the selected sets are databases on themselves, only a total of 48 clips were used in the HUMAINE project under a unified labelling framework. The labels are publicly available to download from the project’s website, although the raw data requires accessing each of the subsets separately. The selected videos were labelled by means of emotions at both global and per-frame levels. The labels also include speech and language descriptors, gesture descriptors, face descriptors, and physiological descriptors.

The MAHNOB-HCI database [Soleymani et al. 2012] has five modalities precisely synchronized, namely, eye gaze data, video, audio, and peripheral and central nervous system physiological signals. Twenty-seven subjects participated in a two-part experiment. During the first part, participants watched fragments of movies, and each were annotating their own emotive state after each fragment, on a scale of valence and arousal. This part was referred to as the “explicit tagging.” During the second part, participants were shown a set of images and videos, along with a tag describing the video, which was either correctly or incorrectly labeling the situation. Users were asked about whether they agreed or not with the tag, and video, audio, gaze, and physiological data was collected in the meantime, with a very precise synchronization between sensors.

The MAHNOB-Laughter database [Petridis et al. 2013] is a database collecting induced smiles of people watching amusing video clips while being recorded by a thermal and RGB camera. In addition, participants were asked to smile and laugh on command, and to speak freely in both their mother tongue and English. The database contains 22 participants (12 male). In total, 180 sessions are available with a total duration of 3 h 49 min. There are annotations for 563 instances of laughter, 849 speech utterances, 51 instances of acted laughter,  $\approx$  50 instances of posed smiles, and 167 other vocalizations.

The Math Data Corpus [Oviatt et al. 2013] was published to support the first Multimodal Learning Analytics grand challenge, in which participants had to predict task performance of students collaborating in solving geometry maths exercises. The students worked in gender and geometry skill-matched groups of three, and could use a set of digital pens and paper to work out the maths problems in a collaborative manner. The students were recorded using five cameras and four microphones, capturing both close-up views and audio as well as overview recordings. In total, 18 students were recorded for a grand total of 29 student-hours, in 12 problem-solving sessions. Transcripts of speech are made available, together with information about the difficulty of the math problems and performance of the groups in solving them. Annotations are also available on the writing of the students, primarily describing the different symbols and graphs created by the student.

The MHI-Mimicry database [Sun et al. 2011] was created to address interactions between humans in detail, with special focus on mimicry and conflict in human-human interaction scenarios. For this purpose, the database was created with the aim of providing a collection of recordings of pairs of people who did not know each other beforehand. One interlocutor was a confederate who was instructed to create conflict. The recordings were made under controlled laboratory conditions using 15 cameras and 3 microphones, to obtain the most favorable conditions possible for analysis of the observed behavior. All sensory data was synchronised with extreme accuracy (less than 10 ns) using hardware triggering.

The Berkeley Multimodal Human Action Database (MHAD) [Ofli et al. 2013] contains a set of 12 subjects (5 female) performing 11 actions (each 5 times), while being recorded by a wide range of devices. The data contains about 660 action sequences corresponding to about 82 min of total recording time. The 11 actions include jumping, jumping jacks, bending, punching, waving two hands, waving one hand, clapping, throwing, sit down/stand up, sit down, and stand up. The actions were recorded using five different systems: an optical motion capture system, four multi-view stereo vision camera arrays, two Microsoft Kinect cameras, six wireless accelerometers, and four microphones. The ground-truth data, i.e. the actual positions, are given by a motion capture system, which captured the 3D position of active LED markers, and a set of eight motion capture cameras arranged in a circular configuration. These cameras were also used to synchronize the captures given by the other sensors. The data is publicly available under a copyright license.

The NoXi dataset ([Cafaro et al. 2017], <https://noxi.aria-agent.eu/>) is a database collected under the EU Project ARIA-VALUSPA, meant to collect natural interactions

between human dyads in an expert-novice knowledge sharing context. Basically, the database is collected while two people interact to each other, where one acts as an expert on a topic of their choice and the other as an interested novice. Experts declared their areas of expertise and novices signed up to hear about what interested them, which ensured they were self-motivated to learn from the expert. The database also contains a collection of controlled and induced interruptions of the interaction, between both the expert and the novice, as an experimental condition. The dataset is recorded in three languages (EN, FR, GE), and includes sensor data from two Kinect RGB cameras, depth data, skeleton data, and audio. In addition to the Kinect microphones, two close-talk microphones were used to collect the audio. The length of the data exceeds 22 h, corresponding to 89 people recorded during 83 interactions.

The RECOLA dataset [Ringeval et al. 2013] presents a remote collaborative scenario that was collected from 46 participants, working in dyadic work teams, which were told they were taking part in a study of people communicating through computer support. After a short introduction, facilitators induced the mood of one of the partners, while keeping the other's neutral. Then participants engaged in a remote discussion according to the survival task paradigm. With the aim to model emotion perception, each participant was asked to continuously label their teammate's emotion. During the experiments, a wide range of measures were taken, including audio, video, and physiological data (EEG and electrodermal activity, EDG). The data was further synchronized through inter-correlation maximization.

The SEMAINE dataset [McKeown et al. 2012] adapted the SAL scenario to incorporate high-quality recordings under three different approaches: a solid SAL, where human operators were playing the agent characters; semi-automatic SAL, where human were selecting sentences to be displayed by a machine; and automatic SAL, where an automated system selects sentences and non-verbal signals. All interactions were recorded with five cameras, and four microphones, and contains multiple annotations, including valence and activation, basic emotions (fear, anger, happiness, etc.), epistemic states (certain/not certain, agreeing/not agreeing, interested/not interested, at ease/not at ease, thoughtful/not thoughtful, concentrating/not concentrating), interaction process analysis (solidarity/antagonism, tension, suggesting/asking for suggestion, giving opinion/asking for opinion, giving information/asking for information), and validity (when the user is not communicating feelings in a straightforward way). Also, transcripts are available, along with annotations of laughs, nods and shakes, and FACS [Ekman et al. 2002]. The

database includes 150 subjects in a total of 959 conversations with SAL characters, each lasting approximately 5 min.

The SMARTKOM corpus [[Wahlster 2006, Schiel et al. 2002](#)] was collected as part of the SMARTKOM project ([www.smartkom.org](http://www.smartkom.org)), which ultimately presented an intelligent computer-user interface allowing natural interaction for the users, under three different scenarios: Public, Home, and Mobile. The system recognizes natural speech as well as gestures above a flat interaction area. Additionally, facial expression is analysed. Subjects were recorded in sessions of 4.5 min each, while interacting with a simulated, but fully functional, version of the system. During sessions, different signals were captured: audio (3 m), video (2 c + 1 infrared), graphical output (for labeling purposes) and gesture coordinates captured by a SIVIT system and the graphical tablet. The sessions were collected in a WOZ experiment in which subjects had to solve certain tasks with the help of the system. The annotations include the transliteration of audio channels, the turn segmentation (between user and machine), the segmentation and labeling of 2D gestures, as well as the user state both by means of facial expressions and speech, and the labeling of prosodic features to recognise emotions. The data includes 45 users, of which 25 were female, 36 German, and whose ages ranged from 6 to 45 years.

The TUM Kitchen dataset [[Tenorth et al. 2009](#)] was recorded in a sensor-equipped intelligent kitchen environment purpose-built to conduct research. The data contains observations of several subjects setting a table in different ways, with each recorded sequence lasting between 1 and 2 min. Sessions for four subjects are publicly available to download from the project site. Actions were recorded using four static overhead cameras. Also, motion capture data was extracted from the videos, using a markerless full-body motion system. The data contains the 6DOF pose and the joint angles. Furthermore, the data is accompanied with RFID tag readings from fixed readers, and data from magnetic reed sensors, detecting when a door or drawer is opened. The data has been manually labeled with a subset of 10 possible actions, such as *Reaching* an object, or *Opening* or *Closing* a door. Each recording is fully segmented into these different actions. Finally, the left hand, the right hand, and the trunk of each person were labeled separately.

The VAM corpus [[Grimm et al. 2008](#)] is made of 12 h of recordings taken from the German TV show “Vera am Mittag” (Vera at Noon). The recordings were further segmented into broadcasts, dialogue acts and utterances, respectively. It contains spontaneous emotional speech recorded from real discussions between guests. The data has been annotated for valence, activation and dominance, using a large number of human evaluators.

## 10.4

### Creating Your Own Database

If after careful deliberation of your data needs and a review of existing datasets you have come to the conclusion that you need to record your own dataset, you will have to prepare well. It is all too easy to invest a lot of time and other resources in your data collection and end up with a large amount of data with little value, resulting in a large waste of time for you, your colleagues, and also the participants that took part in your recordings. This section describes database recording best practice, based on our experience with creating the MMI, MHI-Mimicry, SEMAINE, NoXi, as well as a number of not (yet) publicly available databases. It will cover how to design your database recording (Section 10.4.1), how to record ethically (Section 10.4.2), piloting your design (section 10.4.3), how to store and distribute your data (Section 10.4.4), and finally provide some miscellaneous tips (Section 10.4.5).

#### 10.4.1 Experimental Design

Perhaps the most important aspect of recording a significant dataset is to carefully plan and design every aspect of the data collection process. The best way to keep a record of this is to create a database design document that describes the experimental design, including the recording protocol, recruitment strategy, ethical approval considerations, data storage, and dissemination plans.

*Overview.* Based on the sections above you should by now have the main aims and goals of your data collection—they should go straight and explicitly in the database design document. Try to answer questions such as: What is its purpose? Is it just to evaluate or to build systems? What is its intended audience? Can you generalize the target domain to reach a larger audience? You should certainly establish what type of database you need: development, benchmark, or hypothesis evaluation. It is probably also worth trying to assess what values each of the properties listed in Tables 10.1, 10.2 and 10.3 would take for your database. Try to translate the aims and goals into experimental variables and list these in your document.

In the overview you should describe the experimental scenario clearly; for example, a dyadic mediated interaction of interlocutors engaging in an iterated prisoners' dilemma game. This should give you a good feeling for the space and time requirements, and would be a good moment to decide what room(s) or other spaces you are going to use.

*Technical design.* If so far you haven't decided what sensors and modalities you will be collecting data from, now is the time. You should base the choice of sensors on your aims and goals, but see also the tip at the end of this chapter on over-dimensionalizing—you may want to be conservative and throw in a few

extra sensors, just because you can. With the scenario described and a choice of sensors made, you should now include diagrams of spatial layout of the recording setup. Include other diagrams such as how multiple sensors are to be synchronized and how the data flows from the sensors to the (temporary) storage units, clearly indicating transmission speeds and storage volumes, as suitable for your scenario. You will probably run into your first logistic issues now, perhaps because you find that your capture units (e.g., PCs or Raspberry Pis) are not even theoretically capable of storing the data fast enough (dropping frames) or have insufficient storage volume capacity, or perhaps you find your wires aren't long enough to connect sensor X to unit Y, or because you need 5 USB ports but your recording PC has only 4 ports available. The problems can be many and the number of technical challenges to overcome is sometimes maddening.

Include a description how you are going to transport and store the data, dividing it in logical units depending on the number of participants, sensors, and perhaps separate sections of the experiment. For example, in the SEMAINE database, there were two participants: an operator and a user. The operator played four different characters, in turn, each for about five minutes. We therefore defined the whole interaction between operator and user as one recording, and called each interaction between a user and one of the four characters a session, i.e., each recording consisted of four sessions. Resist the temptation to make these logical units as small as possible—this will only complicate matters later.

*Logistics.* The logistics component of your design document has more to do with people than with the technology and equipment you will use. To start the logistics section, the first thing you should do is to estimate the number of recordings and/or participants you will need or want to record. To determine this, you should at least consider your experimental variables, deriving a minimal number that will allow you to get statistically significant results.

You should create a step-by-step protocol that describes minute-by-minute who is where doing what, starting from the arrival of a participant and the preparation of the recording setup to the moment the participant leaves the building. Try to imagine everything that could go wrong. Doing this for all people involved for every moment in time (perhaps aided by a diagram) will allow you to check if you have created bottlenecks in your logistics flow. For example, if participants need to be let in the building using a key-card by someone, is there a collaborator free who can do this while you record someone else or do you need to do this in between recordings?

Based on this step-by-step protocol you can now gauge how long each recording session takes, and thus how many people/sessions you can record each day.

Together with the required number of sessions/participants this will tell you for how many days/weeks you need to record. Often this is multiple weeks, so make sure you can actually use the space and equipment for that duration! It is not uncommon at this point to realize that the space you allocated for your experiment is insufficient, for example, you may realise you need a waiting room for your participants, or a room where they can be debriefed and fill in questionnaires. This will require you to revisit some of the earlier assumptions of your design. Such iterative improvement of your design is normal.

It is the nature of multimodal, multisensor recordings to generate large volumes of data. It is important that you estimate how much data you generate per minute of recordings, and compare this to the duration of each recording session, the number of sessions planned per day, and the available storage on your recording units. Often you will make use of optimized streaming buses, and/or cannot risk overloading the CPU for fear of dropping data. It is therefore often not possible to transfer the data off the recording unit onto a storage server while recording, and thus you should plan time to transfer data. Luckily this process can usually be automated to run overnight, but do check this as it is another potential limitation on the number of sessions you can record in a day! Clearly this is a good moment to organize your data backup plan.

*Recruitment.* You should also describe your recruitment plan in the database design document. This includes who you aim to recruit (see the note on demographics below), how you intend to reach them, who will be the point of contact for the participants, what information you will store about them, and how you will inform them on the results of your research. This would also be a good place to describe how you will create a well-protected link between the anonymized data and the individuals the data belongs to. This could be done in a single, encrypted text file that lists participant names and contact details associated with an anonymous participant number.

While it may seem a major effort to produce such a detailed database design document, it is well worth it. It is natural and to be expected that you will find problems during the creation of the document, which is one of the main reasons for creating it in the first place. This means you are able to address potentially crippling issues even before recruitment begins. It will also serve as a reference document, to be used when drafting the ethics documents (see Section 10.4.2), to inform collaborators drafted in to help only after the design is finished, and of course it can form the basis for a potential future academic paper describing the dataset. You will not regret investing the time to create this document.

### **10.4.2 Ethics**

An increasing number of countries require due diligence in terms of acquiring ethical approval for the collection of data to go ahead. Usually, an ethics committee will look into issues such as privacy (are people identifiable, what steps have you taken to avoid this), unjust pressure on vulnerable people to participate (e.g., by large financial incentives, but this could also concern supervisors putting pressure on their students to participate), and safety of the recordings. It will almost always feature the collection of informed consent from participants, which means you will need to prepare a document that describes the experiment's goals and what is expected of a participant in language that is appropriate to your target demographics, and a separate consent form that participants or their legal representatives should sign. Note that we have found it useful to add separate tick-boxes to allow the data to be shared with other researchers adhering to similar ethical guidelines, and a tick-box to allow imagery/audio to be used in scientific publications and/or presentations (see Section [10.4.4](#) on how to distribute your data).

Even if your country doesn't have an obligation to request ethics approval for data collection, you may want to consider carrying out a process of ethics anyway, as countries with strong ethical oversight are not only required to obtain ethics approval to record data but, but are also frequently prohibited from using data that was collected without proper ethical oversight in place. Countries/regions that have such regulations include the EU, UK, and the U.S. Thus, if you want your data to be used by as many scientists as possible, you should take this into account and clearly report on the ethics procedure carried out.

One interesting core principle of ethics is that you should not bother participants more than is necessary. This includes that you shouldn't invite the same person to do the exact same experiment twice, if doing it once would suffice. Similarly, you shouldn't use two people where one would suffice in general. This principle is a strong ethical argument for sharing databases between researchers—if you don't share your database, you are essentially forcing other researchers who want to replicate or improve on your research to create their own database, and thus recruit new participants.

### **10.4.3 Pilot**

A pilot study is a dress-rehearsal of your experiment. Doing a pilot study is highly recommended as you will almost always identify problems with your recording procedure. Pilot studies address three main goals: (i) they test technical readiness, (ii) they test logical feasibility/efficiency, and (iii) they provide an indication of

experimental signal to noise ratio, that is, will the data you collect provide you the information you need to reach your goals? You should treat a pilot study as a dress rehearsal for a theatrical performance—there is to be no skipping of steps just because the recordings aren't for real. Only then will you be able to find the small issues in your technical or logical design that slipped through the net during planning. Running an initial quick analysis on the collected data to test the experimental signal is also extremely valuable—in doing so you effectively test the whole recording process end-to-end, and you have verified that you haven't accidentally forgotten to save that crucial file with the start times of each sensor, for example.

The data collected during the pilot serve to test the design and implementation, and should be discarded afterward. But given the nature of pilot studies, it is usually fine to use participants that wouldn't normally be eligible to partake in the study, i.e., lab-mates who are overly familiar with the study. Using them instead of eligible participants means you won't reduce your demographic pool of people to recruit from.

#### 10.4.4 Storage and Distribution

Often, the data is recorded in raw format as compression during recording is too resource intensive, hence compression needs to be done in a post-processing step. Be careful to avoid using lossy compression as this is an irreversible step. If you do use compression, make sure you set the quality level generously high. Often, it is useful to create a low-quality version for quick transfer and visual inspection of your data. This low-quality version can be a mash-up of multiple sensor streams. For example, in the SEMAINE database we created a stream that shows the operator on the left and the user on the right. A neat additional trick is that we added audio using the “balance” function of stereo recordings, with the operator audio on the “left” channel and the user audio on the “right” channel. This allows you to easily mute one of the users if you want.

Given the extreme value of the data you recorded, you should in advance plan how to back up the large amounts of data you recorded. Given the large size of this data, it is usually cost-efficient to use your own dedicated solution rather than cloud storage. A simple setup with two machines, each situated in a different building on your campus in case of fire, kept in sync using rsync is a good solution. Note that depending on your ethics situation you may not be allowed to use off-site data storage at all, including cloud-computing, especially if it means storing data in another jurisdiction.

Note that it is easier to keep fairly large chunks of data combined with files that delimit things that happen in them, rather than a multitude of very short recordings. For example, it would not be a good idea to cut the master recordings into one video per sentence spoken by a participant, as this would result in an unwieldy collection of videos. In later analysis it would be very hard to combine them.

Use meaningful filenames, based on the logical units you defined in your experimental design. Making use of a date-timestamp created at the start of the recording in the filename is particularly useful as it gives a very meaningful human-readable identifier. Continuing the SEMAINE example, we created filenames that started with the timestamp, followed by the participant type (operator/user), followed by the sensor name.

Ideally you should distribute your data as widely as possible, in order to attain the highest possible impact in terms of repeatability of your work, thereby helping science further in general, and to minimize the burden on participants worldwide (see Section 10.4.2). Some people distribute their data by mailing external hard disk drives, but the nicest way of sharing your data is through an easily accessible website that allows people to browse, preview, and select subsets of the data (e.g., the SEMAINE database: <http://semaine-db.eu>). If you do share your data with other researchers, you should make this clear in the participant consent forms. You should ask researchers to sign an EULA prohibiting further distribution, asking for proper acknowledgment of the database creators, and include any other usage constraints you feel necessary. It is a good idea to request that such EULAs are signed by permanent members of staff rather than students, as it may be very hard to hold students legally responsible for misuse of the data, especially after they have completed their studies and moved on.

### 10.4.5 Tips

The following miscellaneous tips may come in useful.

*Focus on your goal.* Be careful not to try and achieve too much—you might end up achieving nothing. Given the large amount of time and preparation needed to create a database, it is sometimes tempting to kill two birds with one stone, that is, trying to achieve two separate goals with one database. This is sometimes possible—for example, you can often easily add an extra sensor to create data for that different research student's needs—but it can sometimes go horribly wrong. As a case-in-point, the MHI-Mimicry database was recorded to record instances of both conflict and mimicry. This dual goal led to an experimental scenario that in the end created recordings of neither convincing conflict nor large amounts of mimicry.

*Over-dimensionalize.* You can always reduce the sampling rate, or the resolution of your images, or the number of sensors you use to train a system. However, you can never add more after the fact. Just because your immediate goal is to create a video analysis tool that runs on smartphones with a  $640 \times 480$  pixels 10 frames per second feed doesn't mean you can't record at  $1920 \times 1200$  60 frames per second. But creating large resolution data will allow you to use the data for different applications where you will have access to better cameras.

*Demographics.* Do you really need participants from all demographics? The more variables you add, the harder it becomes to prove something. Generally speaking, for research you can evaluate a system or hypothesis on a single demographic, if it's reasonable to argue that the same will hold for another demographic.

*Synchronization.* People often say that they need their data streams to be "synchronized," meaning that they can tell afterward what section of the data stream of one sensor belongs to the stream of any other sensor. However, perfect synchronization is almost impossible to achieve. You should therefore ask yourself carefully how much uncertainty in time difference between sensors stream data is acceptable. If delays in the order of hundreds of milliseconds are fine, you can use a clapper board or someone clapping their hands, perceivable by all sensors. If a delay of a few dozen milliseconds is acceptable, you can probably suffice with using the system clock of your devices. If you have multiple devices, synchronize their clocks using Network Time Protocol. If you need delays at the millisecond level, you probably need some form of hardware synchronization. This becomes more difficult if you have heterogeneous sensors.

One trick that you could do with audio-visual sensors is to get cameras that have a trigger-operated shutter. This means that the delay depends on the speed of electricity and the length of your cable. Usually you can set one camera as the trigger and the others as slaves. To synchronize your video with your audio data, you can then record the trigger signal using the same mixer that you plug your microphones in, recording the trigger as a separate audio channel [Lichtenauer et al. 2011].

When creating a mediated interaction scenario, where people see each other through screens in different rooms, you can test synchronization by trying to slow-clap at the same time in both rooms, or by counting to 10, alternating odd and even numbers between the two rooms.

*Partner up.* Obtaining a sufficient number of participants can be difficult, as can convincing other people to start using your data. If you partner up with another lab and record your data at both sites, you can potentially double the number of participants you can reach, double the effort in annotation of your data, and it

is likely that you will receive twice as much exposure to your work, including in terms of citations. If the other lab is in another part of the world, you get a second demographic group for free too!

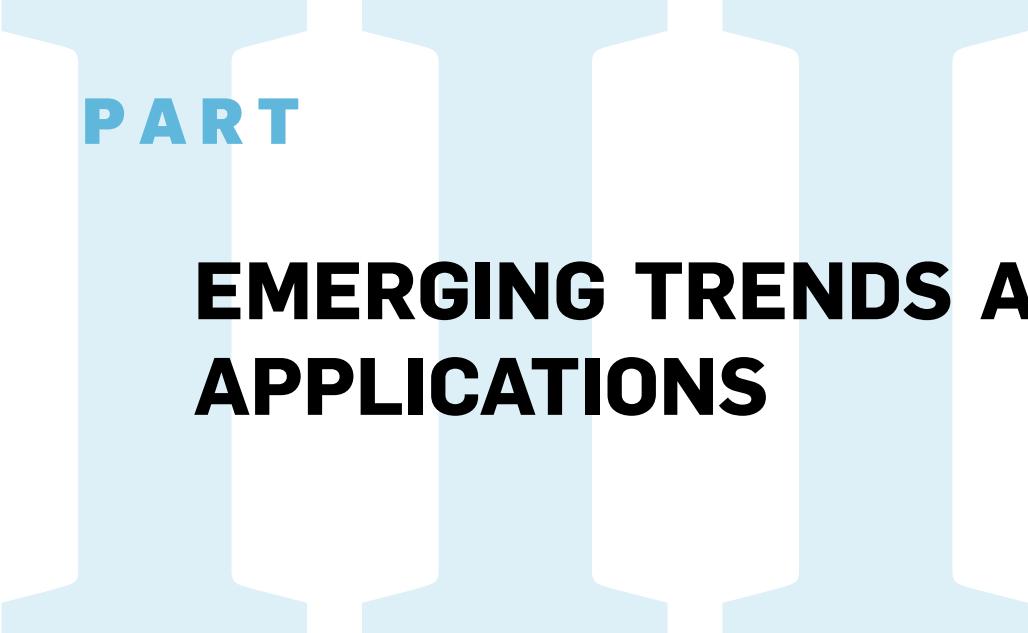
## References

- S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin. 2005. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*. 401, 407, 408
- T. Bänziger and K. R. Scherer. 2010. Introducing the geneva multimodal emotion portrayal (gemep) corpus. In K.R. Scherer, T. Bänziger, and E.B. Roesch, editors, *Blueprint for Affective Computing: A Sourcebook*, pages 271–294. Oxford University Press, Oxford. 401, 407
- A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong. 2004. “You stupid tin box”—children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*. 398, 400, 401
- A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar. 2017. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359. ACM. DOI: [10.1145/3136755.3136780](https://doi.org/10.1145/3136755.3136780). 401, 409
- J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer. DOI: [10.1007/11677482\\_3](https://doi.org/10.1007/11677482_3). 401, 402
- A. Crabtree, A. Chamberlain, M. Davies, K. Glover, S. Reeves, T. Rodden, P. Tolmie, and M. Jones. 2013. Doing innovation in the wild. In *Proc. CHItaly 2013*. DOI: [10.1145/2499149.2499150](https://doi.org/10.1145/2499149.2499150). 394
- F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran. 2008. Guide to the carnegie mellon university multimodal activity (cmu-mmact) database. *Robotics Institute*, page 135, 2008. 401, 403
- E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1–2): 33–60. DOI: [10.1016/S0167-6393\(02\)00070-5](https://doi.org/10.1016/S0167-6393(02)00070-5). 408
- E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and K. Karpouzis. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International Conference on Affective Computing and Intelligent Interaction*. DOI: [10.1007/978-3-540-74889-2\\_43](https://doi.org/10.1007/978-3-540-74889-2_43). 401, 407
- P. Ekman, W. V. Friesen, and J. C. Hager. 2002. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City, UT. 410

- M. Grimm, K. Kroschel, and S. Narayanan. 2008. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE International Conference on Multimedia and Expo*, pages 865–868. DOI: [10.1109/ICME.2008.4607572](https://doi.org/10.1109/ICME.2008.4607572). 401, 411
- M. Kipp et al. 2008. Spatiotemporal coding in anvil. *LREC*. 407
- S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. Deep: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1): 18–31. DOI: [10.1109/T-AFFC.2011.15](https://doi.org/10.1109/T-AFFC.2011.15). 401, 403
- J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic. 2011. Cost-effective solution to synchronised audio-visual data capture using multiple sensors. *Image and Vision Computing*, 29(10): 666–680. DOI: [10.1016/j.imavis.2011.07.004](https://doi.org/10.1016/j.imavis.2011.07.004). 403, 418
- P. Lucey, J. F. Cohn, T. Kanade, J. N. Saragih, and Z. Ambadar. 2010. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 94–101. DOI: [10.1109/CVPRW.2010.5543262](https://doi.org/10.1109/CVPRW.2010.5543262). 397
- G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1): 5–17. DOI: [10.1109/T-AFFC.2011.20](https://doi.org/10.1109/T-AFFC.2011.20). 401, 410
- E. McMahon, R. Cowie, S. Kasderidis, J. Taylor, and S. Kollias. 2003. What chance that a dc could recognise hazardous mental states from sensor outputs? In *Proceedings of the DC Tales Conference*. 408
- T. Mitchell. 1997. *Machine Learning*. McGraw Hill. New York City. 394
- F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *IEEE Workshop on Applications on Computer Vision (WACV)*. DOI: [10.1109/WACV.2013.6474999](https://doi.org/10.1109/WACV.2013.6474999). 401, 409
- S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro. 2000. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human-Computer Interaction*, 15(4): 263–322. DOI: [10.1207/S15327051HCI1504\\_1](https://doi.org/10.1207/S15327051HCI1504_1). 399
- S. L. Oviatt, A. Cohen, and N. Weibel. 2013. Multimodal learning analytics: Description of the math data corpus for icmi data-driven grand challenge workshop. In *Second International Grand Challenge Workshop on Multimodal Learning Analytics*. DOI: [10.1145/2522848.2533790](https://doi.org/10.1145/2522848.2533790). 401, 409
- M. Pantic, N. Sebe, J. F. Cohn, and T. Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676. ACM, 2005. DOI: [10.1145/1101149.1101299](https://doi.org/10.1145/1101149.1101299). 397
- S. Petridis, B. Martinez, and M. Pantic. 2013. The MAHNOB laughter database. *Image and Vision Computing*, 31(2): 186–202. DOI: [10.1016/j.imavis.2012.08.014](https://doi.org/10.1016/j.imavis.2012.08.014). 401, 408

- R Picard. 1997. *Affective Computing*. MIT Press. Cambridge, MA. 397
- R. W. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10): 1175–1191. DOI: [10.1109/34.954607](https://doi.org/10.1109/34.954607). 397
- F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. DOI: [10.1109/FG.2013.6553805](https://doi.org/10.1109/FG.2013.6553805). 401, 410
- Y. Rogers. 2011. Interaction design gone wild: striving for wild theory. *Interactions*, 18(4), July 2011. DOI: [10.1145/1978822.1978834](https://doi.org/10.1145/1978822.1978834). 394
- F. Schiel, S. Steininger, and U. Türk. 2002. The smartkom multimodal corpus at bas. In *LREC*. DOI: [10.1.1.7.9049](https://doi.org/10.1.1.7.9049). 411
- M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. January 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1): 42–55. DOI: [10.1109/T-AFFC.2011.25](https://doi.org/10.1109/T-AFFC.2011.25). 401, 408
- X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, and M. Pantic. 2011. A multimodal database for mimicry analysis. In *International Conference on Affective Computing and Intelligent Interaction*, pages 367–376. Springer, 2011. DOI: [10.1007/978-3-642-24600-5\\_40](https://doi.org/10.1007/978-3-642-24600-5_40). 401, 409
- M. Tenorth, J. Bandouch, and M. Beetz. 2009. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*. DOI: [10.1109/ICCVW.2009.5457583](https://doi.org/10.1109/ICCVW.2009.5457583). 401, 411
- M. F. Valstar and M. Pantic. 2010. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In *Proceedings of the International Conference Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70. 397
- A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. 2009. Canal9: A database of political debates for analysis of social interactions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–4. IEEE. DOI: [10.1109/ACII.2009.5349466](https://doi.org/10.1109/ACII.2009.5349466). 401, 403
- W. Wahlster. 2006. *SmartKom: Foundations of Multimodal Dialogue Systems*, volume 12. Springer. 401, 411
- Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. June 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: [10.1109/CVPR.2016.374](https://doi.org/10.1109/CVPR.2016.374). 401, 403

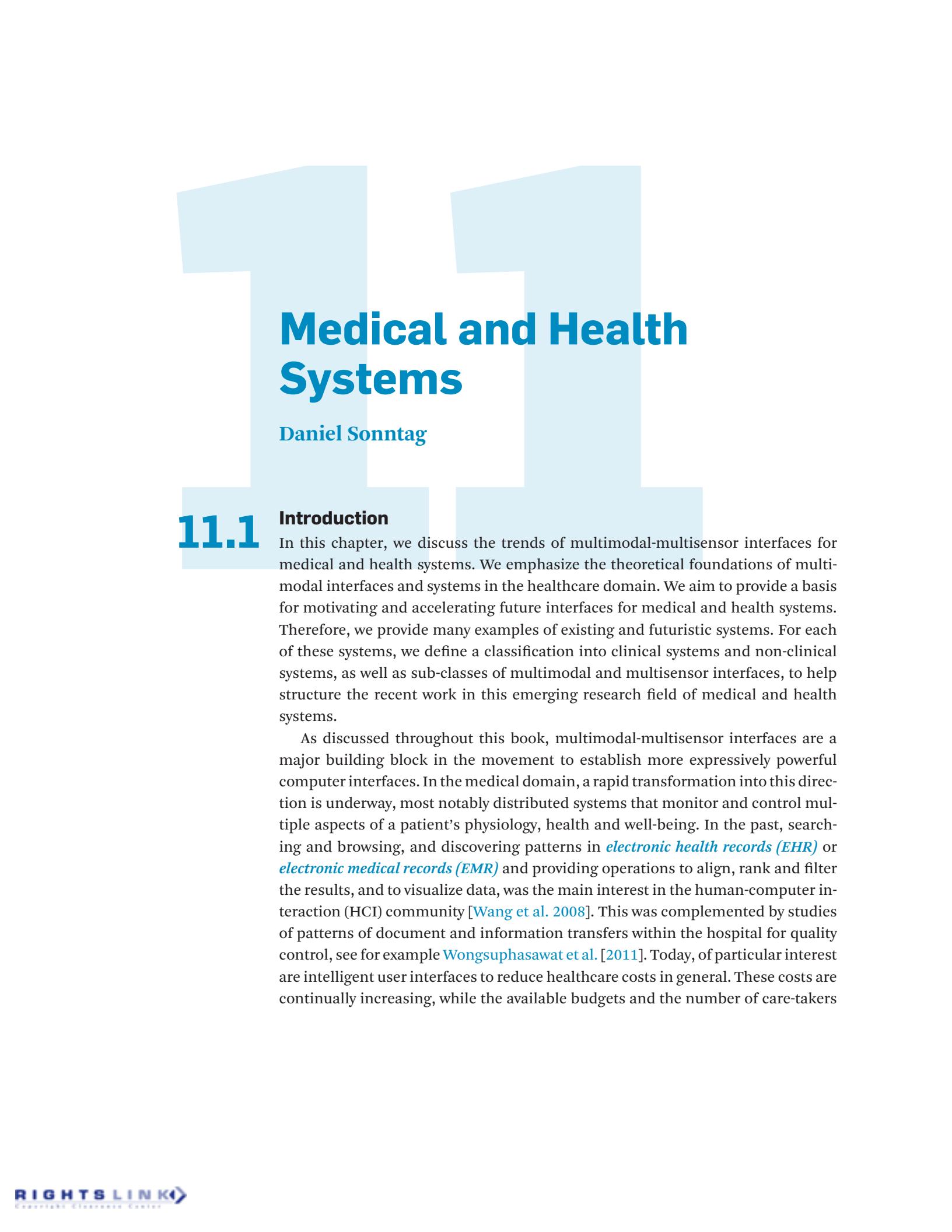




PART

## EMERGING TRENDS AND APPLICATIONS





# Medical and Health Systems

Daniel Sonntag

## 11.1

### Introduction

In this chapter, we discuss the trends of multimodal-multisensor interfaces for medical and health systems. We emphasize the theoretical foundations of multimodal interfaces and systems in the healthcare domain. We aim to provide a basis for motivating and accelerating future interfaces for medical and health systems. Therefore, we provide many examples of existing and futuristic systems. For each of these systems, we define a classification into clinical systems and non-clinical systems, as well as sub-classes of multimodal and multisensor interfaces, to help structure the recent work in this emerging research field of medical and health systems.

As discussed throughout this book, multimodal-multisensor interfaces are a major building block in the movement to establish more expressively powerful computer interfaces. In the medical domain, a rapid transformation into this direction is underway, most notably distributed systems that monitor and control multiple aspects of a patient's physiology, health and well-being. In the past, searching and browsing, and discovering patterns in *electronic health records (EHR)* or *electronic medical records (EMR)* and providing operations to align, rank and filter the results, and to visualize data, was the main interest in the human-computer interaction (HCI) community [Wang et al. 2008]. This was complemented by studies of patterns of document and information transfers within the hospital for quality control, see for example Wongsuphasawat et al. [2011]. Today, of particular interest are intelligent user interfaces to reduce healthcare costs in general. These costs are continually increasing, while the available budgets and the number of care-takers

are shrinking. For example, in developed countries around the world, an ageing population poses challenges to society, but also unique opportunities for HCI and artificial intelligence (AI) methods in health and wellbeing. To use AI methods to a larger extent, we need a systematic collection of patient information in a digital format. Digital records can be shared across different healthcare settings, store data accurately, and capture the state of a patient across time.

We start by classifying multimodal and multisensory interfaces into clinical systems and non-clinical systems. The clinical view includes, most notably: unobtrusive sensing of vital body signals in clinical environments, data mining of contextual clinical data in different modalities (e.g., clinical records and medical images) and semantic annotation of medical texts and images. Applications include text mining in the health and wellbeing domain, big data analysis and clinical data intelligence, personalized schemes for individualized treatment and medication, formalizing clinical guidelines for health and wellbeing, as well as decision support. One of the biggest challenges is to reduce the demand for expensive treatments by detecting small physical and mental health issues early. In addition, avoiding larger health problems by clinical treatment or suitable lifestyle interventions. For example, four specific lifestyle factors (not smoking, maintaining a healthy weight, regular exercise, and following a healthy diet) together are associated with as much as an 80% reduction in the risk of developing the most common and deadly chronic diseases [Ford et al. 2009]. We can act on this challenge by offering AI-based multimodal-multisensor interfaces for integrating self-monitoring sensors (*quantified-self*). This non-clinical view includes, among other things: smart unobtrusive sensing of vital body signals (of, e.g., care home residents), event and task extraction from life logging by, e.g., video capture, data mining of contextual data, smart coaching algorithms for wellbeing, persuasion technologies, and adaptable interfaces that understand the physical and cognitive abilities of the user.

We focus on clinical and non-clinical systems in Sections 11.2 and 11.3, respectively. Three case studies are presented in Section 11.4, followed by future directions of multimodal multisensor combinations and virtual reality in Section 11.5. For a definition of italicized terms in this chapter, see the *Glossary*. For other related terms and concepts, also see the textbook chapter on medical cyber-physical systems [Sonntag 2016], the special issue of the German Journal of Artificial Intelligence [Gelissen and Sonntag 2015], and the overview of the German flagship project on clinical data intelligence [Sonntag et al. 2015]. See this chapter's *Focus Questions* to aid comprehension.

## Glossary

**Application domains** include serious games, conversational agents, or dialogue systems for healthy behavior promotion; intelligent interactive monitoring of patient's environment and needs; intelligent interfaces supporting access to healthcare services; patient-tailored decision support, explanation for informed consent, and retrieval and summarization of on-line healthcare information; risk communication and visualization; tailored access to electronic medical records; tailoring health information for low-literacy, low-numeracy, or under-served audiences; virtual healthcare counselors; and virtual patients for training healthcare professionals.

In addition, we address decision support systems especially for the doctor, which model the diagnostic reasoning and decision-making of medical experts, and systems designed to interact directly with patients as healthcare consumers.

**Biosignals** provide information from a person's biological or physiological structures and their dynamics. Signals measured from the human body typically originate from neural or muscular activity. Neural activity is captured by methods such as EEG, electroencephalogram, a test that detects electrical activity in your brain using small, flat metal electrodes attached to your scalp. Muscular activity is captured by methods such as electromyogram (EMG) electric signals generated by muscles, or electrocardiogram (ECG) electric signals emitted from the human heart. They are the basis for human computing, physiological computing and affective computing. See also [Silva et al. \[2015\]](#). For applications in human computer interaction (HCI) and intelligent user interfaces (IUI), only surface electrodes are used. Signal processing includes, first, time series analysis and, second, the mapping to physical or physiological states [[Schuller 2018](#), [D'Mello et al. 2018](#), [Wagner and André 2018](#), [Martin et al. 2018](#)] toward cognitive states [[Zhou et al. 2018](#), [Cohn et al. 2018](#), [Oviatt et al. 2018a](#)]. Biosignals of future interest include electric conductance, bioimpedance, and bioacoustic signals.

An **electronic medical record** (EMR) is a narrower view of a patient's medical history including laboratory values, for example, while an **electronic health record** (EHR) is a more comprehensive report of the patient's overall health.

**Foundational technologies** are introduced in other chapters in Volumes 1 and 2 of this handbook, namely machine learning [[Panagakis et al. 2018](#), [Baltrušaitis et al. 2018a](#)], deep learning [[Keren et al. 2018](#), [Bengio et al. 2018](#)], and knowledge management [[Alpaydin 2018](#)].

**Medical cyber-physical systems** are real-time, networked medical device systems to improve safety and efficiency in healthcare. The specific advantage of the concepts of cyber-physical systems (CPS) involves the use of both real-time sensor devices (e.g., monitoring devices such as bedside monitors) and real-time actuation devices (such as infusion pumps). In this way, MCPS collect information from the monitoring sensors and actuators by, for example, adjusting the setting of actuation devices, firing an alarm, or providing decision support to caregivers. See [MedicalCPS \[2018\]](#) for intelligent user interface projects that fall into this category.

**Glossary (continued)**

**Medical decision support** systems are guidance services that predict a patient's health status to influence health choices by clinicians. Other functions can be administrative, but we focus on supporting clinical diagnosis and treatment plan processes by for example proposing medical substances with little adverse effects. Future implementations should be integrated into the clinical workflow, provide decision support such as treatment options at the time and location of care as a MCPS rather than prior to or after the patient encounter, and provide recommendations for care, not just assessments.

**mHealth** includes the use of mobile devices in collecting aggregate and patient-level health data.

**Persuasive technologies** focus on the design, development, and evaluation of interactive technologies aimed at changing users' attitudes or behaviors through persuasion, but not through coercion or deception. In general, persuasive technologies are used to change people's behavior. The persuasion approach we support is that choices are not blocked, fenced off, or significantly burdened. The influence on people's behavior in order to make their lives longer, healthier, and better should be subtle. For example, displaying nutrition information at eye-level is a subtle persuasion technology.

**Prevention (primary, secondary, and tertiary)** covers several prevention methods.

Primary prevention aims to prevent disease or injury before it occurs and includes education about health risk factors. Secondary prevention aims to reduce the impact of a disease or injury that has already occurred and addresses an existing disease prior to the appearance of symptoms. Examples include regular exams and screening tests to detect disease in its earliest stages (e.g., mammograms to detect breast cancer) or diet programs to prevent further heart attacks. Tertiary prevention aims to soften the impact of an ongoing illness or injury that has lasting effects. Examples are cardiac or stroke rehabilitation programs and chronic disease management programs (e.g., diabetes). New approaches to improve prevention-related user interaction include *persuasive technologies*.

**Quantified self** is a term used to describe data acquisition on aspects of a person's daily life, e.g., incorporating self-monitoring and self-sensing, which combines wearable *biosignals* sensors and wearable computing.

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications for metadata. It is used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats.

**Telemedicine** subsumes physical and psychological diagnosis and treatments at a distance, including telemonitoring of patient functions.

## 11.2

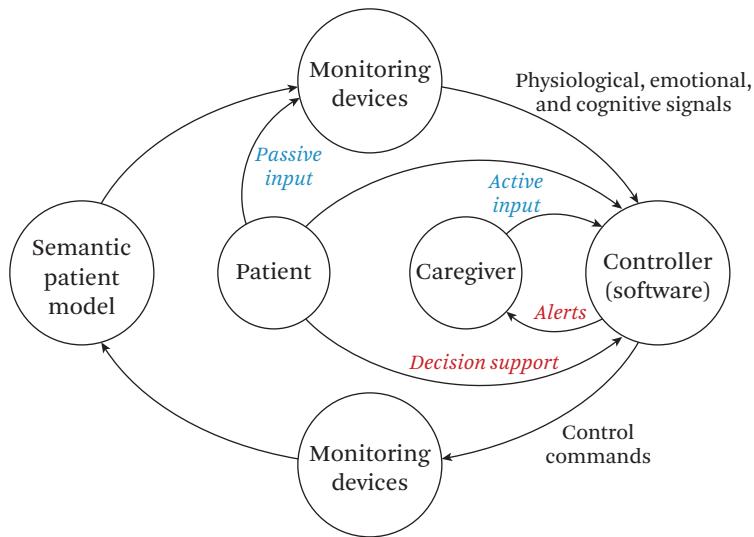
### Clinical Systems

In the future, clinical environments will develop into *medical cyber-physical systems (MCPS)* of their own. This means that patients will get direct treatment according to a direct data acquisition and interpretation workflow. The doctor's decision support will be provided according to the data the MCPS collects from the individual patients. Future MCPS should assist in hospitals, homes, and other settings [Lee et al. 2012, de Man et al. 2013, Carayon 2011, Chen et al. 2014a].

MCPS involve heavier use of sensors and passive user input in terms of biosignals than traditional multimodal interfaces; hence they do not necessarily require explicit, active input from a user (doctor or patient) with an explicit human-computer interface. The development of multimodal-multisensor interfaces that rely heavily on passive user input processing require *foundational technologies* to be effective and reliable without human control via active user input. Active input modes include speech, hand gestures, eye-tracking, digital pens, smartphones, and automatic handwriting recognition. Passive input modes include sensors of the clinical environment, biosignals, or smartphone data. Passive input may involve recognition-based technologies (e.g., gesture) or sensor-based information (e.g., acceleration, pressure). This combination of input sources has not yet been explored in medical environments and is of specific interest because it combines previously unconnected modalities and information sources [Sonntag 2016].

A periodic data collection is important for *primary, secondary, and tertiary prevention* or monitoring chronic symptoms such as asthma or diabetes. So-called cyber-physical system (CPS) controllers can issue alarms for situations that require attention by a doctor in emergency situations or to let clinicians know about the physiological and emotional state of an individual patient. Figure 11.1 shows the resulting conceptual architecture, including monitoring and actuation devices, a semantic patient model [Sonntag and Porta 2014, Sonntag et al. 2014b], and controller software around the patient and the caregiver. MCPS have specific requirements to be met:

1. High confidence medical device software development: This refers to comprehensive verification, validation, and testing as well as robustness and fault-tolerance for clinical systems.
2. Anomaly treatment: The modeling of failures, i.e., anomalies, in using and interacting with medical devices, caregivers, and patient behavior must be accounted for.



**Figure 11.1** Conceptual architecture: networked closed-loop medical cyber-physical system with human-in-the-loop extension. Please note that this extension covers clinical decision support systems for caregivers and patients.

3. Embedded, real-time, networked system development: This includes architecture, platform, middleware, resource management, Quality of Service (QoS) distributed control and functional programming for future *application domains* of health systems.

One of the major application goals is to issue more accurate and targeted alarms, to let the doctors initiate any necessary treatment immediately. The idea is to bring both patients and caregivers into the controlled perception-action loop around the patient; the controller can also start a treatment autonomously. The main concern besides privacy and security [Friedland and Tschantz 2018] is to avoid false alarms. The long-term direction is to build multimodal-multisensor medical systems that simultaneously sense health status (the state-of-the-art in biosignals processing is covered in Volume 2 of this handbook [Oviatt et al. 2018b]), in order to adapt multimodal communication patterns for user-in-the-loop systems and system responses according to users' status. Although having humans-in-the-loop has its advantage, modeling human behaviors is extremely challenging due to the complex physiological, psychological, and behavioral aspect of human beings [Munir et al. 2013, Wood and Stankovic 2008]. We discuss multimodal and multisensor interfaces separately in order to account for the different needs and challenges in the medical domain.

### 11.2.1 Multimodal Interfaces

Multimodal output involves medical system output from two or more modalities, such as a visual display combined with auditory or haptic feedback, which is provided as feedback to the clinician. More modalities allow for more natural communication, which normally employs multiple channels of expression. It is also the case that more modalities constrain the interpretation and, hence, enhance robustness. For the medical image annotation step for example (see Section 11.4.1), predefined speech recognition grammars can be employed. The requirements of medical application domains often include a direct digitalization of multimodal active and passive input data and multimodal feedback in real time. For example, automatic speech recognition and digital pens allow us to transcribe the clinician's spoken and written input. In addition, the requirements often includes knowledge representation and reasoning about medical concepts. Our design principles can be summarized as follows.

1. Representation and Standards: In a complex medical interaction system, a common ground of terms and structures is absolutely necessary. A shared representation and a common knowledge base ease the dataflow within the system, avoiding costly and error-prone transformation processes [Sonntag et al. 2009]. An ontology-based representation combines, for example, formal dialogue and image semantics grammars with an *RDF* repository using the SPARQL query standard *Sesame* [2017]. Linked Open Drug Data (LODD) presents *RDF* based connected medical information graphs that can serve as knowledge repositories for multimodal user queries. The Life Science Interest Group aims to collect, share, and interlink medical data at very detailed levels by harnessing semantic web technologies [Samwald et al. 2011].
2. Encapsulation: Multiple user interfaces can be connected to the multimodal dialogue system. The system also acts as middleware between the multimodal interface and the *RDF* repository [Sonntag et al. 2010a].

One implication of recent research findings is that multimodal interfaces are especially well suited for *mHealth* solutions: multimodal interfaces can directly support the multi-functionality of mobile devices and their applications in different application context with different input and output requirements [Oviatt and Cohen 2000]. In mobile settings, multimodal interfaces will promote the multi-functionality of small devices, in part due to the portability and expressive power

of multiple input modes. These emerging mobile technologies can be used in extended clinical healthcare (e.g., including blood sugar control, heart frequency, movement pattern after hip surgery) and the computer-mediated communication between doctor and patient.

Multimodal interfaces can also be used in semantic search, as the following examples show: personalized search and summarization over multimedia healthcare information [McKeown et al. 2001], and a multimodal dialogue system for medical images [Sonntag and Möller 2010] which integrates a multimodal interface for speech-based annotation of medical images and dialogue-based image retrieval. In addition, Radspeech [Sonntag et al. 2012] is a speech dialogue example which features mutual disambiguation [Oviatt 1999] of recognition errors. Other systems support multimedia queries in medical search in texts and images; see, e.g., Mourão and Martins [2013]. Luz and Kane [2009] investigate the automatic classification of patient case discussions in multidisciplinary medical team meetings recorded in a real-world setting. More sophisticated passive input methods include accelerating meaningful interface analysis through unobtrusive eye tracking of EMRs toward a quantitative and qualitative assessment of EMR interfaces [Rick et al. 2015]. In a further multimodal interface application, an interactive narrative format for clinical guidelines is presented by Cavazza et al. [2015]. In the future, multimodal input and/or output will have another strong application domain: *telemedicine* should facilitate communication between patients and health-care providers and doctors. However, currently, multimodal interface examples are still missing.

### 11.2.2 Multisensor Interfaces

Multisensor interfaces in clinical settings are enabling technologies for future telemedicine, integrating biosignal interpretation (covered in Volume 2 of this handbook), specialized doctors at different locations, and robotic surgery. We will focus on robotic surgery after describing clinical applications with sensors for gesture recognition.

Jacob and Wachs [2014] uses sensor-based contextual cues, i.e., gaze, hand position, and head orientation, to avoid false positive gesture recognitions for navigating MRI images in the operating room. Jacob et al. [2012] developed a prototype of a robotic surgical nurse for handling surgical instruments in the operating room; experimental results show that 95% of the gestures were recognized correctly. Jacob and Wachs [2014] present a sterile system for navigating MRI images in the operating room. The system has been shown to significantly improve task completion performance. Robotic surgery is another multisensor interface example in the op-

erating room: along with [Taurus \[2017\]](#) researchers have been developing a dexterous manipulation interface for telepresent surgical robots for remote surgery. This robot is controlled via direct manipulation through two hand controllers which can give visual and tactile feedback. [Verbsurgical \[2017\]](#) is another robotic surgery platform that integrates sensor technologies with medical imaging, data analysis, and machine learning, in order to introduce more autonomy and control of the robotic arms and its end-effectors. In general, robot CPS systems seek to improve minimally invasive and open surgeries (particularly cardio-thoracic, that have so far not benefited from minimally invasive techniques). The goals are to reduce long hours of operation surgery, increase precision, miniaturize effectors, reduce incision, and decrease blood loss. Additional applications include sensors for example on the effectors for touching soft tissue or bone for computer-assisted surgery or even supervised autonomous robotic soft tissue surgery [[Shademan et al. 2016](#)]. Current test applications are surgeries in urology, gynecology, general surgery, and thoracic surgery. [Turchetti et al. \[2012\]](#) report on cost evaluation studies of robot-assisted operation. Such operations are compared with those performed by a direct manual laparoscopic approach. Evaluations of the endoscopic procedures using this system suggest that it shortens the length-of-stay in the hospital and reduces recovery times. Critics of the systems targeted at HCI aspects focus on the steep learning curve for surgeons who adopt use of the multisensor environment, in the sense that it is difficult and takes much effort to learn the robot-assisted operations. This means a learning curve with a long, fairly flat region, followed by a big, sudden jump. One gains almost no ability until after 50–70 hours of training. Similar surgical interfaces have been developed: MiroSurge [[Tobergte et al. 2011](#)] is a multisensor surgical workstation with several force/torque sensors on haptic input and output devices. It is used in research-based suturing and palpation tasks.

Other multisensor interfaces include robots for basic deliveries or transports of medications, meals, and materials through hospitals; see, for example, [TUG \[2017\]](#). In all these multisensor applications it is to be mentioned that crowdsourcing has many options for quality assurance control of medical procedures done with assisted or autonomous robotics; see, for example, [Chen et al. \[2014b\]](#).

## 11.3

### Non-Clinical Systems

Most non-clinical systems are designed to interact directly with patients. Some non-clinical systems need to understand a patient's intentions, attitude, emotional status, and additional information. In terms of multimodal input processing, facial expression, gaze direction, and emotion as tracked user information are of

particular interest. Multimodal input processing helps to provide a holistic view of the patient. Many new medical education systems use multimodal output (e.g., speech and diagnosis graphs). Medical applications are consistent with the general literature on multimodal processing advantages.

Nutrition, physical exercise, and other non-clinical factors contribute to health and wellbeing. Nowadays, people load their data onto fitness portals such as “MyFitnessPal,” “Fitbit,” or “Garmin Connect.” By integrating those resources, some companies have begun to aggregate the data and show analytics with bars or pie charts, thereby providing new opportunities for user-oriented, personalized non-clinical user interfaces. Web-based solutions can feature an incremental knowledge acquisition process with at least two stages, acquiring fitness data with mobile devices and presenting aggregated data in Web portals. As pointed out in [Friedland and Tschantz \[2018\]](#), advances in multimedia content analysis threatens privacy. People must be made aware that non-clinical data collection and analysis can enable unexpected and invasive inferences about people.

Current research and development efforts of non-clinical systems include wearables (digitally enhanced accessories) that are instantiated in familiar real-world objects like watches, wrist bands, digital pens, and tablets. The design question is to make them more useful, versatile, or attractive for digital input processing. As pointed out in [Oviatt and Cohen \[2015\]](#), one advantage of these interfaces is their transparency to users and ability to leverage existing activity patterns, which minimizes a user’s cognitive load. One long-term interface design direction will be to combine emerging tangible interfaces that support multimodal input with ones that simultaneously sense users’ cognitive load, health status, and similar information in order to adapt system responding to user status. Current non-clinical interfaces for health systems have an emphasis on multisensor interfaces. This extends patient monitoring at hospitals to data collections at home by using portable sensors providing information about a patient’s recovery status.

Because EHRs are used to keep track of medications, allergies, conditions, family history, vitals, and exercise, a speech-based question answering system can take this information as additional input. Vitals tracking using smart devices may offer additional sensors. One example is WatsonPaths [[Lally et al. 2017](#)], a question answering system that can be asked for the most likely diagnosis or most appropriate treatment, over unstructured information where the answer is not contained in documents. Another application example is [GenieMD \[2017\]](#), a telemedicine platform. One may be interested in side effects of a cortisone shot, or recommendations for available treatment options. Users are able to upload medical records such as X-rays and lab results for personalized recommendations

by answering a short questionnaire relating to chief complaints. The system is not yet multimodal at its interface nor does it use sensor input, but future versions may do so.

To provide a better basis for motivating and accelerating future non-clinical systems for medical and health systems, we provide examples of existing and futuristic systems. All of the medical application examples presented in this section have limitations in scope, but collectively they provide converging perspectives on non-clinical systems toward the design of medical multimodal-multisensor interfaces.

### **11.3.1 Multimodal Interfaces**

We summarize the strengths of multimodal interfaces by providing four examples along four dimensions: multimodal data sources, multimodal interaction, method, and goal.

[Sawamoto et al. \[2007\]](#) explore a method for multimodal interaction logs data sources. Gestures and speech are used. Pattern mining methods are applied to medical interviews in order to extract certain doctor-patient interactions.

[Weibel et al. \[2013\]](#) explore how technology can support natural multimodal interfaces for medical information to provide more effective communication in the medical office. The data sources are EMR interaction logs. The system exploits speech interaction together with sensors to track computer-based activity, visual attention, and body movements. The method is pattern mining. The goal is to inform the design of new multimodal healthcare interfaces.

[Bickmore et al. \[2009\]](#) describe an animated, empathic virtual nurse interface for educating and counseling hospital patients in their hospital beds at the time of discharge. It should be emphasized that little research has been done to date on systems to provide information to patients while they are in their hospital beds. Multimodal interaction is provided by a virtual nurse agent (an embodied conversational agent with touchscreen input). The goal is to empower low health literacy hospital patients.

[Lisetti et al. \[2015\]](#) discuss a research project aimed at building socially expressive virtual health agents. Data are collected from interaction logs. They collect data from multiple targets, from obesity to alcohol and drug use, to lack of treatment adherence. Multimodal fusion and fission techniques are primary. The goal is to deliver brief motivational interventions for behavior change in a communication style that individuals and patients not only accept, but also find emotionally supportive and socially appropriate.

### 11.3.2 Multisensor Interfaces

The focus of this subsection is to summarize the strengths of multisensor interfaces to reduce healthcare costs by results from research and application projects, and to describe how they have been applied to date. In this regard, the present list is by no means exhaustive. Robots and their senso-motoric intelligence are not described [Haddadin et al. 2017]. We focus on activity monitoring of humans by non-intrusive sensors, biofeedback and biomarkers, and multisensor interfaces in the context of social and virtual companions. In the future, new multisensor applications, especially for diagnostic reasoning, will arise.

#### 11.3.2.1 Activity Monitoring of Humans by Non-intrusive Sensors

We focus on sensor-based recognition that can be used remotely from cameras or microphones or sensors that are embedded into devices such as smartphones. Chaurasia et al. [2014] discuss a reminder system for carrying out instrumental activities of daily living (iADLs); the system does not focus on interaction with the user, but instead processes data from a network of sensors. An activity probability model is created to prompt the user via a text interface for the next step in the iADL when inactivity is being observed. Similarly, assistants such as Siri and Google Home (TM) can rely on smartphone sensors. Graus et al. [2016] found out that a smartphone's reminder function is an interesting predictor: the creation time is a strong feature in predicting the notification time, and that including the reminder text further improves prediction accuracy with implications for the design of systems aimed at helping people to complete tasks and to plan future activities. Castro et al. [2015] present a research system to predict daily activities from egocentric images using deep learning. They learn a person's behavioral routines and predict daily activities from first-person photos and contextual metadata such as day of the week and time, or contextual information derived from other sensors. Automatic expressive behavior understanding helps to diagnose, monitor, and treat medical conditions that themselves alter a person's social and affective signals. Valstar [2014] describes automatic behavior understanding, based on multiple sensors. Weiss et al. [2016] compare smartwatch and smartphone-based activity recognition, and smartwatches are shown to be capable of identifying specialized hand-based activities which cannot be effectively recognized using a smartphone. Evaluation results show that smartwatch sensors can identify the "drinking" activity with 93.3% accuracy while smartphone sensors achieve an accuracy of only 77.3%. Maurer et al. [2006] report on medical activity recognition and monitoring using multiple sensors on different body positions for patient monitoring. They focus on sensor fusion. Further examples of user state (e.g., alertness, engagement,

physical activity) and trait recognition (e.g., personality, age, gender) where face, fingerprint, and other visual cues are combined, are discussed in [Schuller \[2018\]](#). Multisensory affect detection is described in [D'Mello et al. \[2018\]](#).

### **11.3.2.2 Biofeedback and Biomarkers**

Integrating biomarkers for mental state detection, for example combining emotional and behavioral indicators for autism detection, represent promising research directions. We describe how multisensor interfaces have been applied to date. New sensor networks including Internet-of-things (IoT) devices may produce new multisensor biomarkers for mental disorders. Typical mental disorders can be detected by new multisensor interfaces in the future. [Garbarino et al. \[2014\]](#) describe a wearable wireless multisensor device for real-time computerized biofeedback and sensor data acquisition. [Sriram et al. \[2009\]](#) propose a mobile medical sensor architecture to provide an efficient, accurate, and economic way to monitor patients' health outside the hospital. They provide arguments that patient authentication is a necessary security requirement in remote health monitoring scenarios. [da Silva et al. \[2014\]](#) present Bitalino, a novel development platform for using biosignals. Their low-cost hardware and open-source software toolkit provides streaming functionality of EMG and EDA (electrodermal activity) to build prototypes for future wearable health-tracking devices. [Niemann et al. \[2018b\]](#) use Bitalino to monitor EDA for cognitive assessments for future dementia tests at home. An extended multisensor prototype based on Bitalino will be presented in Case Study 2 (Section 11.4.2). [Picard et al. \[2017\]](#) describe how a commercial wrist sensor reveals sympathetic hyperactivity and hypoventilation in real-time seizure detection by recording wrist motion via three-axis accelerometer and EDA. Extensions to this work for other mental disorders such as dementia will also be presented in Case Study 2.

### **11.3.2.3 Multisensor Interfaces in Social and Virtual Companions**

[Scherer et al. \[2013\]](#) describe audiovisual behavior features for depression assessment during multimodal virtual human interviews. They investigated if audiovisual nonverbal behavior descriptors indicative of depression are observable within semi-structured virtual human interview recordings. Additionally, they assessed the correlation of those behaviors with the assessed depression severity. [Chen et al. \[2016\]](#) conduct a study to motivate patients to exercise. They used multisensor fitness trackers including gyroscope, accelerometer, and EDA. [Mehlmann et al. \[2016\]](#) discuss a research project about modeling grounding for interactive social companions based on sensor input, where common ground is needed for joint action and social speech-based dialogue. Further aspects of common ground are planning to

achieve joint goals and turn-taking. Especially, turn-taking in speech-based dialogue can be informed by sensor input, for example by eye tracking (see Chapter 4).

## 11.4

### Case Studies

We present three medical case studies in the clinical domain. These are chosen because they describe the transition from monomodal to multimodal applications (first study). The use of sensors helps to interpret a clinical patient's physical state, health status, mental status, and engagement in activities relevant for the assessment and monitoring of pathologies such as Alzheimer's (second study). The transition to multimodal-multisensor interfaces is a particularly seminal one in the design of digital tools for behavior characterization in the context of neurodegenerative disorders (third study).

#### 11.4.1 Case Study 1: A Multimodal Dialog System

In this case study, we present a dialogue system for the annotation and retrieval of medical images where different clinicians are involved in the use of multimodal user interfaces.

##### 11.4.1.1 Background

In contemporary, daily hospital work, clinicians can only manually search for "similar" images using outdated desktop search applications. After considering the relevant categories of similarity, they subsequently apply one filter after the other. For instance, a clinician first sets a filter for the imaging modality (e.g., CT angiography), the second filter for the procedure (e.g., coronary angiography), and so on. In addition to the fact that this approach is quite time-consuming, it is neither possible to formulate complex and semantically integrated search queries in a convenient way, nor can a radiologist easily annotate images with new anatomy or disease information. Hence, the need exists for a seamless integration of medical images and different user applications by direct access to image semantics. Semantic image retrieval should provide the basis for the help in decision support and computer-aided diagnosis.

Our solution is a speech-based dialogue system that integrates a multimodal interface for speech-based annotations of medical images and an image annotation tool for manual semantic annotations on a desktop computer [[Sonntag and Möller 2010](#), [Sonntag et al. 2012](#)]. This system implements two of the main applications of medical knowledge acquisition and knowledge integration: first, clinical decision support where some of the clearest opportunities exist to reduce costs by

minimizing the time for finding treatments based on similar patient cases [Bates et al. 2014], and second, treatment optimization for diseases affecting multiple organ systems. In this case study, we also demonstrate the image retrieval (querying) functionality of the multimodal dialogue interface.

#### **11.4.1.2 Problem Description**

Automatic detection of image semantics, i.e., medical annotations of image regions, seems to be feasible, but is too error-prone (at least on the desired annotation level where multiple layers of tissue have to be annotated at different image resolutions or when external expert knowledge is needed). Accordingly, our major challenge is the so-called knowledge acquisition bottleneck. Automatic image recognition cannot easily acquire the necessary medical knowledge about the image contents. As automatic annotation is difficult, we have to address this knowledge acquisition bottleneck problem by concerning ourselves with the question of how to integrate statistical image region annotation (automatic annotation) with manual or semi-automatic annotations.

The requirements discussed with medical experts point to integrate an image annotation tool for annotations on a desktop computer typically performed by medical students (semi-automatic annotation), and a multimodal interface for expert annotations (manual annotation) into a common framework that benefits from manual, semi-automatic, and automatic image annotations. Mainly, a speech-based system for manual annotations of experts should be developed. For new incoming patients, the doctors have to maintain the database and search for similar cases in real-time. Multimodal user interfaces play a significant role in achieving this goal. The system should support the full range of multimodal interaction patterns, such as deictic or cross-modal references in the context of the annotation process. A remote RDF repository which stores the semantic medical image information and connects the annotation and querying task into a common framework, should make the overall architecture relevant to clinical practice.

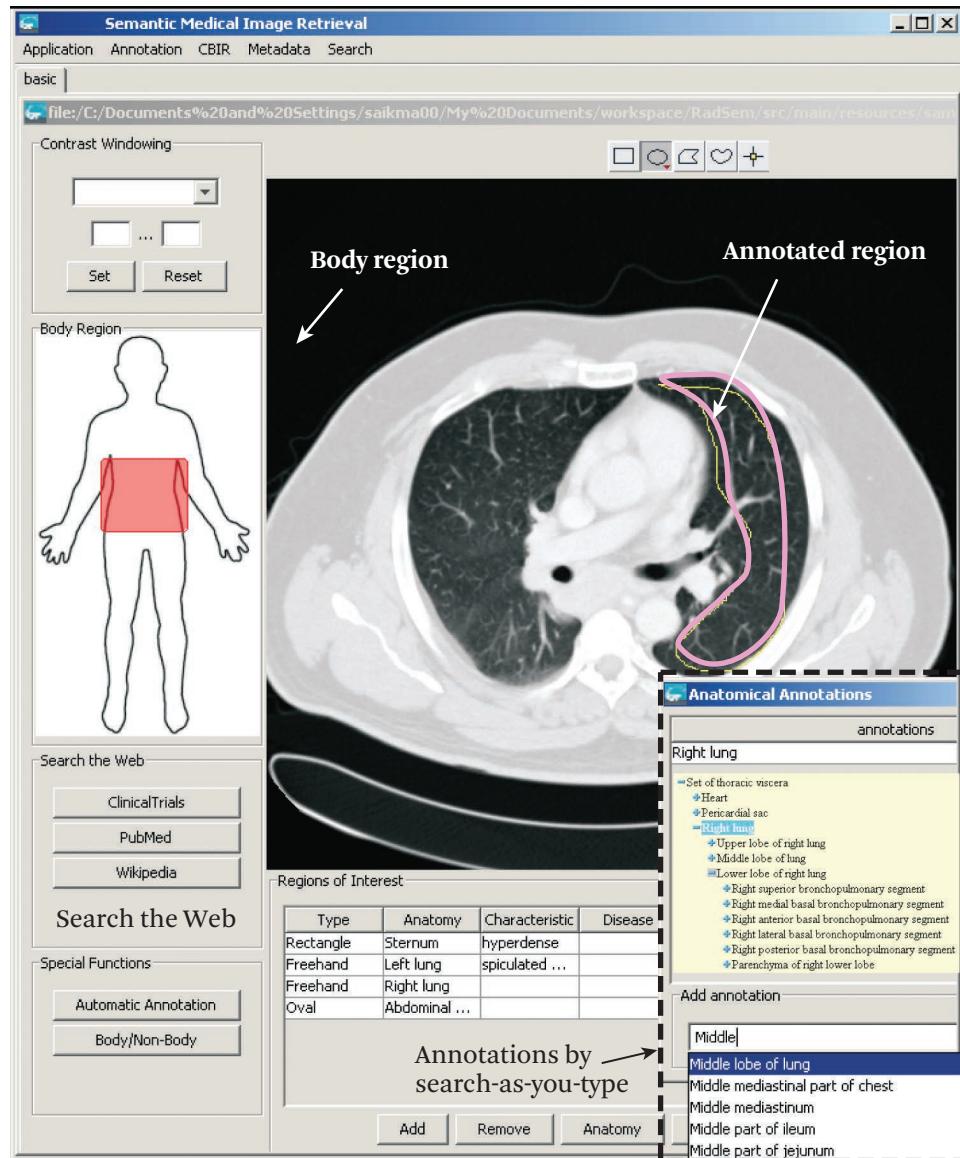
#### **11.4.1.3 Solution**

For the semantic annotation on a regular desktop workstation, Möller et al. [2009] developed RadSem, a medical semantic annotation and retrieval tool. It consists of a component that implements a method to annotate images, and upload/maintain a remote RDF repository with the images and image semantics. For annotations, RadSem reuses existing reference ontologies and terminologies. More precisely, the Foundational Model of Anatomy (FMA) ontology [Mejino et al. 2008] for anatomical annotations, i.e., annotations of body parts. To express features of the visual

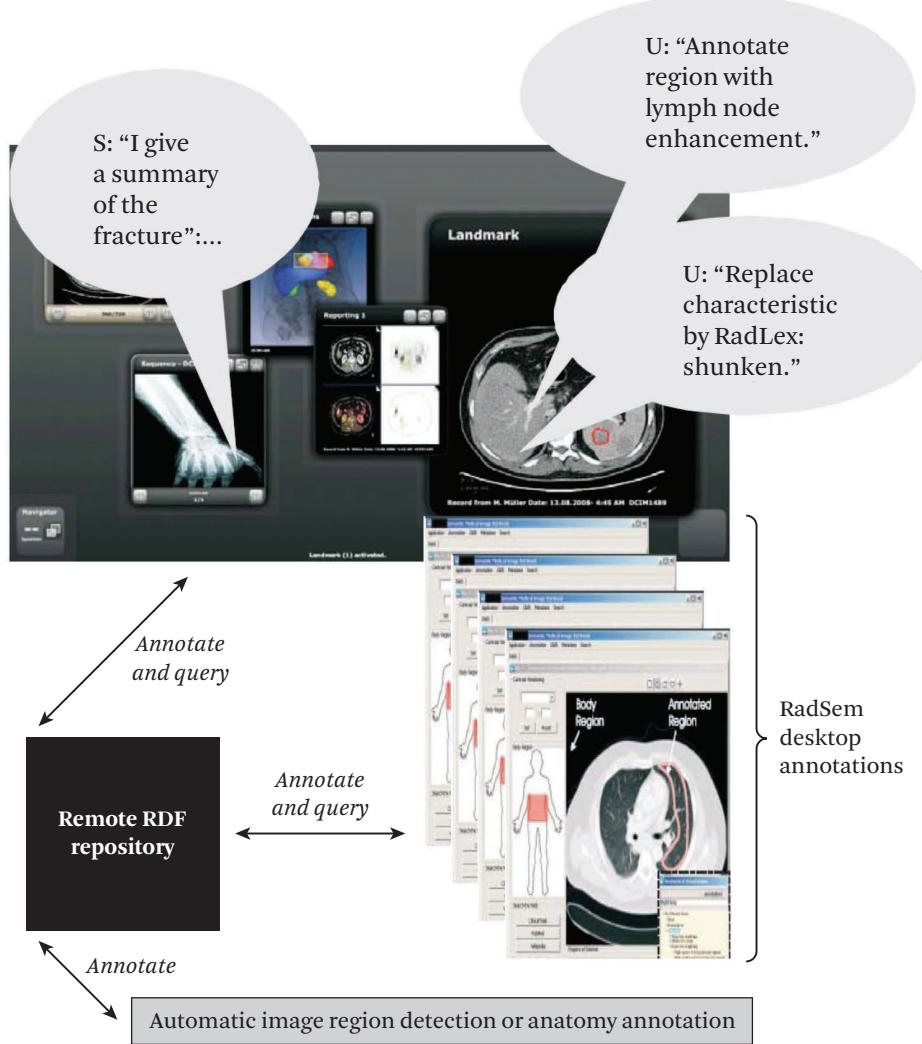
manifestation of a particular anatomical entity or disease of the current image, RadSem uses fragments of the RadLex ontology, see [Langlotz \[2006\]](#). Diseases are formalized using the International Classification of Diseases (ICD-10) [[Möller et al. 2010](#)]. Figure 11.2 shows the graphical user interface of the RadSem annotation tool. Images can be segmented into regions of interest (ROI). Each of these regions can be annotated independently with anatomical concepts (e.g., “lymph node”), with information about the visual manifestation of the anatomical concept (e.g., “enlarged”), and with a disease category using ICD-10 classes (e.g., “Nodular lymphoma” or “Lymphoblastic”). However, any combination of anatomical, visual, and disease annotations is allowed, and multiple annotations of the same region are possible. The resulting annotations (mostly anatomical, performed by medical students) are stored in the RDF repository.

In this usage scenario, the expert user—the radiologist—stands in front of the touchscreen installation (figure 11.3, upper part). The interactive system is based on a generic framework for implementing multimodal dialogue systems. Technically, the generic framework follows an object-oriented programming model that eases the interface to external third-party components (i.e., the automatic speech recognizer (ASR) and the text-to-speech synthesis (TTS) component) while using ontology concepts in a model-based design. Several interfaces for the multimodal framework have been implemented: the multimodal touchscreen interface, the event bus, the speech dialogue system, and the application backend as a remote RDF repository. The multimodal touchscreen interface is implemented as a native application using a special window manager for pointing gestures on a touchscreen display. The client provides means to connect to the dialogue system via an event bus, to notify it of occurred events, to record and playback audio streams, and to render the received display data obtained from the dialogue system. The dialogue system contains an ontology-based rule engine for processing dialogue grammars and an external service connector.

The diagnostic analysis of medical images typically concentrates around two questions: (i) What is the anatomy? (ii) Is it normal or abnormal? To satisfy the radiologist’s information need, he or she can formulate the questions in natural speech when a respective image annotation exists. Most importantly, the multimodal interface helps to annotate the respective images and image regions during the patient finding process. Our prototype systems gives a first answer to the following two research questions. First, what kind of information is relevant for the radiologist’s daily tasks (a combination of annotation and retrieval). Second, at what stage of the workflow should selected information items be offered and aggregated/annotated



**Figure 11.2** Desktop interface of the annotation tool radSem [Sonntag and Möller 2010] for manual semantic annotations of medical images.



**Figure 11.3** Combined multimodal user interface for the semantic annotation and retrieval of medical images.

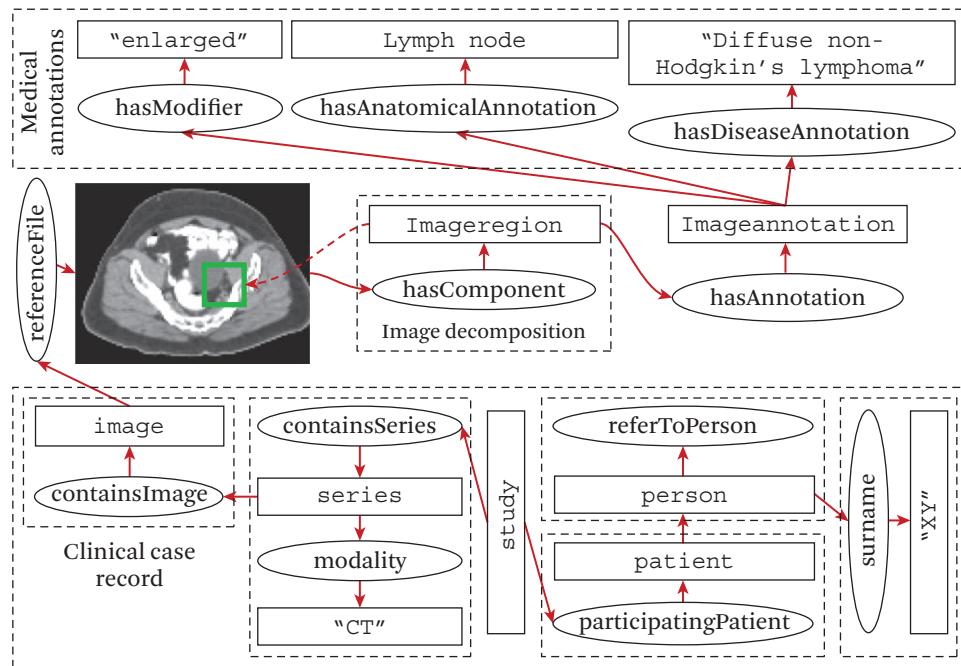
in the diagnostic process while using a touchscreen and speech dialogue interface. A multimodal dialogue example is explained in the following.

1. U: "Show me the CTs, last examination, patient XY." (retrieval stage)
2. S: Shows corresponding patient CT study picture series.

3. **U:** "Show me the internal organs: lungs, liver, then spleen."
4. **S:** Shows patient images according to referral record.
5. **U:** "Annotate this with lymph node enhancement" (+ pointing gesture on region); "so *lymphoblastic*" (expert finding, additional disease annotation (ICD-10)).
6. **S:** "Region has been annotated."
7. **U:** "And replace the characteristic of the other by RadLex: shrunken."
8. **S:** "Region characteristic has been updated." → The radiologist switches to another patient (for illustration purposes with a broken finger) and asks for a summary in this additional retrieval stage.
9. **U:** "Give me a summary of this patient." (retrieval stage)
10. **S:** "This is a summary of the fracture: . . . "
11. **S:** "Five corresponding CTs will be displayed." → The radiologist can now switch again to the differential diagnosis of the suspicious case together with a second medical expert (for the first patient), where the case is examined again and the image annotations can be completed.

A variety of multimodal interaction patterns are implemented in this dialogue, e.g., the resolution of multimodal references. In (5), a deictic reference is resolved (a pointing gesture uniquely singles out an object, it is said to have object-pointing function), whereby in (7) an exophoric reference is given by "the other" annotation already present; it refers to the environment in which the dialogue is taking place, the context of situation is what is displayed on the screen. The command "annotate with" is an implicit reference in the context of the CT image in the current focus. Last but not least, the system builds an own anaphoric reference "corresponding" in (11).

Queries are sent to the open source triple store [Sesame \[2017\]](#). A direct access to the RDF statements is possible while using the query language SPARQL. This allows us to specify queries of almost arbitrary complexity. Queries can span from patient metadata to image annotations to medical domain knowledge and are used to translate the dialogue questions into SPARQL statements. Figure 11.4 shows a graphical representation of the RDF graph that is retrieved when using the query. System responses are based on the retrieved RDF graph. The following SPARQL query example is a translation of the clinician's dialogue question, "Show me the CTs, last examination, patient XY."

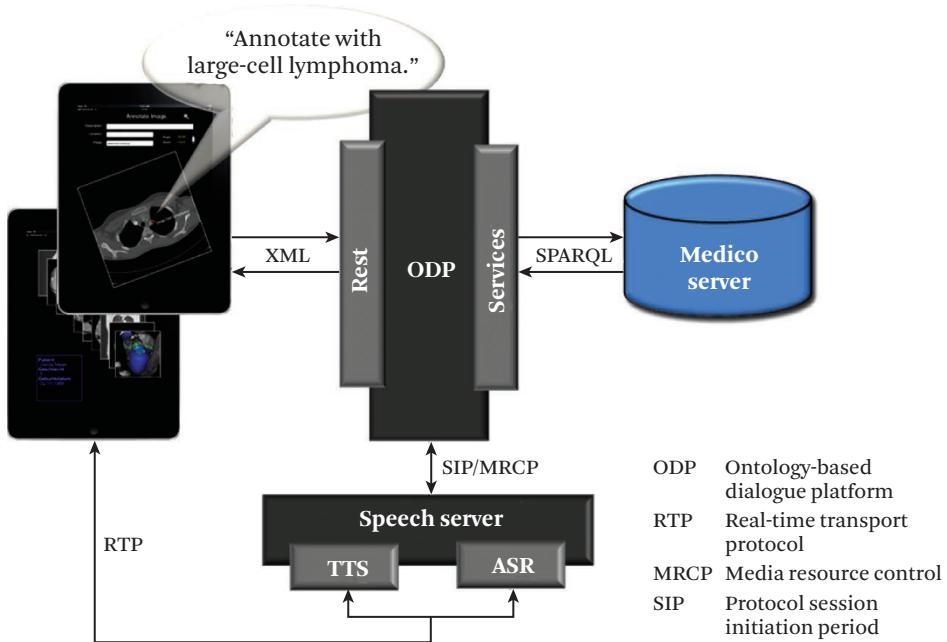


**Figure 11.4** RDF result graph with medical annotations.

```

SELECT ?person ?patient ?imageURL
WHERE {
  ?person mao:surname ?var0 .
  FILTER (regex(?var0, "XY", "i")) .
  ?patient mdo:referToPerson ?person.
  .
  .
  ?series mdo:modality "CT".
  ?series mdo:containsImage ?image.
  ?image mdo:referenceFile ?imageURL.
  .
}
  
```

While using a tablet for interaction (see the architecture in Figure 11.5), additional multisensor information from the tablet such as accelerometer or gyroscope information can be used in future radiology interfaces. The communication between the tablet and the dialogue system is based on state-of-the-art web service protocols: the Representational State Transfer (REST) is an established stan-



**Figure 11.5** Architecture of the Radspeech system showing components and data protocols; see [RadSpeech \[2011\]](#).

dard that defines a set of constraints to be used for creating web services for distributed information retrieval applications. In summary, the multimodal interface allows the user to annotate medical images with ontology-based medical concepts (RadLex); the annotations are directly transferred to a remote RDF repository. At this point, the radiologist can (1) access the images and image region annotations (a summary can also be synthesized), (2) complete them, and (3) refine existing annotations while using a multimodal dialogue system. Finally, the RDF repository is updated again. Future plans include the implementation of the speech-based dialogue system in virtual reality where 3D images can be inspected and annotated (also cf. future directions in Section 11.5).

### 11.4.2 Case Study 2: A Multisensor Digital Pen Interface

This case study describes a categorization and implementation of digital pen sensors for behavior characterization. We focus on the clinical interpretation of time-stamped stroke data from digital dementia tests. Based on using digital pens in breast imaging for instant knowledge acquisition [[Sonntag et al. 2014a](#)], where the

doctor uses the digital pen for reporting, we now begin to use the digital pen for the patient [Prange et al. 2015].

#### **11.4.2.1 Background**

This research is situated within a long-term project Kognit [Sonntag 2015] with the ultimate goal of developing cognitive assistance for patients with automatic assessment, monitoring, and compensation in the clinical and non-clinical context. In the clinical context, we can identify a special target group of interactive cognitive assessment tools as public sector applications: cognitive assistance for doctors in terms of automatically interpreted clinical dementia tests. We think that automatic and semi-automatic clinical assessment systems for dementia have great potential and can improve quality care in healthcare systems. Our new project Interakt [Sonntag 2017] with clinical partners from Charité in Berlin complements previous fundamental research projects for non-clinical interfaces for dementia patients and clinical data intelligence [Sonntag 2015, Sonntag et al. 2015].

Previous approaches of inferring cognitive status from subtle behavior in the context of dementia have been made in a clock drawing test (CDT), a simple pencil and paper test that has proven useful in helping to diagnose cognitive dysfunction such as Alzheimer's disease. This test is the de facto standard in clinical practice as a screening tool to differentiate normal individuals from those with cognitive impairment and has been digitized in a first version with a digital pen only recently [Davis et al. 2014, Souillard-Mandar et al. 2016]. As pointed out in Davis et al. [2014], the use of (1) a digital pen on paper or (2) a tablet and stylus may distort results by its different ergonomics and its novelty. We implement both interfaces for a selection of standard dementia tests in this case study. This should inform future developments of objective neurocognitive testing methods. In particular, we address the issue of what role automation could play in designing multimodal-multisensor interfaces to support precise medical assessments.

#### **11.4.2.2 Problem Description**

Neurocognitive testing assesses the performance of mental capabilities, including for example, memory and attention. Most cognitive assessments used in medicine today are paper-pencil based. A doctor, physiotherapist, or psychologist conducts the assessments. These tests are both expensive and time consuming. Furthermore, the results can be biased. In addition to understanding people, their processes, their needs, their contexts, in order to create scenarios in which Artificial Intelligence (AI) technology can be integrated, we are particularly concerned to assess and predict the healthcare status with unintrusive sensors such as those in digital

pens or in tablets. The goal is to improve the diagnostic process of dementia and other forms of cognitive impairments by digitizing and digitalizing standardized cognitive assessments for dementia. Here digitizing is the process of changing from analog assessments to digital forms with handwriting and gesture recognition. Digitalization is the process to include automatic assessments into the caregiver's task. We aim at weekly procedures in day clinics and base the assessments on clinical test batteries such as the CERAD developed by Morris et al. [1988]. The test is digitized by handwriting recognition and sketch recognition. Additional new parameters are provided by the digital pen's internal sensors. The conducted cognitive walkthrough for digitalization started with a task analysis with experts at the clinic that specifies the sequence of steps or actions a doctor requires to accomplish a pencil-paper-based assessment task as well as the potential system responses to a digitalized version of it. In this case study we identified together with clinical experts that using a digital pen has the following potential benefits:

- the caregiver's time to spend on conducting the test can be reduced;
- the caregiver's time to spend on evaluating the written form can be reduced;
- the caregiver's attention can be shifted from test features while writing (e.g., easy-to-assess completion of input fields) to important verbal test features;
- digital assessments are potentially more objective than human assessments and can include non-standardized tests and features (for example timing information) whereby previous approaches leave room for different subjective interpretations;
- they can be used to get new features of the pen-based sensor environment, to detect and measure new phenomena by more precise measurement;
- they are relevant for new follow-up checks, and they can be conducted and compared in a rigorous and calibrated way;
- they can automatically adapt to intrinsic factors (e.g., sensorimotor deficits) if the user model is taken into account;
- they allow for evidence in the drawing process (e.g., corrections) instead of static drawings that look normal on paper;
- they reduce extrinsic factors (e.g., misinterpreted verbal instructions); and
- they can, in the future, be conducted in non-clinical environments and at home.

The challenges we face are three-fold.

1. To identify interface design principles that most effectively support automatic and semi-automatic digital tests for clinical assessments.
2. At the computational level, it is important to investigate approaches to capture both digital pen features and multimodal-multisensor extensions. Some tests assume content features (what is written, language use, perseveration, i.e., the repetition of a particular response such as a word, phrase, or gesture) in usual contexts, as well as para-linguistic features (how is it written, style of writing, pauses, corrections, etc.). These are potential technical difficulties and/or limitations in the interpretation of the results.
3. At the interface level, it is important to devise design principles that can inform the development of innovative multimodal-multisensor interfaces for a variety of patient populations, test contexts, and learning environments.

#### **11.4.2.3 Solution**

The scenario includes the doctor and the patient at a table in a day clinic (Figure 11.6) which provides most utility. In the following, we focus on the doctor's assessment task. Here, the term utility refers to whether the doctors' intelligent user interface provides the features they need. The conducted cognitive walkthrough started with a task analysis with experts at the clinic that specifies the sequence of steps or actions a doctor requires to accomplish a pencil-paper-based assessment task as well as the potential system responses to a digitalized version of it. According to the requirements, we implement a sensor network architecture to observe states of the physical world and provide real-time access to the state data for interpretation. In addition, this context-aware application may need access to a timeline of past events (and world states) in terms of context histories for reasoning purposes while classifying the input data. The result of the real-time assessment of the input stroke data and context data is presented to the doctor in real-time; see Figure 11.6. The display includes (1) summative statistics of test performances, (2) real-time test parameters of the clock drawing test and similar sketch tests, and (3) real-time information about pen features such as tremor and in-air time of the digital pen.

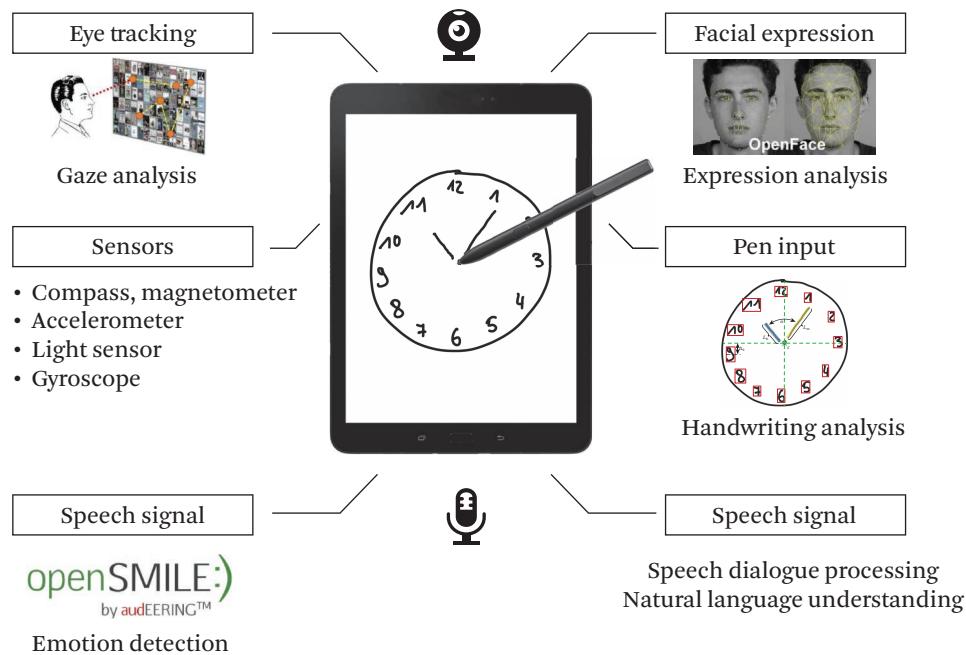
Usability design choices, how easy and pleasant the interface is to use, are made according to industrial usability guidelines [Sonntag et al. 2010b] based on usability inspection methods [Nielsen and Mack 1994] and design heuristics based on the psychophysiology of stress [Moraveji and Soesanto 2012]. They can be summarized as follows. For the patient, the digital pen is indistinguishable from a normal pen. So usability is high and (additional) stress is generally low. But the



**Figure 11.6** Assessment environment with patient and doctor. It also shows the realtime intelligent user interface for the doctor.

psychophysiology of stress needs to be explored. Lupien et al. [2007] suggest that some of the age-related memory impairments observed in the literature could be partly due to increased stress reactivity in older adults to the environmental context of testing. For the doctor, the psychophysiology of stress needs to be explored, too. There needs to be a possibility to control interruptions (e.g., phone calls) [Moraveji and Soesanto 2012]. In general, for both user interfaces, the effects of stress and stress hormones on human cognition are important. Lupien et al. [2007] enumerate the following stressor characteristics (SC) of interfaces that we use to form further design principles: SC1: feels unpredictable, uncertain, or unfamiliar in an undesirable manner; SC2: evokes the perception of losing/lost control; SC3: has potential to cause harm or loss to one's self or associated objects, living things, or property; and SC4: is perceived as judgment or social evaluative threat including threats to one's identity or self-esteem. SC4 especially applies in the situation of the patient assessment. Digital pen on normal paper reduces this effect, whereby using a tablet and stylus might increase SC4 stress levels.

The therapist interface, where the real-time interpretations of the stroke data are made available in RDF, is meant to advance existing neuropsychological testing technology according to our interface design principles. Technical details are



**Figure 11.7** Multimodal-multisensor tablet device.

as follows. First, it provides captured data in real-time (e.g., for a slow-motion playback). Second, it classifies the analyzed high-precision information about the filling process, opening up the possibility of detecting and visualizing subtle cognitive impairments; also it is zoomable to permit extremely detailed visual examination of the data if needed (as previously exemplified in [Davis et al. \[2014\]](#)).

Multimodal-multisensor extensions can be implemented with a tablet device (figure 11.7). Additional modalities can help in the analysis of observed user behavior. When interacting with a tablet computer, multiple built-in sensors can be used in addition.

Besides pen-based input, we consider eye tracking and facial expression analysis via the video signal of the front-facing camera, natural speech captured by the built-in microphone, and additional sensor inputs of modern tablet devices. RGB-based eye tracking is interesting for multimodal interaction with a tablet, because it is deployable using the built-in front-facing camera. However, this gaze estimation is erroneous which should be considered in the interaction design [[Barz et al. 2018](#)]. OpenFace [[Baltrušaitis et al. 2018b](#)] is an open source toolkit for facial behavior analysis using the stream of an RGB-webcam. It provides state-of-the-art performance in facial landmark and head pose tracking, as well as facial action

**Table 11.1** Comparison of the most widely used cognitive assessments

Name	Pen Input	Symbols
AKT - Age-Concentration	100%	cross-out
CDT - Clock Drawing Test	100%	clock, digits, lines
CERAD - Neuropsychological Battery	20%	circles, rectangles, cubes, etc.
DemTect - Dementia Detection	20%	numbers, words
MMSE - Mini-Mental State Examination	9%	pentagrams
MoCA - Montreal Cognitive Assessment	17%	clock, digits, lines
ROFC - Rey-Osterrieth	100%	circles, rectangles, triangles, lines
TMT - Trail Making Test	100%	lines

unit recognition which can be used to infer emotions. The openSMILE toolkit [Eyben et al. 2013] provides methods for speech-based behavior analysis and is distributed under an open source license. It offers an API for low-level feature extraction from audio signals and pre-trained classifiers for voice activity detection, speech-segment detection and speech-based emotion recognition in real-time.

The implemented pencil and paper tests are shown in Table 11.1, namely AKT [Gatterer et al. 1989], CDT [Freedman et al. 1994], CERAD [Morris et al. 1988], DemTect [Kalbe et al. 2004], MMSE [Folstein et al. 1975], MoCA [Nasreddine et al. 2005], ROFC [Canham et al. 2000], and TMT [Reitan 1992].

The pencil and paper tests have been transferred one-to-one, meaning that the digital versions of pen input fields look just as the analog versions. Table 11.1 shows the absolute percentages of the test questions where the pen is used to answer them. The selection of the tests accounts for a variety of patient populations and test contexts. Concerning the test context, a doctor can always switch between the digital pen and the tablet and stylus version. The tablet version can always use multimodal-multisensor input to cover additional test contexts.

#### 11.4.2.4 Lessons Learned

In this section we discuss which choices we have made in the first 18 months of the project Interakt, the analysis of alternatives considered, as lessons learned. We focus on specific designs and decisions that reduce the potential for failures when considering similar applications.

1. The primary motivation of using a digital pen on normal paper stems from the spatial and temporal precision of the obtained stroke data which provides the basis for an unprecedented degree of precision during analysing these data for small and subtle patterns; classifying the strokes for their meaning

is a sketch interpretation task in addition. As a result, we can get assessment data based on what is written or sketched, and how the spatio-temporal pattern looks like. The alternative of using a tablet and stylus turned out to be an additional stress factor for both patients and doctors, as first formative evaluations suggest. As a result, the formative evaluations with patients will be done on the digital paper version. This choice restricts the possibility to gather multimodal data from a tablet, which provides the same spatial and temporal precision of the obtained stroke data.

2. While the tablet version is not always the first choice, the technical implementation is much easier than the digital pen on normal paper version. The reason is the complicated software development kit (SDK) for creating the digital paper forms on normal paper.
3. How will the data from the experiment be gathered without violating privacy regulations [[Friedland and Tschantz 2018](#)]? For example, video capture is currently not allowed. In addition, we need a method to capture assessment results (or corrections/comments) from the doctor while he or she is using the doctor's interface. Will it interfere with the anticipated normal use? Here, the enumeration of the stressor characteristics need to be completed and turned into interface design principles.
4. A version for self-assessment at home for the patient needs to have an ability to control interruptions (e.g., phone calls) [[Moraveji and Soesanto 2012](#)].
5. The digitalization of widely used cognitive assessments has four consecutive steps: (1) the one-to-one transfer from a paper and pencil test to a digital version; (2) the selection of pen features that are relevant for the classification task; (3) the adaptation of the caregivers' instructions to include automatically interpreted test results; and (4) the inclusion of multimodality and multisensor data for additional test parameters.
6. Digital assessments allow for evidence in the drawing process (e.g., corrections) instead of static drawings that look normal on paper. Doctors need to be instructed when to use the slow-motion playback function. To automatically propose replaying a writing scene for further inspection is another interesting classification task where the system can take initiative.
7. The coverage of implemented tests is rather independent of the availability of suitable patient populations and test subjects. It is rather difficult to get the critical amount of conducted tests for machine learning experiments to find subtle patterns that are sensitive or specific to dementia assessment.

Using digital pens for the assessment of dementia can be generalized in several ways, most notably for use by those in the cognitive impairments field. Digitalized dementia tests can be used for the detection of other neurodegenerative diseases such as Parkinson's. Some of the described tests in Table 11.1 have already been used in this direction, such as MMSE and a more sensitive similar test MoCA. In addition, this work could help returning veterans suffering from traumatic brain injuries (TBIs). J Wagner et al. [2011] used CDT to assess cognition and predict inpatient rehabilitation outcomes among persons with TBI. Doctors working in inpatient neurorehabilitation settings are often asked to evaluate the cognitive status of persons with TBI and to give opinions on likely rehabilitation outcomes. In this clinical setting, several other digital pen tests could be used for cognitive assessment and outcome predictor among inpatients receiving neurorehabilitation after TBI. It should be possible to better monitor the rehabilitation outcome. As explained previously, digital assessments could be relevant for new follow-up checks. They can be conducted and compared in a rigorous and calibrated way.

Future research in the clinical domain includes pen-based assessments to treat patients in an automatic fashion and from multimodal input. Concerning multimodal input, for interpreting verbal utterances of the CERAD test battery for example (therapists have problems in taking notes of user answers and comments while conducting a test), a dialogue framework can be used in the future. Combining active speech and pen input should, in the future, be explored toward multimodal approaches to determining cognitive status. This can be done through the detection and analysis of subtle behaviors and skin conductance sensors.

Using digital pens for the assessment of dementia can be generalized for use by those outside the cognitive impairments field. Current research investigates the use of handwriting signal features to predict domain expertise in several educational contexts [Oviatt et al. 2018c]. The trend toward multimodal learning analytics becomes apparent, where natural communication modalities like writing (or speech) are complemented with gestures, facial expressions, and physical activity patterns. The combination of our low-level stroke features with selected components of the implemented cognitive tests, together with the domain expertise prediction task in Oviatt et al. [2018c] might open up opportunities to design new educational technologies based on individualized writing data resulting in better user modeling.

#### 11.4.3 Case Study 3: A Multimodal-Multisensor Framework

In this case study, we report on a multimodal multisensor framework for recording and analyzing handwriting input that is captured using a digital pen (cf. Case Study 2) and electrodermal activity (EDA) captured by the Bitalino sensor board for

extending behavior characterization for cognitive assessments in cognitive impairment cases.

#### **11.4.3.1 Background**

In the future, large-scale community screening programs can arise from multimodal data collections to identify profiles of impairment across different cognitive, psychiatric, and functional disabilities. Multimodal-multisensor data guide differential diagnosis and further assessment, because digital assessments are unbiased to a large degree. Stress and emotion changes reflect the activity of the sympathetic branch of the autonomous nervous system [Boucsein 1992]. Because sweat is an electrolyte solution, changes in the sweat level lead to changes in the skin conductance or electrodermal activity. Changes in EDA, especially the skin conductance response (SCR), can be used to detect stress, affect, and arousal [Pecchinenda 1996, Saitis and Kalimeri 2016, Kurniawan et al. 2013, Zhai and Barreto 2006]. Because of this correlation of stress (and cognitive load) and EDA, we believe it to be a suitable tool for the indication of cognitive impairments, in combination with digital pen features.

#### **11.4.3.2 Problem Description**

To include EDA into future digital pen based screening methods is very interesting because it is a process tracing method (unobtrusive and continuous measure) for neural activity and can reflect psychological processes, but context and sensor fusion is needed because it is a multifaceted phenomenon (sensitive but not specific). The digital pen-based environment provides such a sensor fusion context for its interpretation. Toward a multimodal cognitive assessment framework, three challenges need to be addressed. First, the selection of a useful subset of multimodal digital tests together with their implementations. Second, the inclusion of EDA and pen data into a multimodal-multisensor platform. Third, the study design for a future multimodal and digital cognitive assessment framework.

#### **11.4.3.3 Solution**

One of the most often used assessments for cognitive impairment is the CDT (also see Case Study 2), where the subject is asked to draw an analog clock including the numbers of the clock face and a specific time [Freedman et al. 1994]. Most of the existing tools concentrate on automating existing scoring schemes, which were originally designed to be processed and evaluated by therapists. In this described solution, Niemann et al. [2018a] combine EDA with digital pen data for a direct

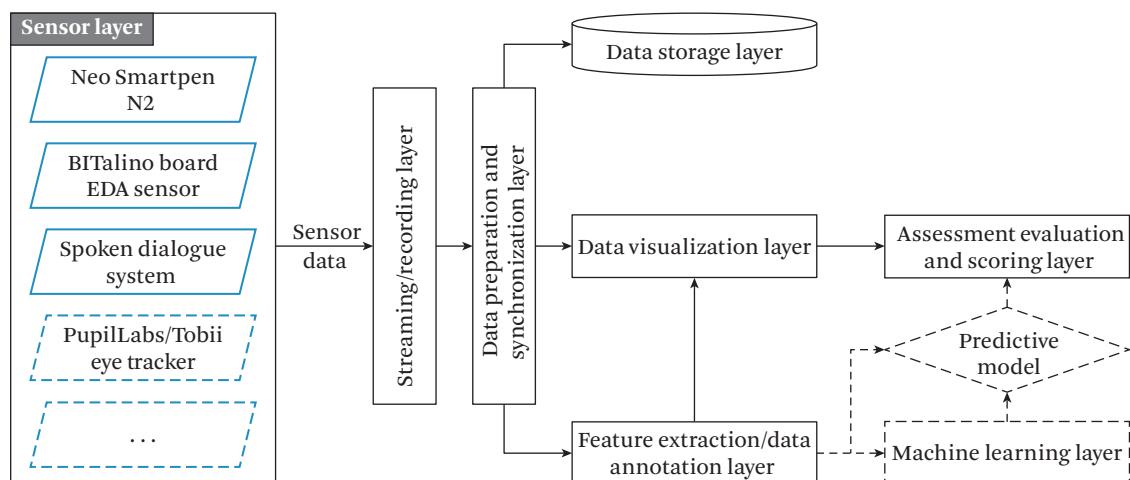
interpretation of writing behavior and biosignals: In addition to the traditional assessment categories (e.g., clock face numbers being in the correct place) they also take into account the EDA sensor data. In order to do so, writing tasks need to be split into semantic categories first. Figure 11.9 shows the visualization of a selected semantic feature set in the context of CDT. Participants are asked to draw a clock face with the time set to 11:10. The drawn clock is then examined by a trained physician and rated based on a predefined scoring scheme, reflecting the visual appearance and integrity of the clock using a numerical score. In CDT, we extract the following semantic features from the traditional scoring system.

- $c$  denotes the center point of the clock (centroid), the closer it is to the center of the clock's circle, the more points are awarded.
- $L_h$  and  $L_m$  represent the lengths of the hour and minute hands respectively. If the clock is well drawn, the hour hand should be shorter than the minute hand.
- The angle between the hour and minute hands is denoted as  $\alpha$ , together with the orientation of the hands it can be used to determine if the correct time was set.
- $\Delta_9$  is the displacement of clock face digits relative to their ideal location. In this example it is the vertical offset of digit number 9 to its correct center position.

With the help of semantic features and appropriate time stamps, the EDA signal can be fused with the digital pen interpretation and high EDA amplitudes traced back to writing and sketching tasks. In other words, the deviation of semantic feature interpretations from the norm indicate stress and cognitive load points that can be synchronized and validated by the sensitive, but non-specific, EDA signal. Similar semantic features can be extracted from similar tests [Prange et al. 2018b]. The test overview in Table 11.2 includes how much time the assessment usually takes, multimodal-multisensor aspects of active/passive pen input, active speech input, and passive EDA input, as well as how the assessment is evaluated by experts. The selection is based on the trade-off between the amount of tasks including handwriting input, the amount of movement during tasks (which might interfere with EDA signal data), and the need for additional speech input by the patient toward multimodal-multisensor interaction. An interesting point is how the evaluation is done: generalized standard scoring (g-scores) and/or individual scoring (i-scores) and/or time scoring (t-scores) or a combination of those. G-scores

**Table 11.2** Comparison of the most widely used cognitive assessments concerning multimodal input.

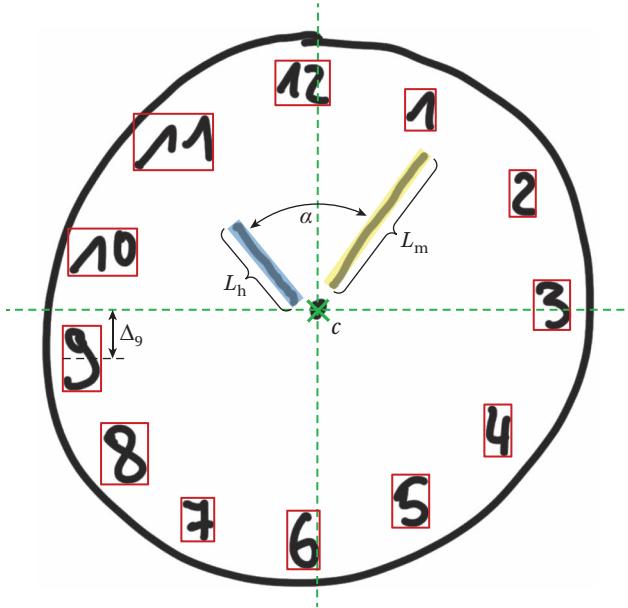
Name	Approx. time needed	Speech Input	Pen Input	EDA Input	Evaluation
AKT	15 min	NO	YES	YES	i-scores
DemTect	6–8 min	YES	YES	YES	s-scores
MMSE	5–10 min	YES	YES	YES	s-scores
TMT	3–5 min	NO	YES	YES	s-i-t-scores
CDT	2–3 min	NO	YES	YES	i-t-scores



**Figure 11.8** Architecture of the multimodal-multisensor platform used for behavior characterization. Dotted lines indicate work in progress.

are calculated by adding points for successfully solved questions and tasks. Final g-scores take age, sex, and similar factors into account. The digitalization of g-scores is straightforward. I-scores are individual interpretations of the doctor, as in the CDT for example. There is a huge potential to interpret those i-scores in a standardized way in the future. T-scores track the time needed to solve specific sub-tasks. They can be calculated automatically with a digital pen.

Now we focus on the multimodal-multisensor platform (Figure 11.8): To capture real-time handwriting input, we employ the Neo Smartpens N2 and M1 [Neosmartpen 2018]. We use the NeoSmartpen SDK that is available on Github for connecting the N2 digital pen and streaming the data using bluetooth. The



**Figure 11.9** Visualization of multimodal semantic features of the clock drawing test.

input is captured as a series of ink samples, which are grouped together forming strokes in pen-up and pen-down events, including timestamps and pen pressure data. The obtained digital ink gets automatically analyzed by the handwriting recognition component, which is based on the [Myscript \[2018\]](#) recognition engine. To distinguish between handwritten input and correction gestures, we employ mode detection described by [Sonntag et al. \[2014a\]](#). [Niemann et al. \[2018a\]](#) use several sensors to capture input and biosignals of the subject during the assessment and are planning to include more in the near future. The *Streaming and Recording Layer* serves as an abstraction layer for the specific hardware components used in the *Sensor Layer*, its main task is to stream and record the raw sensor data. Depending on the exact sensor type different libraries are used to connect the sensor to the overall framework. Synchronization, resampling, and input fusion take place during the *Data Preparation & Synchronization Layer*. Resulting data are visualized (*Data Visualization Layer*) and stored together with the raw input (*Data Storage Layer*) for later usage. The *Feature Extraction/Data Annotation Layer* prepares the handwriting input and EDA sensor data for machine learning tasks (interpretation and late fusion).

Now the focus is on the study design: Werner et al. [2006] used computerized handwriting evaluation to discriminate Mild Alzheimer Disease (AD) and Mild Cognitive Impairment (MCI). They observed that participants with MCI and mild AD spent a significantly longer time with the pen in the air than healthy participants and that all kinematic measures (except for velocity) differ between healthy and impaired participants. Findings by Schroter et al. [2003] suggest that it is possible to distinguish between different forms of cognitive impairment and healthy subjects by analyzing the kinematic aspects of handwriting movements. For the digital pen recording, we prepare the paper on the table, positioned in a comfortable position to the participant, and ask the patients to hold a digital pen in the same hand they hold a normal pen while writing. Inside the tip of the Neo pen an infrared camera recognizes the special microdot pattern printed onto the paper, which is merely visible and therefore similar to normal, white paper. If needed the subject is allowed to fixate the paper using the non-dominant hand and instructed to avoid movement with that hand if possible.

Movement of the hand results in either more or less pressure on the electrodes and, therefore, in noisy data. In order to minimize the chance of false signal peaks from unwanted movement, we suggest putting additional tape on the electrodes to fixate them in-place. The usage of additional cycling gloves has proven to be useful when recording tasks that might include frequent movement of the non-dominant hand (e.g., turning pages). From our experience, the amount of movement is highly dependent on the individual subjects. As EDA is a relatively slow signal (latency of about 0.5–5 s) [Boucsein 2012] in combination with the recovery time of the amplitude, there should be at least 6 s between each task of interest. It is advised to run a baseline measurement period between 2–4 min. when the participant is not engaged in any given task [Braithwaite et al. 2013].

Preliminary evaluations of the study design on the CDT, DemTect, and MMSE suggest that, through the automation of these cognitive assessments for dementia, the caregiver's time spent on conducting the tests can be reduced and his or her attention can be shifted from test features while writing (e.g., easy-to-assess completion of input fields) to other more subtle observations. EDA is displayed to the caregiver and helps in interpreting those subtle observations (yet to be quantified).

## 11.5

### Future Directions

Future directions include applications of multimodal-multisensor combinations and virtual reality applications which we will discuss in the rest of this chapter.

### 11.5.1 Multimodal-Multisensor Combinations

These interfaces combine multiple user input modalities with multiple sensor information (e.g., location, acceleration, proximity, tilt). Sensor-based cues may be used to interpret a user's physical state, health status, mental status, and many other types of information. Sensors may capture biosignals in addition, such as EDA (see case Study 3). On the other hand, users may engage in intentional actions when deploying sensor controls, such as hand gestures which are captured by a video sensor.

[Simsensei \[2014\]](#) introduced by [DeVault et al. \[2014\]](#) is an example of a multimodal-multisensor combination, i.e., a virtual agent-based interface with an additional collection of body sensors. This application should recognize and identify psychological distress from multiple signals in a multimodal dialogue. The mental status subsystem automatically tracks and analyzes in real-time facial expressions, body posture, acoustic features, linguistic patterns, and higher-level behavior descriptors (e.g., attention and fidgeting). It is very interesting to mention the two roles the multisensor system has. First, it contributes to the indicator analysis to identify psychological distress from those multiple signals. Just as in Case Study 2 with digital pen features, these distress indicators can allow the clinician or healthcare provider to make a more informed diagnosis. Second, the sensors' outputs are broadcasted to the other components of the multimodal interface: sensor outputs assist the virtual human with turn taking, listening feedback, and building rapport by providing appropriate non-verbal feedback.

Other applications are intelligent tutoring systems for educational healthcare. These systems use multimodal presentation of information to allow users (e.g., medical students) with different preferences and abilities to use information in their preferred way. In addition, multisensor processing might include speaker traits. For example, [Chatterjee et al. \[2015\]](#) analyze the most discriminative elements of a speaker's non-verbal behavior that contribute to the perceived credibility or passionateness.

Another example is Kognit [[Sonntag 2015](#)], a research project about multisensor input processing to counteract cognitive impairments, based on episodic memory construction through activity recognition. Kognit includes eye tracking sensors for activity recognition and multimodal speech-based dialogue in augmented reality applications. Eye tracking and activity recognition are explored in multiple upcoming research projects, but in Kognit, a robot senses a patient at home and interacts with him or her in a multimodal way. The patient can use speech or a digital pen to communicate with the robot. Multimodal output involves the robot's output from two or more modalities: A head-mounted visual display for the patient

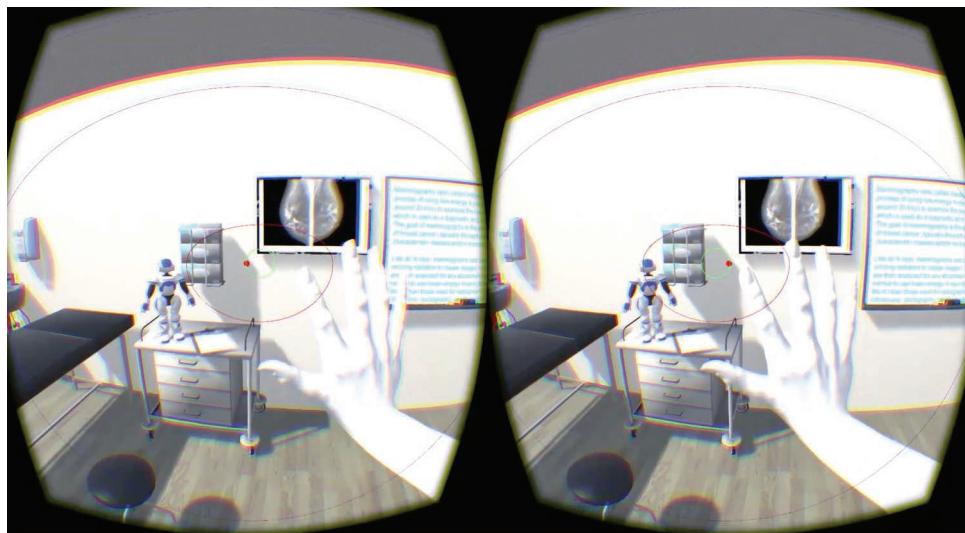
is combined with auditory feedback, which is provided as multimodal feedback to the user. Other important directions of multimodal-multisensor combinations include clinical multimodal-multisensor systems for doctors, where incremental knowledge acquisition, multimodal dialogue constraints, and virtual reality applications are brought together.

### 11.5.2 Virtual Reality

The design, development, and evaluation of virtual reality (VR) systems targets the areas of clinical diagnosis and decision support, clinical assessment, and rehabilitation. VR headsets provide a powerful new tool for future exploration of sensors, for example an Oculus Rift-integrated binocular eye tracking system [[Oculus 2018](#)]. In addition, digital pen-based interfaces can be combined with multiple approaches to determining cognitive status through the detection and analysis of subtle behaviors [[Davis et al. 2014](#)].

Examples of futuristic VR applications include clinical training [[Rizzo and Talbot 2016](#)]. Multimodal dialogues include avatars that respond to pre-selected choices, for example in the context of VR exposure therapy for combat-related post-traumatic stress disorder (PTSD) [[Cukor et al. 2016](#)]. [Greenleaf \[2016\]](#) states that significant impact of VR technology will be in the area of clinical medicine and healthcare, mostly because VR can address and ameliorate some of the most difficult problems in healthcare, i.e., ranging from mood disorders such as anxiety and depression to post traumatic stress disorder, addictions, autism, cognitive aging, and physical rehabilitation. VR examples include interactive visualization of shared electronic patient records, previously acquired with a remote tablet device, in a virtual environment. Hand tracking, eye tracking, and vision-based peripheral view monitoring can be integrated. [Luxenburger et al. \[2016\]](#) provide a combination of hand gesture and eye tracking recognition in order to assess whether all regions of a medical image have been explored by the doctor in the VR environment (Figure 11.10).

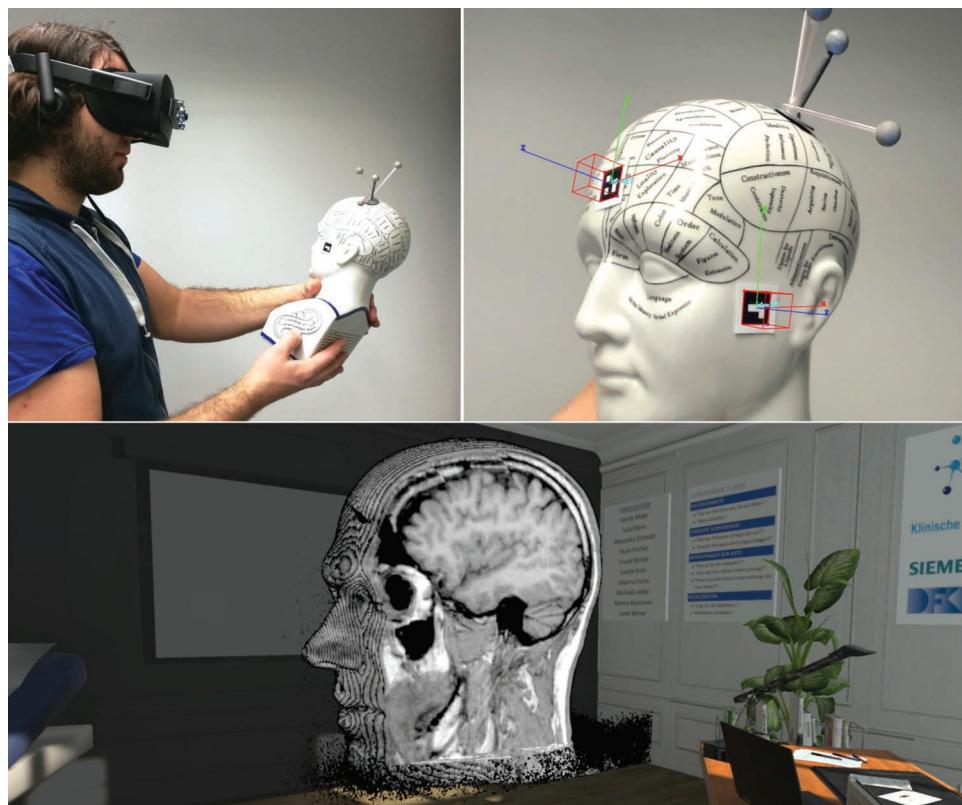
One of the main research questions for the multimodal interaction community is how to match affordances in VR with the medical task domain to (1) stimulate learning and understanding, (2) stimulate cognition, and (3) improve overall performance in medical decision support applications. Based on the case study presented in Section 11.4.1, [Prange et al. \[2018a\]](#) developed a multimodal VR prototype for the doctor. The system uses a headset google in a remote collaboration setting. In the application scenario, the radiologist starts with a patient examination form by using a tablet with built-in stylus for notes and drawing. The handwritten multi-stroke



**Figure 11.10** Medical remote collaboration using eye gaze and hand gesture input in VR.

sketches are transcribed by using handwriting and gesture recognition,<sup>1</sup> then analyzed and stored based on common medical ontologies [Sonntag et al. 2009]. The doctor then examines the patient records in VR, and he can interact with the 3D MRI medical images of the patient. Ard et al. [2017] present a similar scenario where neurology images are displayed in VR. The end-to-end system of Prange et al. [2018a] provides a GPU-accelerated machine learning model for automated decision support that computes therapy predictions in real-time. This video<sup>2</sup> shows the complete workflow. The dialogue system supports task-based interaction with the patient data shown on the virtual display (e.g., “Open the patient file for Gerda Meier.”, “Show the next page.”), question answering functionality about factoid contents of a patient record (e.g., “When was the last examination?”), and the therapy prediction component (“Which therapy is recommended?”). This prototype suggests that in future work, it is worth investigating how a VR application, together with dialogue-based therapy prediction, impacts the medical findings process in daily hospital routine, in particular when 3D images can be observed by haptic objects in a natural mapping (Figure 11.11). Recent advancement of VR technology for clinical purposes, i.e., added value over traditional diagnosis, decision support, or assess-

- 
1. <http://medicaleps.dfgi.de/www/wp-content/uploads/BIRADS-30-seconds.mp4>
  2. [http://medicaleps.dfgi.de/www/wp-content/uploads/KDI\\_V2\\_Pro\\_v04\\_2.mp4](http://medicaleps.dfgi.de/www/wp-content/uploads/KDI_V2_Pro_v04_2.mp4)



**Figure 11.11** Immersion by haptic objects in VR.

ment approaches, may lead to improved immersion effect. Multimodality aspects that can potentially lead to immersion are for example the following: pen, speech, (head) movement, VR controllers for input and a 3D VR scene, 3D image material, animated 3D graphics, and speech for input and output. Since people's object and concept perceptions are multisensor, people are influenced by an array of object affordances (e.g., auditory, tactile) and their visual properties. In addition, the acoustic qualities of a computer voice can influence a user's immersion and engagement. Future work includes additional input modalities such as eye tracking to improve the multimodal interaction in VR by other physiological sensors.

## 11.6 Conclusion

We discussed the trends of multimodal-multisensor interfaces of medical and health systems and emphasized the theoretical foundations of multimodal inter-

faces and systems in the healthcare domain. We started with a discussion of the background of medical and health systems, defined MCPS, and focused on the distinction of clinical and non-clinical systems, followed by three clinical use case studies. The first study described a multimodal dialogue system in the radiology domain, the second focused on a multisensor digital pen interface for cognitive assessment, and the third described a multimodal-multisensor framework including EDA. Future directions include multimodal-multisensor combinations.

For MCPS, in addition to the specific recommendations of multimodal-multisensor interactions, prototypes will have to go through product lifecycles, including the design, development, distribution, verification, validation, deployment, and maintenance of these devices. The exact challenges for real-world MCPS developers, in particular for verification and validation of such medical interfaces, remain mostly unknown. For example, EDA-based sensor input is easy to capture, but one gets motion artifacts if the hand with the electrodes moves as well as when the state of the hand changes, for instance, from open to close. The changed state of the hand results in either more or less pressure on the electrodes and, therefore, in more or less skin conductance. Future research in medical and health systems should include research on wearable body area networks for continuous monitoring of patients at home, based on wireless sensor networks for healthcare [[Alemdar and Ersoy 2010](#)]. [Darwish and Hassanien \[2011\]](#), for example, explain the important role of body sensor networks in medicine to minimize the need for caregivers and help the chronically ill and elderly people live an independent life.

In designing future architectures for multimodal-multisensor interfaces for medical and health systems, important insights clearly can be gained from cognitive principles of sensory integration of passive and active input modes. One challenge will be to create human-in-the-loop medical cyber-physical systems that incorporate a broad range of information by data-driven approaches of large multimodal databases. Those data-intensive systems can fuse multiple modalities. These infrastructures are capable of integrating multisensor input, potentially increasing the reliability of a percept through multisensor integration. A further consideration is improved robustness, and hence trust, in future assistive MCPS, where the higher level automatic intent recognition leads to collaborative action with humans in medical care. This also suggests that future research should explore whether individual patients' biomarkers may provide a useful signature for adaptation purposes (of advanced fusion-based multimodal interfaces for example). Such an approach is currently limited by diagnostic tools that are insensitive to changes in behavior future systems should adapt to automatically. Special, new biomarkers are of particular interest, i.e., markers of emotion regulation, social response and social attention. For example, learning representations of affect from speech [[Ghosh et al.](#)

2015] can be used in autism detection and treatment and prove that multimodal-multisensor interfaces have medical applications not only in sensomotoric and cognitive intelligence aspects, but also emotional and social intelligence aspects of medicine.

It is beneficial to take a broad perspective. We described the AI perspective that prime applications include clinical decision support, patient monitoring, and automated devices to assist in surgery or patient care. Concerning clinical decision support, advances of multimodal-multisensor interfaces can promise to change the cognitive tasks assigned to human clinicians by cognitive assistants and structured patterns of inference.

### **Focus Questions**

**11.1.** Why is WCET (worst-case execution time) an important consideration for medical CPS?

**11.2.** To support individuals on their personal health we must take a life-time perspective. Why?

**11.3.** What is the paradigm shift in healthcare provisions and how can we manage patients more effectively with multisensor information and multimodal communication technologies?

**11.4.** Why is it important to focus initially on people with certain risk factors such as cardiovascular risks like high blood pressure or high blood glucose?

**11.5.** How can emergency response services at home be improved by a multimodal-multisensor interface?

**11.6.** Why are automatic prevention services so cost-effective? How can multimodal-multisensor interfaces support primary and secondary prevention?

**11.7.** What information can be extracted from active input modalities to support medical applications based on biosignals?

**11.8.** How can multimodal-multisensor systems be integrated into daily life situations?

**11.9.** What adaptation approaches will result in effective multimodal and multisensor interfaces for real-word deployment with patients?

## References

- H. Alemdar and C. Ersoy. October 2010. Wireless sensor networks for healthcare: A survey. *Computer Networks*, 54(15): 2688–2710. DOI: [10.1016/j.comnet.2010.05.003](https://doi.org/10.1016/j.comnet.2010.05.003). 463
- E. Alpaydin. 2018. Classifying multimodal data. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107994](https://doi.org/10.1145/3107990.3107994). 427, 772
- T. Ard, D. M. Krum, T. Phan, D. Duncan, R. Essex, M. Bolas, and A. Toga. March 2017. NIVR: Neuro Imaging in Virtual Reality. In *Proceedings of Virtual Reality (VR), 2017 IEEE*, pp. 465–466. IEEE, Los Angeles, CA. DOI: [10.1109/VR.2017.7892381](https://doi.org/10.1109/VR.2017.7892381). 461
- T. Baltrušaitis, C. Ahuja, and L.-P. Morency. 2018a. Multimodal machine learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107993](https://doi.org/10.1145/3107990.3107993). 427, 772
- T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency. May 2018b. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 59–66. DOI: [10.1109/FG.2018.00019](https://doi.org/10.1109/FG.2018.00019). 450
- M. Barz, F. Daiber, D. Sonntag, and A. Bulling. 2018. Error-aware gaze-based interfaces for robust mobile gaze interaction. In *Proceedings of the International Symposium on Eye Tracking Research and Applications (ETRA)*. DOI: [10.1145/3204493.3204536](https://doi.org/10.1145/3204493.3204536). 450
- D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar. July 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs (Millwood)*, 33(7): 1123–1131. 439
- S. Bengio, L. Deng, L.-P. Morency, and B. Schuller. 2018. Multidisciplinary challenge topic: Perspectives on predictive power of multimodal deep learning: Surprises and future directions. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108006](https://doi.org/10.1145/3107990.3108006). 427, 772
- T. W. Bickmore, L. M. Pfeifer, and B. W. Jack. 2009. Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pp. 1265–1274. ACM, New York. DOI: [10.1145/1518701.1518891](https://doi.org/10.1145/1518701.1518891). 435
- W. Boucsein. 1992. *Electrodermal Activity*. Plenum Press. 454
- W. Boucsein. 2012. *Electrodermal Activity*. Springer Science & Business Media. 458
- J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe. 2013. A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1): 1017–1034. 458

- R. Canham, S. Smith, and A. Tyrrell. 2000. *Automated scoring of a neuropsychological test: The Rey-Osterrieth Complex Figure*, pp. 406–413. IEEE. [451](#)
- P. Carayon. 2011. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety, Second Edition*. CRC Press, Boca Raton, FL. [429](#)
- D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. 2015. Predicting daily activities from egocentric images using deep learning. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ISWC '15, pp. 75–82. ACM, New York. DOI: [10.1145/2802083.2808398](#). [436](#)
- M. Cavazza, F. Charles, A. Lindsay, J. Siddle, and G. Georg. 2015. An interactive narrative format for clinical guidelines. *KI - Künstliche Intelligenz*, 29(2): 185–191. DOI: [10.1007/s13218-015-0354-3](#). [432](#)
- M. Chatterjee, S. Park, L.-P. Morency, and S. Scherer. 2015. Combining two perspectives on classifying multimodal data for recognizing speaker traits. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pp. 7–14. ACM, New York. DOI: [10.1145/2818346.2820747](#). [459](#)
- P. Chaurasia, S. I. McClean, C. D. Nugent, and B. W. Scottney. 2014. A duration-based online reminder system. *International Journal of Pervasive Computing and Communications*, 10(3): 337–366. DOI: [10.1108/IJPC-07-2014-0042](#). [436](#)
- C. Chen, J. Favre, G. Kurillo, T. P. Andriacchi, R. Bajcsy, and R. Chellappa. 2014a. Camera networks for healthcare, teleimmersion, and surveillance. *IEEE Computer*, 47(5): 26–36. DOI: [10.1109/MC.2014.112](#). [429](#)
- C. Chen, L. White, T. Kowalewski, R. Aggarwal, C. Lintott, B. Comstock, K. Kuksenok, C. Aragon, D. Holst, and T. Lendvay. 2014b. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research*, 187(1): 65–71. [//www.sciencedirect.com/science/article/pii/S0022480413008998](http://www.sciencedirect.com/science/article/pii/S0022480413008998). DOI: <http://dx.doi.org/10.1016/j.jss.2013.09.024>. [433](#)
- Y. Chen, Y. Chen, M. Randriambelonoro, A. Geissbühler, and P. Pu. 2016. Peer influence on the engagement of fitness tracker usage: A diabetic and obesity study. *2016 IEEE International Conference on Healthcare Informatics*, ICHI 2016, Chicago, IL, USA, October 4–7, 2016, pp. 192–197. DOI: [10.1109/ICHI.2016.28](#). [437](#)
- J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer. 2018. Multimodal assessment of depression and related disorders based on behavioural signals. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108004](#). [427](#), [764](#)
- J. Cukor, M. Gerardi, S. Alley, C. Reist, M. Roy, B. O. Rothbaum, J. Difede, and A. Rizzo. January 2016. Virtual reality exposure therapy for combat-related PTSD. In *Posttraumatic Stress Disorder and Related Diseases in Combat Veterans*, pp. 69–83. Springer International Publishing, Cham, Switzerland. DOI: [10.1007/978-3-319-22985-0\\_7](#).

- H. P. da Silva, A. Fred, and R. Martins. 2014. Biosignals for everyone. *IEEE Pervasive Computing*, 13(4): 64–71. DOI: [10.1109/MPRV.2014.61](https://doi.org/10.1109/MPRV.2014.61). 437
- A. Darwish and A. E. Hassanien. 2011. Wearable and implantable wireless sensor network solutions for healthcare monitoring. *Sensors*, 11(6): 5561–5595. <http://www.mdpi.com/1424-8220/11/6/5561>. DOI: [10.3390/s110605561](https://doi.org/10.3390/s110605561). 463
- R. Davis, D. Libon, R. Au, D. Pitman, and D. Penney. 2014. Think: Inferring cognitive status from subtle behaviors. *Innovative Applications of Artificial Intelligence (IAAI)*. <http://www.aaai.org/ocs/index.php/IAAI/IAAI14/paper/view/8626>. DOI: [10.1609/aimag.v36i3.2602](https://doi.org/10.1609/aimag.v36i3.2602). 446, 450, 460
- F. de Man, S. Greuters, C. Boer, D. Veerman, and S. Loer. 2013. Intra-operative monitoring, many alarms with minor impact. *Anaesthesia*, 68. DOI: [10.1111/anae.12289](https://doi.org/10.1111/anae.12289). 429
- D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pp. 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. <http://dl.acm.org/citation.cfm?id=2617388.2617415>. 459
- S. K. D'Mello, N. Bosch, and H. Chen. 2018. Multimodal-multisensor affect detection. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107998](https://doi.org/10.1145/3107990.3107998). 427, 437, 764
- F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pp. 835–838. ACM, New York. DOI: [10.1145/2502081.2502224](https://doi.org/10.1145/2502081.2502224). 451
- M. Folstein, S. Folstein, and P. McHugh. November 1975. “Mini-mental state.” A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3): 189–198. 451
- E. S. Ford, M. M. Bergmann, J. Kröger, A. Schienkiewitz, C. Weikert, and H. Boeing. 2009. Healthy living is the best revenge: Findings from the european prospective investigation into cancer and nutrition?potsdam study. *Archives of Internal Medicine*, 169(15): 1355–1362. DOI: [10.1001/archinternmed.2009.237](https://doi.org/10.1001/archinternmed.2009.237). 426
- M. Freedman, L. Leach, E. Kaplan, G. Winocur, K. Shulman, and D. Delis. 1994. *Clock Drawing: A Neuropsychological Analysis*. Oxford University Press, Oxford, UK. 451, 454
- G. Friedland and M. Tschantz. 2018. Privacy concerns of multimodal sensor systems. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA. 430, 434, 452

- M. Garbarino, M. Lai, D. Bender, R. Picard, and S. Tognetti. November 2014. Empatica E3 - A wearable wireless multisensor device for real-time computerized biofeedback and data acquisition. *2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*, pp. 39–42. DOI: [10.1109/MOBILEALTH.2014.7015904](https://doi.org/10.1109/MOBILEALTH.2014.7015904). 437
- G. Gatterer, P. Fischer, M. Simanyi, and W. Danielczyk. 1989. The A-K-T (“Alters-Konzentrations-Test”) a new psychometric test for geriatric patients. *Functional Neurology*, 4(3): 273–276. 451
- J. Gelissen and D. Sonntag. 2015. Special issue on health and wellbeing. *KI - Künstliche Intelligenz*, 29(2): 111–113. DOI: [10.1007/s13218-015-0360-5](https://doi.org/10.1007/s13218-015-0360-5). 426
- GenieMD, 2017. GenieMD. <http://geniemd.com>. 2018-03-21. 434
- S. Ghosh, E. Laksana, L. Morency, and S. Scherer. 2015. Learning representations of affect from speech. *CoRR*, abs/1511.04747. <http://arxiv.org/abs/1511.04747>. 463, 464
- D. Graus, P. N. Bennett, R. W. White, and E. Horvitz. 2016. Analyzing and predicting task reminders. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP ’16, pp. 7–15. ACM, New York. DOI: [10.1145/2930238.2930239](https://doi.org/10.1145/2930238.2930239). 436
- W. Greenleaf. 2016. How VR Technology Will Transform Healthcare. In *ACM SIGGRAPH 2016 VR Village*, SIGGRAPH ’16, pp. 5:1–5:2. ACM, New York. DOI: [10.1145/2929490.2956569](https://doi.org/10.1145/2929490.2956569). 460
- S. Haddadin, A. D. Luca, and A. Albu-Schäffer. December 2017. Robot collisions: A survey on detection, isolation, and identification. *IEEE Transactions on Robotics*, 33(6): 1292–1312. DOI: [10.1109/TRO.2017.2723903](https://doi.org/10.1109/TRO.2017.2723903). 436
- P. J Wagner, H. Wortzel, K. Frey, C. Alan Anderson, and D. Arciniegas. 2011. Clock-drawing performance predicts inpatient rehabilitation outcomes after traumatic brain injury. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 23: 449–53. DOI: [10.1176/jnp.23.4.jnp449](https://doi.org/10.1176/jnp.23.4.jnp449). 453
- M. Jacob, Y.-T. Li, G. Akingba, and J. P. Wachs. 2012. Gestonurse: a robotic surgical nurse for handling surgical instruments in the operating room. *Journal of Robotic Surgery*, 6(1): 53–63. DOI: [10.1007/s11701-011-0325-0](https://doi.org/10.1007/s11701-011-0325-0). 432
- M. G. Jacob and J. P. Wachs. January 2014. Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters*, 36: 196–203. DOI: [10.1016/j.patrec.2013.05.024](https://doi.org/10.1016/j.patrec.2013.05.024). 432
- E. Kalbe, J. Kessler, P. Calabrese, R. Smith, A. P. Passmore, M. Brand, and R. Bullock. February 2004. DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *International Journal of Geriatric Psychiatry*, 19(2): 136–143. DOI: [10.1002/gps.1042](https://doi.org/10.1002/gps.1042). 451
- G. Keren, A. E.-D. Mousa, O. Pietquin, S. Zafeiriou, and B. Schuller. 2018. Deep learning for multisensorial and multimodal interaction. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of*

- Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107996](https://doi.org/10.1145/3107990.3107996). 427, 772
- H. Kurniawan, A. V. Maslov, and M. Pechenizkiy. June 2013. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pp. 209–214. DOI: [10.1109/CBMS.2013.6627790](https://doi.org/10.1109/CBMS.2013.6627790). 454
- A. Lally, S. Bagchi, M. Barborak, D. W. Buchanan, J. Chu-Carroll, D. A. Ferrucci, M. R. Glass, A. Kalyanpur, E. T. Mueller, J. W. Murdock, S. Patwardhan, and J. M. Prager. 2017. Watsonpaths: Scenario-based question answering and inference over unstructured information. *AI Magazine*, 38(2): 59–76. <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2715>. 434
- C. P. Langlotz. 2006. Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26: 1595–1597. <http://radiographics.rsnajnl.org/cgi/content/full/26/6/1595>. DOI: [10.1148/radiographics.266065168](https://doi.org/10.1148/radiographics.266065168). 440
- I. Lee, O. Sokolsky, S. Chen, J. Hatcliff, E. Jee, B. Kim, A. King, M. Mullen-Fortino, S. Park, A. Roederer, and K. Venkatasubramanian. January 2012. Challenges and research directions in medical cyber-physical systems. In *Proceedings of the IEEE*, 100(1): 75–90. DOI: [10.1109/JPROC.2011.2165270](https://doi.org/10.1109/JPROC.2011.2165270). 429
- C. Lisetti, R. Amini, and U. Yasavur. 2015. Now all together: Overview of virtual health assistants emulating face-to-face health interview experience. *KI - Künstliche Intelligenz*, 29(2): 161–172. DOI: [10.1007/s13218-015-0357-0](https://doi.org/10.1007/s13218-015-0357-0). 435
- S. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. Schramek. 2007. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition*, 65(3): 209–237. <http://www.sciencedirect.com/science/article/pii/S0278262607000322>. DOI: <https://doi.org/10.1016/j.bandc.2007.02.007>. 449
- A. Luxenburger, A. Prange, M. M. Moniri, and D. Sonntag. 2016. Medicalvr: Towards medical remote collaboration using virtual reality. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pp. 321–324. ACM, New York. DOI: [10.1145/2968219.2971392](https://doi.org/10.1145/2968219.2971392). 460
- S. Luz and B. Kane. 2009. Classification of patient case discussions through analysis of vocalisation graphs. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pp. 107–114. ACM, New York. DOI: [10.1145/1647314.1647334](https://doi.org/10.1145/1647314.1647334). 432
- J.-C. Martin, C. Clavel, M. Courgeon, M. Ammi, M.-A. Amorim, Y. Tsalamllal, and Y. Gaffary. 2018. How do users perceive multimodal expressions of affects? In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108001](https://doi.org/10.1145/3107990.3108001). 427, 764
- U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. 2006. Activity recognition and monitoring using multiple sensors on different body positions. In *International*

*Workshop on Wearable and Implantable Body Sensor Networks (BSN'06).* DOI: [10.1109/BSN.2006.6](https://doi.org/10.1109/BSN.2006.6). 436

- K. R. McKeown, S.-F. Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivas-siloglou, S. Johnson, D. A. Jordan, J. L. Klavans, A. Kushniruk, V. Patel, and S. Teufel. 2001. Persival, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '01, pp. 331–340. ACM, New York. DOI: [10.1145/379437.379722](https://doi.org/10.1145/379437.379722). 432
- MedicalCPS, 2018. Medical cyber-physical systems. <http://dfki.de/MedicalCPS/>. 2018-08-22. 427, 778
- G. Mehlmann, K. Janowski, and E. André. 2016. Modeling grounding for interactive social companions. *KI-Künstliche Intelligenz*, 30(1): 45–52. DOI: [10.1007/s13218-015-0397-5](https://doi.org/10.1007/s13218-015-0397-5). 437
- J. L. Mejino, D. L. Rubin, and J. F. Brinkley. 2008. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In *Proceedings of the AMIA Symposium*, pp. 465–469. <http://stanford.edu/~rubin/pubs/097.pdf>. 439
- M. Möller, S. Regel, and M. Sintek. June 2009. Radsem: Semantic annotation and retrieval for medical images. In *Proceedings of the 6th Annual European Semantic Web Conference (ESWC2009)*. <http://www.manuelm.org/publications/wp-content/uploads/2009/02/eswc2009.pdf>. 439
- M. Möller, M. Sintek, R. Biedert, P. Ernst, A. Dengel, and D. Sonntag. 2010. Representing the international classification of diseases version 10 in OWL. In J. Filipe and J. L. G. Dietz, editors, *KEOD 2010 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, Valencia, Spain, October 25–28, 2010, pp. 50–59. SciTePress. DOI: [10.1007/978-3-642-29764-9\\_16](https://doi.org/10.1007/978-3-642-29764-9_16). 440
- N. Moraveji and C. Soesanto. 2012. Towards stress-less user interfaces: 10 design heuristics based on the psychophysiology of stress. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pp. 1643–1648. ACM, New York. DOI: [10.1145/2212776.2223686](https://doi.org/10.1145/2212776.2223686). 448, 449, 452
- J. Morris, R. Mohs, H. Rogers, G. Fillenbaum, and A. Heyman. 1988. Consortium to establish a registry for Alzheimer's disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacol Bulletin*, 24(4): 641–52. 447, 451
- A. Mourão and F. Martins. 2013. Novamedsearch: A multimodal search engine for medical case-based retrieval. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pp. 223–224. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, Paris, France. <http://dl.acm.org/citation.cfm?id=2491748.2491798>. 432
- S. Munir, J. A. Stankovic, C.-J. M. Liang, and S. Lin. 2013. Cyber physical system challenges for human-in-the-loop control. emphPresented as part of the 8th International

- Workshop on Feedback Computing. USENIX, Berkeley, CA. <https://www.usenix.org/conference/feedbackcomputing13/workshop-program/presentation/Munir>. 430
- Myscript, 2018. Myscript SDK product. <https://www.myscript.com>. 2018-11-05. 457
- Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4): 695–699. DOI: [10.1111/j.1532-5415.2005.53221.x](https://doi.org/10.1111/j.1532-5415.2005.53221.x). 451
- Neosmartpen. 2018. Neo N2 smartpen product. <https://www.neosmartpen.com>. 2018-11-05. 456
- J. Nielsen and R. L. Mack, editors. 1994. *Usability Inspection Methods*. John Wiley & Sons, Inc., New York. 448
- M. Niemann, A. Prange, and D. Sonntag. 2018a. Towards a multimodal multisensory cognitive assessment framework. In J. Hollmén, C. McGregor, P. Soda, and B. Kane, editors, *31st IEEE International Symposium on Computer-Based Medical Systems, CBMS 2018*, Karlstad, Sweden, June 18–21, 2018, pp. 24–29. IEEE Computer Society. DOI: [10.1109/CBMS.2018.00012](https://doi.org/10.1109/CBMS.2018.00012). 454, 457
- M. Niemann, A. Prange, and D. Sonntag. 2018b. Towards a multimodal multisensory cognitive assessment framework. *IEEE 31th International Symposium on Computer-Based Medical Systems (CBMS)*. DOI: [10.1109/CBMS.2018.00012](https://doi.org/10.1109/CBMS.2018.00012). 437
- Oculus. 2018. Oculus Rift product. <https://www.oculus.com>. 2018-11-05. 460
- S. Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '99*, pp. 576–583. ACM, New York. DOI: [10.1145/302979.303163](https://doi.org/10.1145/302979.303163). 432
- S. Oviatt and P. Cohen. March 2000. Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3): 45–53. DOI: [10.1145/330534.330538](https://doi.org/10.1145/330534.330538). 431
- S. Oviatt and P. R. Cohen. 2015. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. Morgan & Claypool Publishers, San Rafael, CA. <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7087855>. DOI: [10.2200/S00636ED1V01Y201503HCI030](https://doi.org/10.2200/S00636ED1V01Y201503HCI030). 434
- S. Oviatt, J. F. Grafsgaard, L. Chen, and X. Ochoa. 2018a. Multimodal learning analytics: Assessing learners' mental state during the process of learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108003](https://doi.org/10.1145/3107990.3108003). 427, 764
- S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors. 2018b. *The Handbook of Multimodal-Multisensor Interfaces, Volume 2. Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 430

- S. L. Oviatt, K. Hang, J. Zhou, K. Yu, and F. Chen. 2018c. Dynamic handwriting signal features predict domain expertise. *TiiS*, 8(3): 18:1–18:21. DOI: [10.1145/3213309](https://doi.org/10.1145/3213309). 453
- Y. Panagakis, O. Rudovic, and M. Pantic. 2018. Learning for multimodal and context-sensitive interfaces. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107995](https://doi.org/10.1145/3107990.3107995). 427, 772
- A. Pecchinenda. 1996. The affective significance of skin conductance activity during a difficult problem-solving task. *Cognition & Emotion*, 10(5): 481–504. 454
- R. W. Picard, M. Migliorini, C. Caborni, F. Onorati, G. Regalia, D. Friedman, and O. Devinsky. 2017. Wrist sensor reveals sympathetic hyperactivity and hypoventilation before probable sudep. *Neurology* August 8, 2017; 89 (6). <http://n.neurology.org/content/89/6/633>. DOI: [10.1212/WNL.0000000000004208](https://doi.org/10.1212/WNL.0000000000004208). 437
- A. Prange, I. P. Sandrala, M. Weber, and D. Sonntag. 2015. Robot companions and smartpens for improved social communication of dementia patients. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, pp. 65–68. ACM, New York. DOI: [10.1145/2732158.2732174](https://doi.org/10.1145/2732158.2732174). 446
- A. Prange, M. Barz, and D. Sonntag. 2018a. Medical 3d images in multimodal virtual reality. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, IUI'18, pp. 19:1–19:2. ACM, New York. DOI: [10.1145/3180308.3180327](https://doi.org/10.1145/3180308.3180327). 460, 461
- A. Prange, M. Barz, and D. Sonntag, 2018b. A categorisation and implementation of digital pen features for behaviour characterisation. DFKI Technical Report. 455
- RadSpeech. 2011. RadSpeech Project. <https://www.dfki.de/RadSpeech>. 2018-10-31. 445
- R. Reitan. 1992. *Trail Making Test*. Reitan Neuropsychology Laboratory. 451
- S. Rick, A. Calvitti, Z. Agha, and N. Weibel. 2015. Eyes on the clinic: Accelerating meaningful interface analysis through unobtrusive eye tracking. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth '15*, pp. 213–216. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium. <http://dl.acm.org/citation.cfm?id=2826165.2826197>. DOI: [10.4108/icst.pervasivehealth.2015.259276](https://doi.org/10.4108/icst.pervasivehealth.2015.259276). 432
- A. Rizzo and T. Talbot. January 2016. Virtual reality standardized patients for clinical training. In *The Digital Patient*, pp. 255–272. John Wiley & Sons, Inc, Hoboken, NJ. DOI: [10.1002/9781118952788.ch18](https://doi.org/10.1002/9781118952788.ch18). 460
- C. Saitis and K. Kalimeri. 2016. Identifying urban mobility challenges for the visually impaired with mobile monitoring of multimodal biosignals. *International Conference on Universal Access in Human-Computer Interaction*, pp. 616–627. Springer. DOI: [10.1007/978-3-319-40238-3\\_59](https://doi.org/10.1007/978-3-319-40238-3_59). 454
- M. Samwald, A. Jentzsch, C. Bouton, C. Kallesøe, E. L. Willighagen, J. Hajagos, M. S. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens. 2011. Linked open

- drug data for pharmaceutical research and development. *Journal of Cheminformatics*, 3: 19. DOI: [10.1186/1758-2946-3-19](https://doi.org/10.1186/1758-2946-3-19). 431
- Y. Sawamoto, Y. Koyama, Y. Hirano, S. Kajita, K. Mase, K. Katsuyama, and K. Yamauchi. 2007. Extraction of important interactions in medical interviews using nonverbal information. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, pp. 82–85. ACM, New York. DOI: [10.1145/1322192.1322209](https://doi.org/10.1145/1322192.1322209). 435
- S. Scherer, G. Stratou, and L.-P. Morency. 2013. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 135–140. ACM, New York. DOI: [10.1145/2522848.2522886](https://doi.org/10.1145/2522848.2522886). 437
- A. Schroter, R. Mergl, K. Burger, H. Hampel, H. J. Moller, and U. Hegerl. 2003. Kinematic analysis of handwriting movements in patients with Alzheimer's disease, mild cognitive impairment, depression and healthy subjects. *Dementia and Geriatric Cognitive Disorders*, 15(3): 132–142. DOI: [10.1159/000068484](https://doi.org/10.1159/000068484). 458
- B. Schuller. 2018. Multimodal user state & trait recognition: An overview. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107997](https://doi.org/10.1145/3107990.3107997). 427, 437, 764
- Sesame. 2017. RDF4J sesame. <http://rdf4j.org>. 2017-05-22. 431, 443
- A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. W. Kim. 2016. Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine*, 8(337): 337ra64–337ra64. <http://stm.sciencemag.org/content/8/337/337ra64>. DOI: [10.1126/scitranslmed.aad9398](https://doi.org/10.1126/scitranslmed.aad9398). 433
- H. P. D. Silva, S. Fairclough, A. Holzinger, R. Jacob, and D. Tan. January 2015. Introduction to the special issue on physiological computing for human-computer interaction. *ACM Transactions on Computer-Human Interaction*, 21(6): 29:1–29:4. DOI: [10.1145/2688203](https://doi.org/10.1145/2688203). 427, 764
- Simsensei. 2014. Project description. <http://ict.usc.edu/prototypes/simsensei/>. 2017-01-16. 459
- D. Sonntag. 2015. Kognit: Intelligent cognitive enhancement technology by cognitive models and mixed reality for dementia patients. In *Proceedings of the AAAI Fall Symposium Series*. <http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11702>. 446, 459
- D. Sonntag. 2016. Medical cyber-physical systems. In *Cyber-Physical System Design with Sensor Networking Technologies*, Control, Robotics and Sensors, pp. 311–333. Institution of Engineering and Technology. 426, 429
- D. Sonntag. 2017. Interakt - A multimodal multisensory interactive cognitive assessment tool. *CoRR*, abs/1709.01796. <http://arxiv.org/abs/1709.01796>. 446
- D. Sonntag and M. Möller. 2010. A multimodal dialogue mashup for medical image semantics. In *Proceedings of the 15th International Conference on Intelligent User*

*Interfaces*, IUI '10, pp. 381–384. ACM, New York. DOI: [10.1145/1719970.1720036](https://doi.org/10.1145/1719970.1720036).  
[432](#), [438](#), [441](#)

- D. Sonntag and D. Porta. 2014. Intelligent semantic mediation, knowledge acquisition and user interaction. In W. Wahlster, H. Grallert, S. Wess, H. Friedrich, and T. Widenka, editors. 2014. Towards the Internet of Services: The THESEUS Research Program. *Cognitive Technologies*. pp. 179–189. Springer. DOI: [10.1007/978-3-319-06755-1\\_14](https://doi.org/10.1007/978-3-319-06755-1_14).  
[429](#)
- D. Sonntag, P. Wennerberg, P. Buitelaar, and S. Zillner. 2009. Pillars of ontology treatment in the medical domain. *Journal of Cases on Information Technology*, 11(4): 47–73. [431](#), [461](#)
- D. Sonntag, N. Reithinger, G. Herzog, and T. Becker. 2010a. A discourse and dialogue infrastructure for industrial dissemination. In *Proceedings of the Second International Conference on Spoken Dialogue Systems for Ambient Environments*, IWSDS'10, pp. 132–143. Springer-Verlag, Berlin, Heidelberg. <http://dl.acm.org/citation.cfm?id=1925948.1925961>. DOI: [10.1007/978-3-642-16202-2\\_12](https://doi.org/10.1007/978-3-642-16202-2_12).  
[431](#)
- D. Sonntag, C. Weihrauch, O. Jacobs, and D. Porta. 2010b. Theseus ctc-wp4 usability guidelines for use case applications. Technical report, Bundesministerium für Wirtschaft und Technologie.  
[448](#)
- D. Sonntag, C. Schulz, C. Reuschling, and L. Galarraga. 2012. Radspeech's mobile dialogue system for radiologists. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pp. 317–318. ACM, New York. DOI: [10.1145/2166966.2167031](https://doi.org/10.1145/2166966.2167031).  
[432](#), [438](#)
- D. Sonntag, M. Weber, A. Cavallaro, and M. Hammon. 2014a. Integrating digital pens in breast imaging for instant knowledge acquisition. *AI Magazine*, 35(1): 26–37. [445](#), [457](#)
- D. Sonntag, S. Zillner, P. Ernst, C. Schulz, M. Sintek, and P. Dankerl. 2014b. Mobile radiology interaction and decision support systems of the future. In W. Wahlster, H. Grallert, S. Wess, H. Friedrich, and T. Widenka, editors. 2014. Towards the Internet of Services: The THESEUS Research Program. *Cognitive Technologies*, pp. 371–382. Springer. DOI: [10.1007/978-3-319-06755-1\\_28](https://doi.org/10.1007/978-3-319-06755-1_28).  
[429](#)
- D. Sonntag, V. Tresp, S. Zillner, A. Cavallaro, M. Hammon, A. Reis, A. P. Fasching, M. Sedlmayr, T. Ganslandt, H.-U. Prokosch, K. Budde, D. Schmidt, C. Hinrichs, T. Wittenberg, P. Daumke, and G. P. Oppelt. 2015. The clinical data intelligence project. *Informatik-Spektrum*, pp. 1–11. DOI: [10.1007/s00287-015-0913-x](https://doi.org/10.1007/s00287-015-0913-x).  
[426](#), [446](#)
- W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, C. C. Price, M. Lamar, and D. L. Penney. March 2016. Learning classification models of cognitive conditions from subtle behaviors in the digital Clock Drawing Test. *Machine Learning*, 102(3): 393–441. DOI: [10.1007/s10994-015-5529-5](https://doi.org/10.1007/s10994-015-5529-5).  
[446](#)
- J. C. Sriram, M. Shin, T. Choudhury, and D. Kotz. 2009. Activity-aware ecg-based patient authentication for remote health monitoring. In *Proceedings of the 2009 International*

- Conference on Multimodal Interfaces*, ICMI-MLMI '09, pp. 297–304. ACM, New York. DOI: [10.1145/1647314.1647378](https://doi.org/10.1145/1647314.1647378). 437
- Taurus. 2017. Taurus. <https://www.youtube.com/watch?v=cqBm97jBvuY>. 2015-05-06. 433
- A. Tobergte, P. Helmer, U. Hagn, P. Rouiller, S. Thielmann, S. Grange, A. Albu-Schäffer, F. Conti, and G. Hirzinger. 2011. The sigma.7 haptic interface for mirosurge: A new bi-manual surgical console. In *IROS*, pp. 3023–3030. IEEE. <http://dblp.uni-trier.de/db/conf/iros/iros2011.html#TobergteHHRTGACH11>. DOI: [10.1109/IROS.2011.6094433](https://doi.org/10.1109/IROS.2011.6094433). 433
- TUG. 2017. Tug robots in healthcare. <http://www.aethon.com/tug/tughealthcare/>. 2017-05-22. 433
- G. Turchetti, I. Palla, F. Pierotti, and A. Cuschieri. 2012. Economic evaluation of da vinci-assisted robotic surgery: a systematic review. *Surgical Endoscopy*, 26(3): 598–606. ISSN 1432-2218. DOI: [10.1007/s00464-011-1936-2](https://doi.org/10.1007/s00464-011-1936-2). 433
- M. Valstar. 2014. Automatic behaviour understanding in medicine. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, RFMIR '14, pp. 57–60. ACM, New York. DOI: [10.1145/2666253.2666260](https://doi.org/10.1145/2666253.2666260). 436
- Verbsurgical. 2017. Verbsurgical. <http://www.verbsurgical.com>. 2015-05-22. 433
- J. Wagner and E. André. 2018. Real-time sensing of affect and social signals in a multimodal framework: a practical approach. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108000](https://doi.org/10.1145/3107990.3108000). 427, 764
- W. Wahlster, H. Grallert, S. Wess, H. Friedrich, and T. Widenka. 2014. *Towards the Internet of Services: The THESEUS Research Program*. Cognitive Technologies. Springer.
- T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. 2008. Aligning temporal data by sentinel events: Discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 457–466. ACM, New York. DOI: [10.1145/1357054.1357129](https://doi.org/10.1145/1357054.1357129). 425
- N. Weibel, C. Emmenegger, J. Lyons, R. Dixit, L. L. Hill, and J. D. Hollan. 2013. Interpreter-mediated physician-patient communication: Opportunities for multimodal healthcare interfaces. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '13, pp. 113–120. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium. DOI: [10.4108/icst.pervasivehealth.2013.252026](https://doi.org/10.4108/icst.pervasivehealth.2013.252026). 435
- G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber. 2016. Smartwatch-based activity recognition: A machine learning approach. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 426–429. DOI: [10.1109/BHI.2016.7455925](https://doi.org/10.1109/BHI.2016.7455925). 436

- P. Werner, S. Rosenblum, G. Bar-On, J. Heinik, and A. Korczyn. 2006. Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 61(4): P228–P236. [458](#)
- K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. 2011. Lifeflow: Visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 1747–1756. ACM, New York. DOI: [10.1145/1978942.1979196. 425](#)
- A. D. Wood and J. A. Stankovic. January 2008. Human in the loop: Distributed data streams for immersive cyber-physical systems. *SIGBED Review*, 5(1): 20:1–20:2. DOI: [10.1145/1366283.1366303. 430](#)
- J. Zhai and A. Barreto. 2006. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. DOI: [10.1109/IEMBS.2006.259421. 454](#)
- J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad. 2018. Multimodal behavioural and physiological signals as indicators of cognitive load. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3108002. 427, 764](#)

# 12

# Automotive Multimodal Human-Machine Interface

Dirk Schnelle-Walka, Stefan Radomski

## 12.1

### Introduction

The majority of user interfaces in the automotive domain were not developed as the result of user-centered design iterations, but are a compromise of continuously balancing usability and established expectations within a conservative product development cycle. However, users are already interacting *multimodally*, e.g. with their smartphones, and expectations with regard to natural interaction have increased dramatically in the past years. Moreover, users have started to project these expectations towards all kind of interfaces encountered in their daily lives. Currently, these expectations are not yet fully met by car manufacturers. While there are some showcases that realize smart, natural interaction, a widespread deployment on any meaningful scale is still missing. In this chapter we describe some ongoing efforts for multimodal interfaces applicable in an automotive environment.

This chapter is organized as follows. First, we will describe the evolution of human-machine interfaces (HMIs) in cars. After discussing challenges and opportunities resulting from this evolution, we will describe today's efforts to add support for additional modalities before we conclude the chapter with a discussion and outlook.

## 12.2

### HMI Evolution

As outlined in the blog of ustwo<sup>1</sup> about future HMI trends in cars, the biggest innovations with regard to interaction in today's cars are related to the way in which

---

1. <https://ustwo.com/blog/hmi-where-we-are-now/>

### Glossary

**Cognitive load** expresses a user's mental load and mental effort to solve a given problem.

Mental load is imposed by problem's parameters such as task and structure of sequence of information. Mental effort refers to the amount of capacity that is allocated to the problem's demands. [Sweller 1988, Paas 1992]

An **infotainment system** combines entertainment and information in cars for drivers and passengers. It originated in car audio systems but now also includes navigation systems, telephony, and comparable functionality.

The **interaction space** is restricted to the area that the driver can reach with his hands or with the input devices. The design goal is to position the controls such that interaction can be performed comfortably, i.e. the driver is seated behind the wheel. The space encloses the windshield, the dashboard, center stack, steering wheel, floor, and periphery.

**Mental models** are a pervasive property of humans. People form internal, mental models of themselves and of the things and people with whom they interact. These models provide predictive and explanatory power for understanding the interaction. Mental models evolve naturally through interaction with the world and with the particular system under consideration. [Norman 1986]

**Modality.** A modality identifies a perceptual mode of a human or a comparable concept simulated by a computer.

**Multimodal.** A human-computer interface is multimodal if the computer supports two or more perceptual modes (1) to decode information relevant to the interaction and (2) to encode its response.

**Muscle memory** is a form of procedural memory that involves consolidating a specific motor task into memory through repetition. When a movement is repeated over time, a long-term muscle memory is created for that task, eventually allowing it to be performed without conscious effort.

**Natural interaction** is a more intuitive way of interacting with a computing device. Recent advancements in HCI have facilitated this kind of interaction and its development is expected to make it easy for users to learn how to use the interface in the quickest possible way. [Villaroman et al. 2011]

A **nomadic device** is a device for information including entertainment, and/or communication that can be used outside of the vehicle and inside the vehicle while driving. It is not supplied or installed by the vehicle manufacturer. [Kulmala and Mäuerer 2005]

**Glossary** (*continued*)

**Pervasive distributed computing** promises a computing infrastructure that seamlessly and ubiquitously aids users in accomplishing their tasks and that renders the actual computing devices and technology largely invisible. The basic idea behind pervasive computing is to deploy a wide variety of smart devices throughout our working and living spaces. These devices coordinate with each other to provide users with universal and immediate access to information and support users in completing their tasks. [Grimm et al. 2001]

A **spoken dialog system** (SDS) converses with the user via voice. It consists of a number of components to function successfully: Speech recognition to recognize the words a user said, natural language understanding to assign meaning to these words, a dialog manager to decide how the utterance fits into the dialog so far and decide which actions to perform next, external communication to access external data, a response generation to choose the words and phrases to be used in a response and a speech synthesizer to actually speak the response. [McTear 2004]

**Tangible user interfaces** are concerned with providing tangible representations to digital information and controls, allowing users to quite literally grasp data with their hands. Serving as direct, tangible representations of digital information, these augmented physical objects often function as both input and output devices providing users with parallel feedback loops: physical, passive haptic feedback that informs users that a certain physical manipulation is complete; and digital, visual or auditory feedback that informs users of the computational interpretation of their action. [Shaer and Hornecker 2010]

In **Wizard-of-Oz** studies subjects are told that they are interacting with a computer system, although in fact they are not. Instead, the interaction is mediated by a human operator, the wizard, with the consequence that the subject can be given more freedom of expression, or be constrained in more systematic ways, than is the case for existing implementations.

we control the integrated devices, i.e., the **infotainment system**. This evolution over the past few years has introduced a new level of complexity to the interaction.

Around 1970, the only interaction offered by cars was direct control and manipulation of buttons, switches, and sliders (see Figure 12.1).

The haptic and spatial arrangement allowed users to develop a mental mapping, however arbitrary, of the user interface by relying on their **muscle memory**. While the sparseness of controls promised simplicity, and these mechanical elements offered all the aforementioned advantages of tangible controls, it still required some effort to explore and learn the available **interaction space**.



**Figure 12.1** Interior of a Mercedes in 1970. Source <https://ustwo.com/blog/hmi-where-we-are-now/>, accessed 06/14/2016.

Things changed significantly with the introduction of computers controlling the HMI, as shown in Figure 12.2.

There are considerably more controls for far more functionalities available. While some interactions are still available via *tangible interfaces*, they are accompanied by a graphical user interface (GUI), which can adapt to the current task and present the information required. The GUI takes over the role as the primary source of feedback, decoupling user in- and output and leading to indirect control. Sometimes the GUI can also be controlled by a mouse-like device (a *knob*) and more recent systems also allow for a direct manipulation by touch.

This leads to additional effort for the user to establish a *mental model*. With the plethora of available modalities requiring multiple mental maps, learnability decreased considerably and distraction increased. In the following section, we will discuss the challenges and opportunities of multimodal interaction to help in reducing the *cognitive load* and increase learnability again.



**Figure 12.2** Interior of a Mercedes SL. Source <https://ustwo.com/blog/hmi-where-we-are-now/>, accessed 06/14/2016.

## 12.3

### Challenges and Opportunities

In a meta-analysis of respective research, Radomski [2016] identifies two major benefits of multimodal interfaces:

1. combining many error-prone modalities increases overall accuracy; and
2. providing different means of interaction reduces cognitive load.

Accuracy can be improved by using those modalities that best express a certain input. For instance, voice is not very precise in describing a location, while deictic gestures can be more easily resolved to spatial references [Oviatt 2003].

The term *cognitive load* was first coined by John Sweller to express a user's mental effort to solve a given problem [Sweller 1988] and can be decomposed into the following three categories:

1. Intrinsic cognitive load as the complexity inherent to the information presented with the current task at hand.

2. Extraneous cognitive load as induced by the manner of presentation, e.g., when unsuited or hard-to-use interfaces divert mental resources from the actual task.
3. Germane cognitive load, which is also induced by the manner of presentation, but experienced when the interfaces are supportive and help the user in establishing a suitable mental model.

In order to account for the subjective cognitive load perceived, Sweller introduced the human working memory as a normalization factor via the *central executive* introduced by Baddeley and Hitch in their model of the human working memory:

$$\text{Cognitive Load} := \frac{\text{CL}_{\text{intrinsic}} + \text{CL}_{\text{extraneous}} - \text{CL}_{\text{germane}}}{\text{Human Working Memory}}$$

Hence, cognitive load can be reduced by: (1) suitable presentations and interaction paradigms that reduce the extraneous cognitive load and decrease germane cognitive load or (2) by increasing the human working memory. For the latter, various modalities can be combined beneficially as some modalities are assumed not to share memory according to Baddeley and Hitch's model, which was verified in experiments conducted by [Oviatt \[2006\]](#). While the formulae suggests that a reduction of intrinsic cognitive load could also reduce the overall cognitive load, this component is inherent to the given task and thus cannot be modified.

In the car, the driver is dealing with tasks that [Salvucci \[2001\]](#) divides into

- primary tasks, involved in maneuvering (e.g., turning the steering wheel and operating the pedals);
- secondary tasks, involved in maintaining safety (e.g., turn signals, windshield wipers); and
- tertiary tasks, involving all other comfort, information, and entertainment functions.

Secondary and tertiary tasks compete for the drivers attention with the primary task of driving [[Normark et al. 2009](#)], thus intrinsically increasing cognitive load. Furthermore, operating third-party *nomadic devices*, such as smartphones, imposes yet an additional load onto the driver. Especially for safety reasons, drivers must focus their attention to what is outside of the car rather than on in-vehicle displays and controls, but without limiting the inspection and interaction possibilities of a driver with its car [[Baber and Wankling 1992](#)]. Focusing on tertiary tasks is also referred to as *driver distraction*, that is “diversion of attention away from activities critical for safe driving toward a competing activity” [[Lee et al. 2008](#)]. The 100-car

study from Neale et al. showed that distraction played a major role in car crashes or near-crashes [Neale et al. 2005] and the Virginia Technology Transportation Institute revealed, more precisely, that distraction within three seconds prior to an incident is one of the main causes of accidents [Pickering et al. 2007].

In order to limit distraction, the US National Highway Traffic Safety Administration (NHTSA) recommends to limit each interaction with the infotainment system requiring visual focus to less than 2 s and the overall interaction cycle to 12 s [NHSTA 2010]. With the same intention, the Commission of the European communities refined a set of guidelines, first published in 1999 to improve safe and efficient in-vehicle information interaction, in [CEC 2007]. These guidelines are based on the following five design goals.

1. The system supports the driver and does not give rise to potentially hazardous behavior by the driver or other road users.
2. The allocation of driver attention while interacting with system displays and controls remains compatible with the attentional demand of the driving situation.
3. The system does not distract or visually entertain the driver.
4. The system does not present information to the driver, which results in potentially hazardous behavior by the driver or other road users.
5. Interfaces with systems intended to be used in combination by the driver while the vehicle is in motion are consistent and compatible.

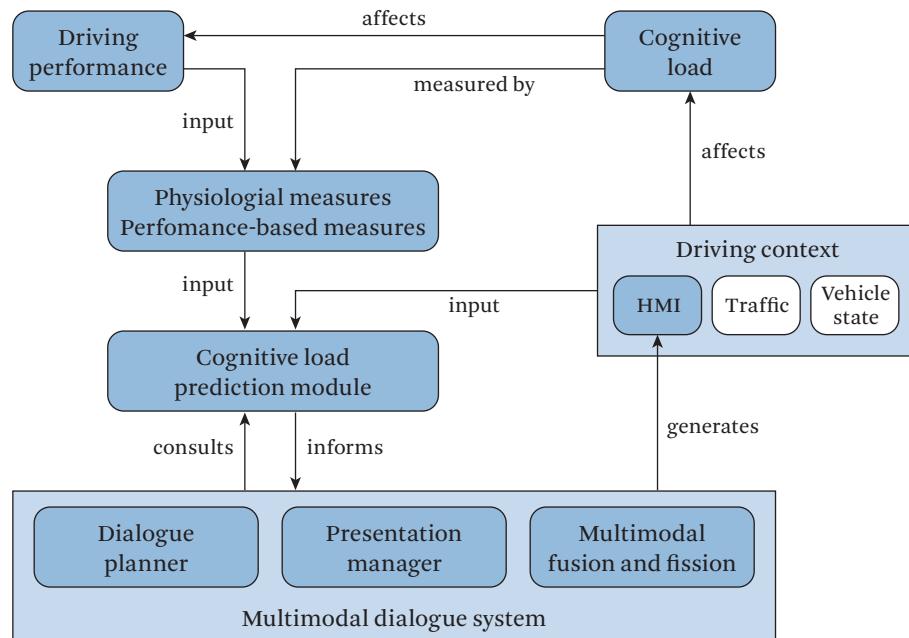
These guidelines do not propose precise durations as does the NHTSA but they do also recommend limiting glance-away times and overall interaction cycles. However, this unquantified “timeliness” is hard to measure and validate.

In order to perform such a validation, Neßelrath and Feld [2013] identified the relevant components that have an influence on cognitive load and driving performance and their interrelation in a multimodal dialog system in cars. An overview of the components they identified and how they influence each other is shown in Figure 12.3.

Their model also includes task inherent difficulties, e.g., finding a radio station broadcasting a favorite sporting event.

Neßelrath and Feld [2013] name the following categories for measures for cognitive load:

- subjective measures for the subjective workload, e.g. by a questionnaire;
- physiological measures for cognitive stress reflected in the human physiology, e.g. by heart rate;



**Figure 12.3** Components influencing cognitive load and driving performance of an in-car multimodal dialog system (after [Neßelrath and Feld 2013]).

- performance measures for the effects of cognitive load onto task performance, e.g. dialog processing performance or driving performance; and
- behavioral measures for changes in interaction behavior, e.g., changes of pitch and speech rate under high cognitive load.

Subjective measures, like the subject's cognitive load, are not always applicable for real-time assessment. Due to their invasive nature, physiological measures are used under lab conditions, e.g., physiological sensors for heart rate, brain activity, galvanic skin response and eye activity employed in laboratory tests are inapplicable in a real driving situation. The correlation between physiological measure like heart rate, skin conductance and respiration rate and “changes in workload before the appearance of clear decrements in driving performance” was shown, e.g., by [Mehler et al. \[2009\]](#) and [\[Palinko et al. 2010\]](#) estimated the cognitive load based on pupillometry, where the Mean Pupil Diameter Change (MPDC) was shown to correlate to driving performance [\[Palinko et al. \[2010\]\]](#). However, the models required to infer cognitive load from raw sensor data is still an open research issue. For instance

changing illumination while driving severely affects the performance of pupillometry. Consequently, most of today's studies start with subjective and performance measures to rate the driver's distraction and driving performance [Nefelath and Feld 2013]. For instance, recent studies, like the Utah study [Strayer et al. 2013], employ a device that sits on one's head and turns on a light in the driver's peripheral vision. Then, the driver needs to react by pushing a button on the steering wheel. Reaction times are measured.

Other aspects that need to be considered are mentioned in Nielsen's criticism.<sup>2</sup> of the BMW iDrive<sup>3</sup> He bemoans, among other things, the following usability defects.

- **Response times are incredibly slow.**<sup>4</sup>. Users should not be forced to wait until a subsequent screen shows up. Slow response times require an increased attention to operate the user interface (UI). A review of the 2017's X1 model measured response times exceeding 1 s.
- **Clumsy task flows.** Users should not be forced to take additional steps to achieve their goal. The system should be supportive by offering the most likely follow-up tasks while others are left as exceptional options.
- **Misleading mapping between input device and screen.** Be consistent in the way the UI is operated.
- **Obscure abbreviations.** Avoid abbreviations that the user may not be familiar with and spell out commands when there is sufficient space available.
- **Lack of situational awareness.** When selecting from a list rank those options higher that can be derived from the context.

Long response times are not a unique property of embedded HMIs but are well known in other domains. However, the limiting factors like storage size and available computing power in the car are steadily increasing and the situation may relax in the coming years.

An alternative way of overcoming these restrictions is using off-board technology, i.e. server-based ASR and TTS for voice-based interfaces. This has become a

---

2. <https://www.nngroup.com/articles/why-consumer-products-have-bad-ux/>

3. <http://www.bmwusa.com/Standard/Content/Innovations/Engineering/iDrive/Overview/default.aspx>

4. <http://www.caranddriver.com/reviews/2017-bmw-x1-in-depth-model-review-2017-bmw-x1-idrive-infotainment-review-car-and-driver-page-6>

realistic option since the automotive market is clearly preparing for the car to be “connected”, i.e. to have internet access, as a common feature.<sup>5</sup>

Off-board systems do not submit to tight restrictions in computing power and complexity. Hence, they allow for new functions and services in in-car applications like message dictation and POI search (points of interest). Furthermore, it is a general advantage of cloud-based services that they are easy to maintain, update, and adapt to fit new conditions or provide so-called over-the-air updates. This is of great value, especially in the automotive domain for which long lifecycles are a characteristic. However, the increased response times for results delivered from the cloud and the problems with only intermittent high-speed mobile internet connectivity relativize these benefits.

In the following, we will discuss cognitive load as well as usability for several modalities that can be found in today’s cars or in research projects.

## 12.4

### Multimodal In-Car Interaction

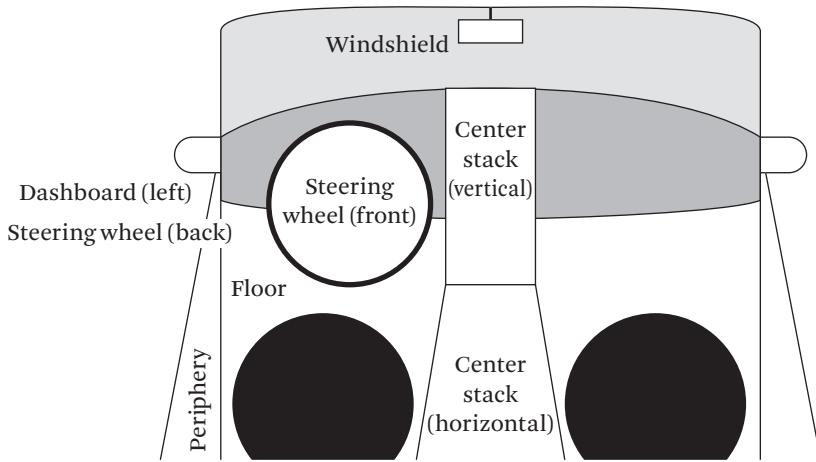
This section introduces some trends to add multimodal interaction in today’s dashboards. We will discuss haptic controllers, touch screens, 3D gestures, voice, secondary displays and gaze. In general, for many of those modalities, the question of *where* to interact within the confines of a car arises. To this effect, [Kern and Schmidt 2009] identified the following interaction areas in a car, shown in Figure 12.4:

- windshield: used for example for head-up displays;
- dashboard: for driver-based user interfaces we focus on the left part of the dashboard that is directly in front of the driver in left-hand cars;
- center Stack: divided into the vertical part (on the right side of “dashboard left” in front of the driver) and the horizontal part (between the front seats);
- steering wheel: divided into front and back side of the steering wheel;
- floor: virtually exclusively used to interact with gas, brake, and clutch pedals; and
- periphery: includes the side-/rear-view mirrors up.

These physical environments provide the design space for the placement of physical interactive elements in the car. Ergonomic issues play a major role when deciding upon the placement of such interactive elements, e.g., displays should be

---

5. <http://www.economist.com/node/13725743>



**Figure 12.4** Interaction areas of a driver (after Kern and Schmidt [2009]).

arranged to minimize the need to remove the eyes from the road. It also has a direct impact on the driver's ability to use direct manipulation, e.g., on touch screens as they need to be reachable at an arm's length at the maximum. An example of the problems within this design space is the use of the 17" touch screen in the Tesla Model S. However, it forces the driver to look at the screen and does not comply with the 30° viewing angle as is expressed in the NHTSA guidelines NHSTA [2010].

In the following sections, we will introduce each *modality* separately before we discuss how they can be combined for current and future HMIs in cars.

### 12.4.1 Haptic Controls

The growth of automotive functionality makes it impossible to be controlled solely by increasing the number of buttons, switches and sliders placed on the dashboard Döring et al. [2011]. Rather, such an approach leads to a setup wherein the buttons are spread all over the available space, requiring constant visual attention while interacting. An example of this problem is found in the control panel of the Opel Zafira, as seen in Figure 12.5.

Since physical buttons cannot adapt to content dynamically, other modalities such as screens should be used in addition [Riener and Rossbory 2011]. For such a combined use of haptic input and screen output, dedicated buttons can also be substituted by a knob acting as a generic button and slider with a dynamic function depending on the interaction context (Figure 12.6).



**Figure 12.5** Button control panel in the Opel Zafira.

Drivers are able to select a displayed item on the screen by simply turning or flicking the knob to the left or the right. The selection of the item currently highlighted on the screen can be activated by simply pressing the button.

A study by [Rogers et al. 2005] showed that this kind of interaction is useful for repetitive tasks, e.g., to navigate menu structures and the occasional text entry. This kind of operation has become a de-facto standard in middle-class or high end cars. Although it was designed to ease access, its operation causes high distraction, especially with first time users Bowler [2013]. Drivers may be forced to navigate a tree of items back and forth if they are not familiar with the hierarchical structure of the menu. One of the advantages of the multifunctional knob is that the muscle memory learns the position of the knob easily and it can eventually be operated without looking at it. However, these advantages are tempered by the fact that the feedback of the selection is entirely visual, requiring the driver to permanently observe the changes on the screen.

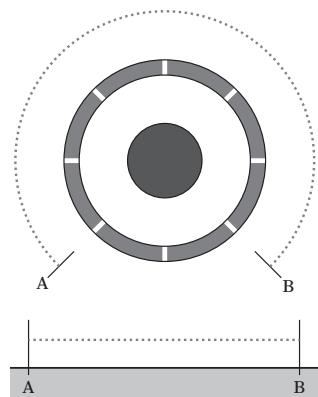
In 2007, the Immersion TouchSense PR-1000 device<sup>6</sup> enhanced the knob with force-feedback functionality aimed at emulating the feel of different conventional controls. The manufacturers classify the following control modes:

6. <http://www.upcindex.com/608819325170> (accessed 06/13/2016)

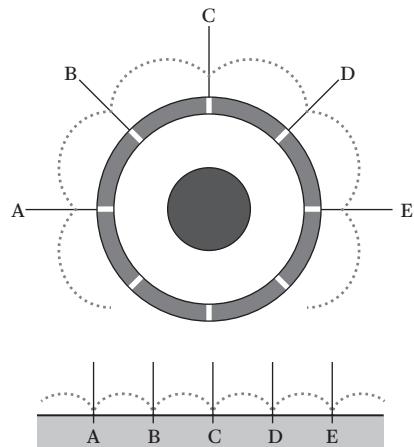


**Figure 12.6** Knob control in a car. Source <http://www.proctorcars.com/in-car-technology-the-ultimate-guide-to-infotainment-systems/>

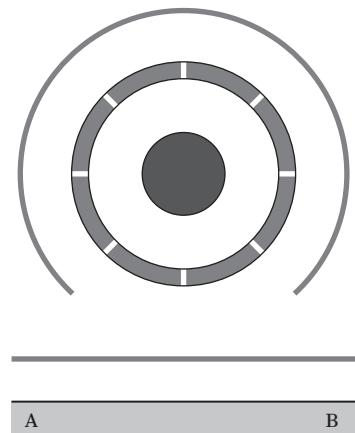
- barrier: Restriction of the motion by a hard stop, e.g., to indicate the first and last item of a list;



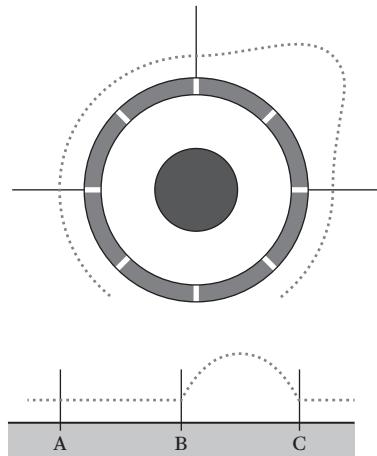
- detent: Use of notches of adaptable size and shape, e.g., to indicate selections in a list or increments;



- constant force: Use of a force feedback, e.g., to indicate friction or momentum;



- hill: A special case of a detent, e.g., to indicate a menu wrap-around; and



- compound: any combination of the above mentioned effects to better match operational steps.

This kind of multimodal addition lowers the requirement for visual feedback and is implemented, e.g., in the BMW iDrive system.<sup>7</sup>

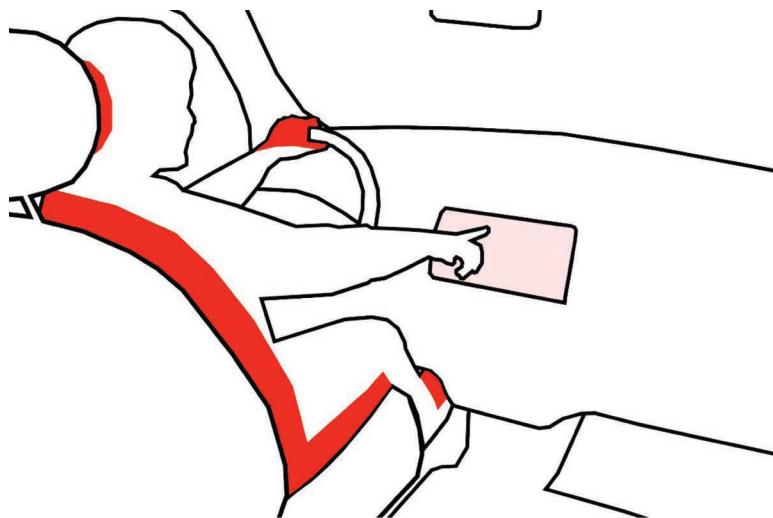
Apart from the omnipresent knob, haptic controls are also under investigation to be used as an output modality. Figure 12.7, taken from [Richter and Wiethoff 2011], shows potential areas of tactile feedback. These are: (i) the steering wheel, (ii) the driver's seat, and (iii) the gas pedal. In the following, we will focus on the first two.

Enriquez et al. [2001] introduced such a system for the steering wheel. Inflatable pads at specific areas on the steering wheel were used in a simulator to alert the driver via fixed frequency pulsation of possible problems, e.g., low fuel warnings. The usage of different frequencies was proposed to convey extra information. Since the driver has continuous contact with the steering wheel and with hands being quite sensitive to touch, it was thought to provide a good means for providing information. However, the driver is required to have both hands on the wheel to get all information. In case of longer trips, drivers tend to only have a light touch on the wheel so that the driver may miss conveyed information. A major drawback of this approach is the experimental finding that this kind of feedback increased reaction times [Enriquez et al. 2001].

Another element of the car that the driver has continuous contact with is the driver's seat. Different haptic messages on the seat can be distinguished by having

---

7. [http://www.bmw.com/com/en/insights/technology/technology\\_guide/articles/idrive.html](http://www.bmw.com/com/en/insights/technology/technology_guide/articles/idrive.html)



**Figure 12.7** Tactile feedback in cars according to [Richter and Wiethoff 2011]. Colored areas indicate locations on the body for tactile stimuli.

vibrations in different places. As the back is not very sensitive, stimuli in this region must be stronger to be unambiguous. Google filed a patent about such a system to provide a stimulus when the vehicle crosses a lane while being on overspeed [Tijerina and Pilutti 2002].

Additionally, safety belts can be employed to provide haptic feedback. The Citroen C4 Picasso provides haptic feedback for its Lane Departure Warning System.<sup>8</sup> If the speed exceeds 80 km/h, the safety belt vibrates when the vehicle accidentally crosses an unbroken white line. In the case of the C4, the location of the vibration is not used to convey meaning, so the driver just receives the information that a line has been crossed but does not know which one. The accuracy of the image processing to recognize the lines works very well and false alarms are reported only rarely. While the system is useful when driving on the highway or on the road, it still reports many false positives in case of ongoing construction work with preliminary lane markings.

Haptic alert systems must be designed with care and are required to match the current situation. Alarms issued too frequently decrease the driver's acceptance

---

8. <http://www.citroen.co.uk/about-citroen/technology/lane-departure-warning-system>

and trust in the system. Incorrect design may lead to unwanted responses of the driver, such as ignoring the message altogether or even causing dangerous situations. Inappropriate use may also actually increase the driver's distraction [[Lee et al. 2004](#), [Summerskill et al. 2004](#)].

### **12.4.2 Touch Screens**

Touch screens are well known from smartphones and are currently being adopted by car manufacturers. One of the biggest advantages over knob controls is that interaction happens directly by pointing on an interactive element on a screen. No further interaction like rotating the knob or multiple pushes of a button are required, leading to faster, more direct interaction [Forlines et al. \[2007\]](#). [Harvey et al. \[2011\]](#) investigated 20 different tasks performed by knobs and touchscreens while driving. They found that touchscreens outperformed the knob in all aspects, including secondary task time, errors and subjective usability. One reason lies in the fact that vision is the dominant modality for feedback and can be processed faster and more precisely than haptic input [[Ernst and Bülthoff 2004](#)]. Some touch screens provide haptic feedback, e.g., by vibrating when a valid action was triggered in order to increase learnability through muscle memory. However, [[Rydström and Bengtsson 2007](#)] observed in a desktop study that haptic information was often ignored if visual information was available.

Because the majority of deployments add touch capabilities to the existing infotainment screen, the interaction space is usually very limited. However, there are some exceptions like the Tesla Model S with a huge 17-inch display or the Porsche 918 Spyder with a touch screen replacing the center stack (see Figure 12.8). This kind of interface promotes touch interaction.

However, touch-based interaction strictly relies on permanent visual feedback and thus precise hand-eye coordination and visual attention with serious consequence for safety. Moreover, cars' swings and bounces severely influence the driver's precision in touching the desired region. It is not unlikely that the driver may trigger an unwanted action, leading to additional interaction steps to undo it.

As for the benefits, only one hand is required for the interaction with the displayed controls and developers and designers have great freedom for arranging them on the screen. This allows for information and controls to be tailored to the current task and displayed as needed. However, the usual absence of any haptic feedback prevents a user from establishing a muscle memory. Consequently, learnability with this kind of modality is low as the layout of controls usually differs between screens.



**Figure 12.8** Interior of the Porsche 918 Spyder.

With over-the-air updates, it would also be possible to change the UI on the fly. Elon Musk announced such a UI design update for their Firmware Version 7.0 at a press conference. Consequently, users would be forced to relearn the layout and may initially be confused by different appearance of the controls.<sup>9</sup>

One approach to overcome the drawback of missing acknowledgements found with haptic controls is the concurrent use of audible feedback. This is usually accomplished by a short beep or clicking sound that is played back when the driver triggered an action as a result of touch interaction. A study by [Pitts et al. 2009] showed a preference of drivers for multimodal feedback over visual feedback only. For haptic feedback, they employed a Touchsense 8.4" LCD touchscreen demonstration unit from Immersion Corporation<sup>10</sup> equipped with haptic actuators. A combination of touch with audible feedback or with haptic feedback was clearly preferred. However, a combination of all three modalities showed the best results. In contrast, [Rydström and Bengtsson 2007] observed, in a desktop study, that this kind of haptic information was often ignored, if visual information was available.

Richter and Wiethoff [2011] discuss different scenarios with a remote tactile feedback, wherein actuators are placed in the driver seat. They suggest different

9. <http://www.teslarati.com/no-home-run-tesla-emphasizes-software-updates-press-conference/>

10. <http://www.immersion.com/touchsense-products/touchsense-haptic-enabling-kit-for-mobile-oems/>

stimuli before, during and after a touch interaction. As an example, consider a scrolling interaction: before the driver touches the screen, this kind of feedback may be used to inform about proximity to interactive part of the screen with the controls. A change in the driver's finger's location can be reflected by a change of the haptic stimulus. The driver can then swipe on the screen to scroll. Once the finger moves away from the touch screen, they suggest another stimulus indicates how many items have passed without detailing how this would be realized. They found that this was suitable to reduce visual load. However, following the discussion above, this is also likely to be ignored by drivers since the number of different stimuli tends to be high.

Pfleging et al. [2012] investigated a combination of touch-based interaction and spoken input. Drivers had to select an object by voice and manipulate it via touch in a follow-up step. For instance, the driver was expected to say "passenger window" and then perform a sliding top-down gesture on a touch-pad mounted at the steering wheel to open the window. A study conducted in a driving simulator showed that this approach was not suited to decrease distraction but offered greater flexibility. In essence, they confirmed the potential of using multiple modalities, even if they are used as rather primitive inputs.

### 12.4.3 Gestures

Gesture controls employ different technologies, such as infrared image analysis of depth cameras or capacitive proximity sensors. Compared to touch input, gestures do not require the driver to interact in a confined area of a given screen, but may be performed wherever the driver's hands are anyway. Riener et al. [2013] found that most gestures were performed "in a limited region in the car, bounded by the "triangle" steering wheel, rear mirror, and gearshift." Riener showed that "in-vehicle gestural interfaces are easy to use and increase safety by reducing visual demand on the driver". They are also not too distracting and can be performed with one hand [Patten 2007]. Bach et al. showed that, although not directly reducing driving task errors, gestures are able to lower the visual demand during interaction [Bach et al. 2008]). The error-rates for the detection of gestures depend on the sensors employed to perform the recognition and reliable quantitative studies are still missing.

There are two variants of gestures that have been investigated: macro gestures and micro gestures. The system developed by Rahman et al. [2011] enables the control of media devices by macro gestures, i.e., hand moves in the air. In order to perform this type of gesture, the driver has to remove one hand from the steering wheel.

In their meta-study [Young and Regan \[2007\]](#) categorize this as Biomechanical (Physical) Distraction with a negative impact onto driving performance. Biomechanical distraction occurs, when driver removes one hand from the steering wheel for an extended time period in order to physically manipulate an object not required to drive safely.

If they are properly designed, gestures feel familiar and natural [Vatavu \[2012\]](#). Currently, the selection of gestures is customized per manufacturer or even car model. Although a set of gestures to trigger certain actions have been established for smartphones, so far, no standard set of gestures exist for in-car usage. [Riener et al. \[2013\]](#) made first steps toward proposing common gestures, but these did not yet find their way into an industry standard.

Micro gestures are usually performed near the steering wheel and appear to be safer than full-scale macro gestures [Carrino et al. \[2012\]](#)). For instance, the system of [\[Endres et al. 2011\]](#) reduces gesturing to finger gestures so that the driver's hands can remain on the steering wheel and trigger actions by finger movements. Continental already introduced a production-ready support of micro gestures near the steering wheel,<sup>11</sup> as shown in Figure 12.9.

In order to precisely define the gesture interaction area, they employ two transparent plastic panels inside the steering wheel, which the user can operate with his thumbs, similar to a touchpad. Gesture detection is realized with the help of a 3D camera system to translate infrared depth images into a 3D image with a resolution in the range of millimeters. The user may have contact with the plastic panels but can also gesture in the air without touching the panels at all. The system is currently capable of detecting different gestures, e.g., to set the navigation, browse through apps, start music, answer calls, and control the on-board computer using gestures that are similar to those available on smartphones. For example, moving his fingers up and down in a uniform movement while keeping his hands on the steering wheel, the driver can accept calls or reject them. In their press release, Continental claims that preliminary user studies verified that the system was easy to operate with low distraction. Comparable interactive elements could also be employed near the gear shift, with capacitive sensing used to detect this kind of gestures via proximity sensing [Braun et al. \[2015\]](#).

---

11. <https://www.continental-automotive.com/en-gl/Passenger-Cars/Interior/Comfort-Security/Driver-Status/Gesture-Control>



**Figure 12.9** Gesture control on the steering wheel. ([http://www.continental-corporation.com/www/pressportal\\_com\\_en/themes/press\\_releases/3\\_automotive\\_group/interior/press\\_releases/pr\\_2016\\_05\\_10\\_wheel\\_gestures\\_en.htm](http://www.continental-corporation.com/www/pressportal_com_en/themes/press_releases/3_automotive_group/interior/press_releases/pr_2016_05_10_wheel_gestures_en.htm), last accessed 6/14/2016)

#### 12.4.4 Handwriting Recognition

Given the obvious safety concerns, there are only a few handwriting recognition tasks conceivable in cars. Audi and BMW, for instance, enhanced the surface of the existing knob with touch capabilities to write single characters as they can be written without requiring visual attention [Pickering et al. 2007]. Figure 12.10 shows a rear-seat passenger operating the touch-enabled knob of a BMW.

However, explicit feedback is required to indicate to the driver that her input was recognized successfully. As such, this kind of input can be used to shorten lists by entering the starting letters, e.g., for a destination in a navigation system. Longer text entries would consume too much time and increase distraction.

Additional challenges arise with lower recognition rates under driving conditions on bumpy roads. Also, users must be able to use their preferred hand. Given



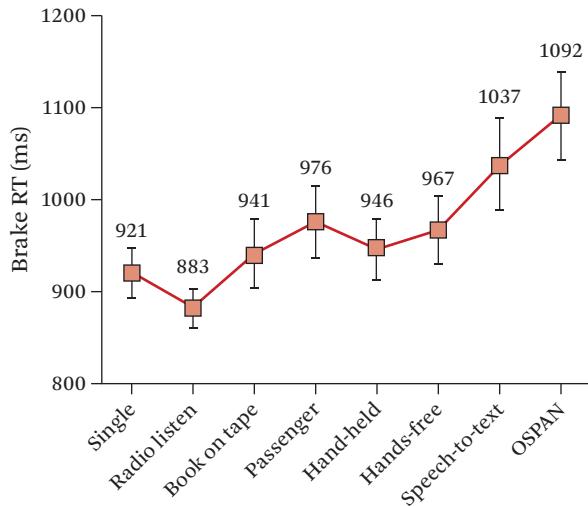
**Figure 12.10** Handwriting recognition on the touch-enabled knob from BMW. (<http://www.bmwblog.com/2012/07/10/introducing-the-bmw-idrive-touch-controller/>)

the wealth of problems associated with this input modality, it remains one of the more gimmicky input interfaces with only few actual installations.

### 12.4.5 Voice

*Spoken dialog systems* (SDS) promise to reduce hands and eyes free access while driving since they interfere to a lesser extent with the resources needed for the primary tasks [Wickens 2002]. It also has been shown that voice based interaction is less distracting than interaction with visual and haptic interfaces [Maciej and Vollrath 2009, Tsimhoni et al. 2004]. Furthermore, voice has been proven an ideal modality for secondary tasks [Salvucci 2001]. By contrast, the results of the more current AAA or Utah studies [Strayer et al. 2013] claim a 20-s residual effect to regain situational awareness after a voice interaction. They claim an increase in crash risk based on the observation of brake reaction times. These attentional effects via the cognitive load of vocal tasks are shown in Figures 12.11 and 12.12. However, no actual crashes were reported.

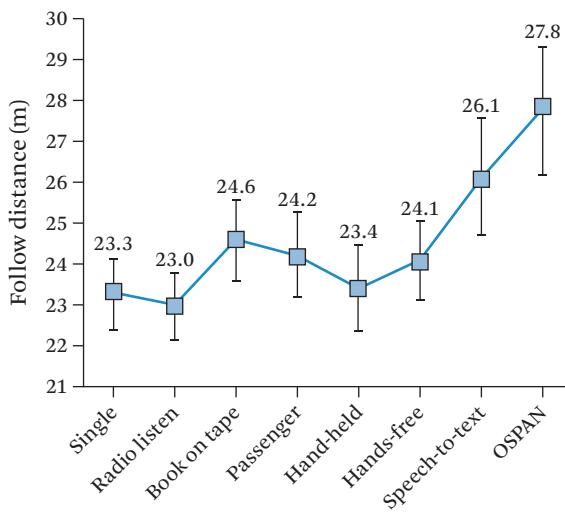
The tasks in the Utah study were described by [Strayer et al. 2013] as follows:



**Figure 12.11** Mean brake reaction time in the Utah study to a lead vehicle brake light onset for the eight task conditions in the Utah simulator. Error bars are standard errors taken from Young [2014].

- single: a baseline single task without any secondary tasks;
- radio listen: concurrent listening to radio;
- book on tape: concurrent listening to a book on tape;
- passenger: concurrent conversation with a passenger sitting next to the participant;
- hand-held: concurrent conversation on a hand-held cell phone;
- hands-free: concurrent conversation on a hands-free cell phone;
- speech-to-text: concurrent interaction with a speech-to-text interfaced e-mail system; and
- OSPAN: concurrent performance with an auditory version of the Operation Span task.

OSPA measures the attempts to recall single syllable words in serial order while solving mathematical problems [Turner and Engle 1989]. While executing the tasks, the drivers were able to keep their eyes on the road and their hands on the steering wheel with the hand-held task being the only exception.



**Figure 12.12** The mean following distance to the lead vehicle for the eight task conditions in the Utah study. Error bars are standard errors from [Young \[2014\]](#).

The measures like brake reaction times have been criticized by [\[Young 2014\]](#) on the grounds that “drivers self-regulated their following distances to compensate for the slight increases in brake response time while performing auditory-vocal tasks” [\[Young 2014\]](#) which was not taken into account in the study.

Most of the current SDSs in cars do not take advantage of the positive effects of voice-based interaction. Dialog structures that do not meet their expectations lead to perceptions of a dialog as exhausting and complicated [\[Hackenberg et al. 2013\]](#).

Usually, a dialog is initiated if the user presses a Push-to-Talk button on the steering wheel (see Figure 12.11). A sample dialog for a command-based dialog, taken from [\[Hofmann 2014\]](#) is the following.

```

Driver: Enter Navigation.
System: Please say the name of the town, the street, and the house number.
Driver: Stuttgart.
System: Stuttgart accepted. Would you like to enter a street?
Driver: Yes.
System: Please say the street.
Driver: Koenigsstrasse.
System: Koenigsstrasse accepted. Would you like to enter a house number?

```



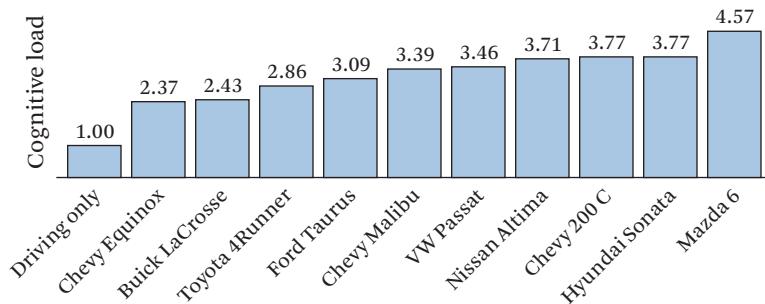
**Figure 12.13** PTT button on a steering wheel. (From Hofmann [2014])



**Figure 12.14** Infotainment screen in an ongoing spoken dialog. (From Hofmann [2014])

The way users provide their input may be ungrammatical or consist of mainly elliptical phrases. In the step-by-step guidance following the initial request, the user is restricted to short answers in a system initiative manner. More complex dialog phenomena, such as answering out-of-focus questions or user adaptation, are usually unsupported as well as any contextual adaptation.[21]

The concept of “speak-what-you-see” [Niemann 2013] provides the driver with visual clues about valid spoken input. As an example, the screen shown in Figure 12.14 provides hints, like “Navi” or “Radio”, about what the user may say to start a dialog.



**Figure 12.15** Cognitive load while utilizing infotainment system of different cars by voice on a scale from 1–5 according to the second Utah study [Strayer et al. 2015].

In-car spoken dialog systems also limit the set of available commands upfront to decrease recognition errors in the presence of a noisy environment. There are, to our knowledge, no systems on the market that dynamically perform such adaptations dependent on noise measurements. They usually rely on effective noise cancellation to account for the driving noise.

Hence, accounting for recognition errors has to be considered, maybe even more than in office settings, from the early stages when designing the voice interface in these environments. This has also been a result of the second Utah study where the authors relate higher cognitive load demands to recognition errors. In this case the baseline was the *Wizard-of-Oz* study of the first Utah study with a recognition rate of almost 100%. Figure 12.15 shows that the cognitive load demands range from moderate to high[22] depending on the vehicle[23]. It remains unclear whether or not these results are due to the noise profiles of the vehicles or their dialog designs. Reimer et al. [2013] showed that the cognitive load of the driver and task completion time significantly increases not only when the speech recognition accuracy drops, but also when the Back End System (BES)—which interacts with the SDS—fails to fulfill the driver's command. This is in accordance with another finding of the second University of Utah's study which states that “robust, intuitive systems with lower levels of complexity and shorter task durations result in less cognitive distraction” [Strayer et al. 2015].

A recent American Automobile Association report by Strayer et al. [2017] tested voice and touch interactions (in 2 physical locations) in 30 model 2017 vehicles for the tasks of: adjusting the audio entertainment subsystem, placing phone calls, creating text messages, and entering destinations to the navigation subsystem. These were compared with a low reference demand (simply driving without any other

secondary tasks) and a known high cognitive demand task (listening to a list of numbers and replying with the second to last one). The measures included cognitive demand, visual demand, subjective workload, and interaction time. Voice interaction was found to result in the lowest cognitive demand (but still greater than the high demand reference task), the lowest visual demand, and the lowest subjective workload (both below the high reference demand load). However, auditory interactions were longer on average than the touch-screen-based ones, in large part because the vocal navigation commands were much longer than those of the touch-screen modalities, all of which were significantly longer than the 24-s reference target set by the US National Highway Transportation Safety Agency [NHSTA 2013]. For the other three tasks, voice interaction was the fastest. There was no good interaction modality for navigation, resulting in a recommendation that navigation systems only be active when the car is at rest. A combined measure of overall demand showed that vocal interactions were the lowest among the three modalities, although still higher than the reference task.

Voice interfaces, as [Schnelle 2007] among others pointed out, feature some modality inherent characteristics.

- **Transience:** Remembering voice output is mainly controlled by the short-term memory. Hence, listening to long utterances has the effect that users forget most of the information that was given in the beginning. For instance, a voice-only output application has to face the lost-in-space problem. Consequently, voice is not an ideal modality for delivering large amounts of data.
- **Invisibility:** Users have only little clues what they may say to perform a wanted action. This also leads to the effect that users do not feel in control of the system. As an advantage, visual and aural perception may happen in parallel. It is also possible to deliver information without the need to switch context.
- **Asymmetry:** People can speak faster than they can type but listen much more slowly than they read. This affects the amount of audio data and the information being delivered.
- **One-dimensionality:** While the eye is an active organ, the ear is passive. Hence, the ear cannot browse a set of recordings as the eye can browse a set of documents.

As these characteristics are inherent to the modality of voice, special considerations for voice user interface design to cope with these issues are required. Especially, asymmetry hints at supplementing a basic audio interface with additional displays as a permanent, visible, and two-dimensional counterpart to audio.

Upcoming systems will increasingly employ more unconstrained natural language understanding (NLU) capabilities that users are already familiar with from personal assistants. Bobrow [1964] defines natural language understanding as follows:

“A computer understands a subset of English if it will accept input sentences which are members of this subset, and correctly answer questions based on information contained in these sentences. This ability must extend to deductions based on implicit information contained in several sentences”

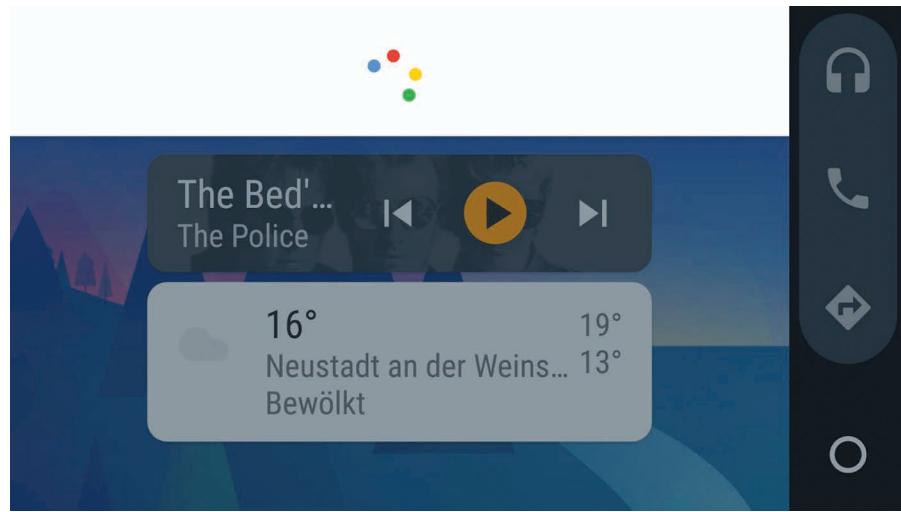
NLU is a subtopic of natural language processing (NLP) that has its roots in the “1950s as the intersection of artificial intelligence and linguistics” [Nadkarni et al. 2011]. A prominent example to demonstrate the utilization of knowledge learned or deducted is *Question Answering* or *Voice Search* (e.g., [Bobrow 1964, Winograd 1972, Wang et al. 2008]). Here, users formulate their request to perform a certain action via a spoken query, e.g., for cheap Chinese restaurants:

- Driver: Navigate me to Cheap chinese restaurants near Leicester Square in London.
- System: I found 5 inexpensive restaurants near Leicester Square in London that serve Chinese cuisine. These are . . . Which one do you want?
- Driver: . . .

The automotive industry has a clear preference to such so-called one-shot entries where the drivers enters all needed query parameters in a single utterance. This is motivated by the aim for only few dialog turns. Traditionally, speech recognition in embedded systems used to be time consuming with higher time delays for the recognition process resulting in unacceptable lengths of the overall dialog. As computational power and memory increased in cars, or as they become more connected to the internet themselves and the computing power to connected cars available there, it is also acceptable to add or remove additional query parameters as needed in follow-up utterances to refine the results.

A sample dialog for voice search, i.e. local search when running embedded in the car, utilizing subsequent addition of query parameters, taken from [Wang et al. 2008] is as follows.

- Driver: Tell me about restaurants in London.
- System: I know of 596 restaurants in London. All price ranges are represented.  
Some of the cuisine options are Italian, British, European, and French.



**Figure 12.16** Voice input with Android Auto.

Driver: I'm interested in Chinese food.

System: I know of 27 restaurants in London that serve Chinese cuisine. All price ranges are represented. Some are near the Leicester Square tube station.

Driver: How about a cheap one?

System: I know of 14 inexpensive restaurants that serve Chinese cuisine. Some are near the Leicester Square tube station. Some are in Soho.

One of the main innovators in the area of voice search is Google. They also started to direct their efforts towards spoken user interfaces in an automotive environment with Android Auto<sup>12</sup> shown in Figure 12.16.

Similarly, Apple has entered the market with their CarPlay.<sup>13</sup> It is still questionable how much of the smartphone experience can be transferred to the car. The Utah study [Strayer et al. 2013] discussed above employed traditional smartphone applications that performed worse in terms of distraction than systems that were built for the purpose of in-vehicle usage (c.f. Figure 12.15). However, Android Auto and Apple CarPlay were not explicitly tested.

12. <https://www.android.com/auto/>

13. <http://www.apple.com/ios/carplay/>

Another drawback of these approaches is that they currently require an online connection to work. This may result in erroneous conditions when internet connectivity get disrupted, e.g., when the car enters a tunnel. Therefore, voice interfaces ought to be hybrid, with recognition and semantic interpretation happening, to different degrees, in the car and in the cloud in parallel. This allows a system to take advantage of current technology that require computational resources that are not yet available in the car while always being able to respond to user input by falling back to the embedded engines.

Reimer et al. [2014] tried to reduce cognitive load by introducing an *Expert Mode*. In this mode, several confirmation steps were removed to decrease the overall listening time. A navigation task dialog taken from [Reimer et al. 2014] for the non-expert mode is the following.

Driver: [Driver presses voice button].  
System: Please Say a Command.  
Driver: “Destination Street Address.”  
System: Destination Street Address, is that correct?  
Driver: “Yes.”  
System: Destination Street Address, please say the city name.  
Driver: “Cambridge.”  
System: Cambridge, is that correct?  
Driver: “Yes.”  
System: Cambridge, please say the street name.  
Driver: “Massachusetts Avenue.”  
System: Massachusetts Avenue, is that correct?  
Driver: “Yes.”  
System: Massachusetts Avenue, please say a house number.  
Driver: “1-7-7.”  
System: The number you said was 1-7-7, is that correct?  
Driver: “Yes.”  
System: Yes. Press the voice button to say set as destination or back.  
Driver: [Driver presses voice button]  
Driver: “Set as destination.”  
System: Set as destination, is that correct?  
Driver: “Yes.”  
System: [Set as destination]

In the expert mode 8 steps were removed. The reduced dialog is as follows

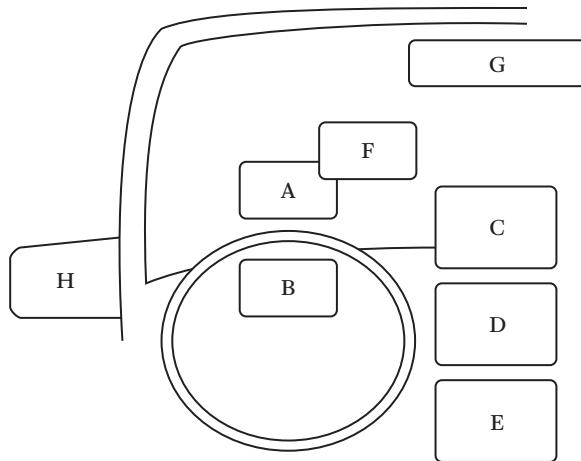
Driver: [Driver presses voice button]  
 Driver: "Destination Street Address."  
 System: Destination Street Address, please say the city name.  
 Driver: "Cambridge."  
 System: Cambridge, please say the street name.  
 Driver: "Massachusetts Avenue."  
 System: Massachusetts Avenue, please say a house number.  
 Driver: "1-7-7"  
 System: The number you said was 1-7-7, is that correct?  
 Driver: "Yes."  
 System: Yes. Press the voice button to say set as destination or back.  
 Driver: [Driver presses voice button]  
 Driver: "Set as destination."  
 System: [Set as destination.]

While the system still relies on state-based dialog flows, it was able to reduce the overall task completion time. However, it did not help to decrease the requirement for visual feedback as the time the driver's eyes were off the road. More potential of this approach in terms of a better user experience and reduced cognitive load could be available by employing one-shot address entries or voice-search based dialogs as described above. These are still available in these approaches.

#### **12.4.6 Secondary Displays**

Displays can be made available at different positions in a car. [[Wittmann et al. 2006.](#)] identified the following display positions shown in Figure 12.17:

- A. above speedometer at front window;
- B. at speedometer;
- C. above middle console;
- D. middle console in a middle position;
- E. At the bottom of the middle console;
- F. On the windshield;
- G. Rear mirror.



**Figure 12.17** Available display positions in a car.

In addition, the following position is used

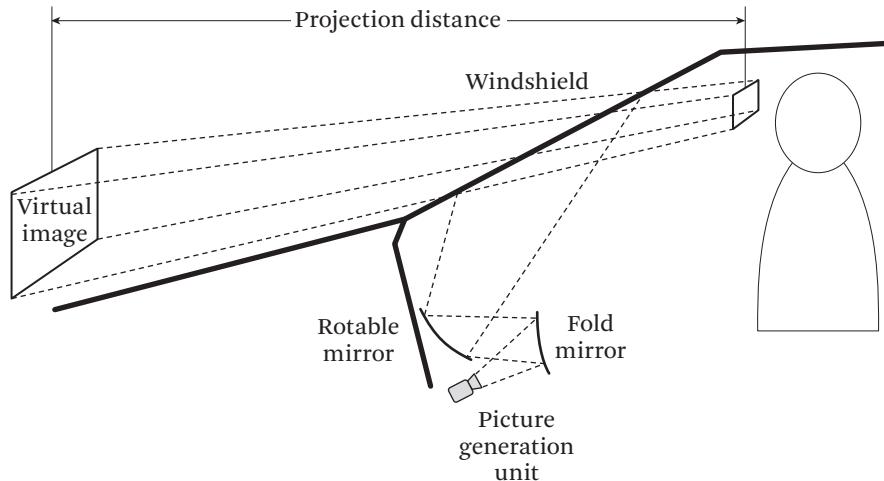
#### H. side mirror.

These positions are categorized into head-down-displays (HDD) where the driver directs his gaze away from the street, usually downward and head-up-displays (HUD) where the driver can keep his eyes on the road. HUDs usually feature shorter display-road transition and eye accommodation times [Lauber et al. 2014]. While primarily used in planes, installations in a car were already available 1998 in the Corvette C5 [Bowler 2013].

Figure 12.15 illustrates a possible approach. A projector displays an image onto a special layer on the windshield. Because of the windshield's curved nature, this requires the usage of special mirrors to compensate for the optical distortions. As a result, the driver perceives a virtual image floating over the hood at a distance of about two meters.

With this technique, the driver is not forced to take the eyes off the road to receive the information, which decreases cognitive load and increasing reaction times [Normark et al. 2009]. These displays are also more efficient in presenting information, measured as gaze retention period depending on age, than classical displays in cars [Ablassmeier et al. 2007].

However, there are also problems with HUDs. Prinzel and Risser [2004] refer to this as *cognitive attention capture* or *perceptual tunneling* and there are some issues inherent with the accommodation of the human eye when switching visual focus



**Figure 12.18** Virtual image projection with a head-up display. (After <http://continental-head-up-display.com/de/>)

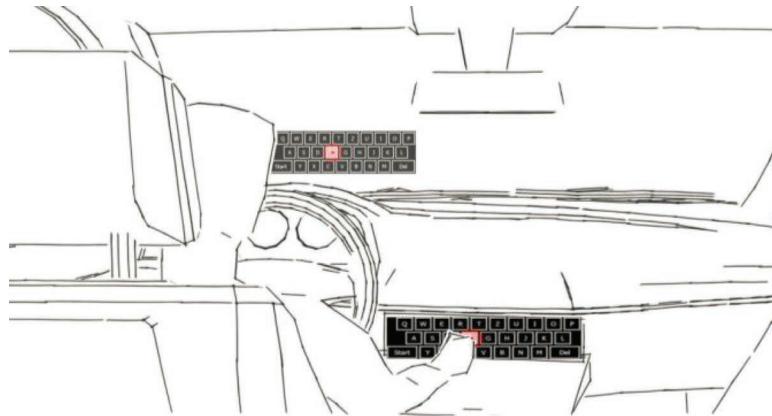
between the HUD projected onto the windshield and the street, a process that may take anywhere from 200–800 ms depending on the age of the driver and the lighting conditions [Iavecchia et al. 1988, Wolffsohn et al. 1998].

A more thorough analysis of benefits and drawbacks of HUDs is still an ongoing research issue. Current installations are found in, e.g., the BMW 5 and 7 series and provide the driver with information about the vehicle's speed and navigational information.<sup>14</sup>

Lauber et al. [2014] added touch capabilities to HUDs in a concept that they call *What You See Is What You Touch* (WYSIWYT). They use the head-down displays located in the middle console and at the steering wheel to enable touch interaction. The finger and the control itself is displayed in the head-up display, as shown in Figure 12.19.

The basic idea of this approach is to liberate the driver from the requirement to avert the eyes from the road. In their study, performed in a driving simulator, they compared various touch interaction concepts like *direct touch selection*, *hover*, and *touch*, where the selection was made with a hovering gesture and activated by touching the screen, or *hover and click*, where the selection was made with a hovering pointing gesture activated by a button on the steering wheel for alphanumeric

14. <http://www.bmw.de/de/footer/publications-links/technology-guide/head-up-display.html>.



**Figure 12.19** Visualization of touch interaction in a head-up display. [Lauber et al. \[2014\]](#)

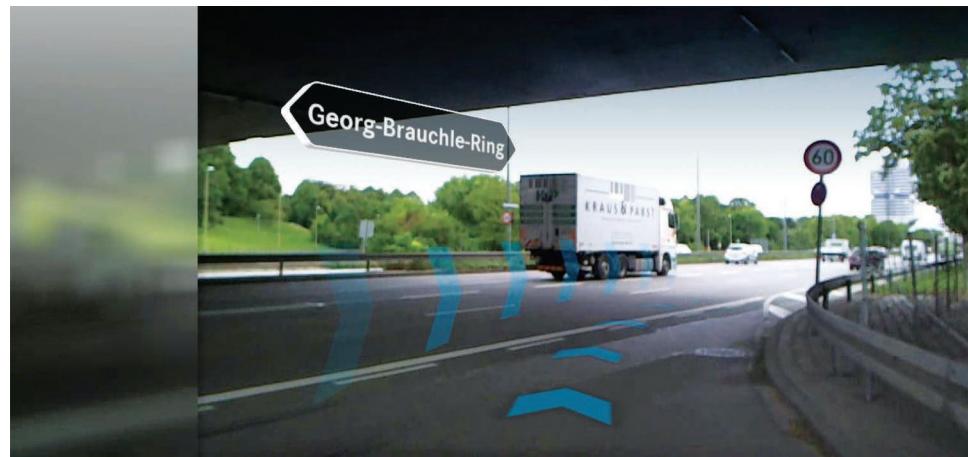
text input. For the selection, direct touch selection worked best, while the button on the steering wheel turned out to be the best choice for the activation. However, they did not investigate the cognitive load of the driver, especially the additional load that came along with the multitude of items displayed on the head-up display. A study by [\[Sun et al. 2015\]](#) showed that “when visual attention must be divided and allocated to distinct but spatially commingled stimuli, participants’ performance is compatible with a biased competition mode.” They state that attention to the secondary task may hijack the resources of the primary task. Especially, in cognitive high-load conditions when the traffic situation requires the driver’s full attention, warnings that pop-up compete with the driver’s attention for the primary task.

Newer systems employ augmented reality to enhance the surrounding with information. For instance, [\[Rao et al. 2014\]](#) present such an approach for navigation systems, as shown in Figure 12.20.

The benefit of these systems is that the information is located where it is relevant, thus decreasing the cognitive load for the user. However, these concept studies usually work with a large field-of-view. Current systems are “still limited to the relative small area floating above the instrument cluster” [\[Rao et al. 2014\]](#). Hence, the benefit of these systems is potentially high but almost impossible to implement with the currently available technology.

#### 12.4.7 Gaze

An important goal of secondary displays is to minimize the times the user’s eyes are not focused on the road, i.e. to minimize gaze direction changes. However,



**Figure 12.20** Navigation aid based on augmented reality projections on the windshield. Rao et al. [2014] <http://www.3drealms.de/ismar-2014-ar-in-vehicle-implementation/>

the driver's gaze can also be employed as a modality. First systems are available to determine if a driver is fatigued depending on her head position and eye closure [Haworth et al. 1988]. Assistance systems in today's vehicles suggest the driver take a break in this case. For instance, the Volkswagen Golf VII shows a coffee cup icon in the center of the dashboard.<sup>15</sup>

Gaze could also be used to influence other modalities, like voice input. Microsoft's patent US20130307771 A1 limits the available inputs based on user gaze [Parker et al. 2012]. It does not focus on vehicles but there is a high potential for a deployment in this environment. Imagine a menu displaying several options that are currently available. In a speak-what-you-can see approach, the spoken dialog manager uses a grammar to capture the labels on these options. This approach extends it in a way that as the sensor detects the focused object, it will choose the appropriate instruction set, grammar, or recognition action.

In combination with road scene activity and vehicle information, gaze can also be employed to reduce the amount of warnings of driver assistance systems [Fletcher et al. 2005]. Fletcher et al., [2005] combined eye gaze with the detection of traffic signs and informed the user of any missed signs if the vehicle's state is not in compliance with the sign. They conclude that "if the driver has not seen the sign, and the car's state is not in compliance with the sign, the driver can be informed

15. <http://www.volksvagen.co.uk/technology/passive-safety/driver-alert-system>

with high priority. If, however, the driver appears to be aware of the sign, or the vehicle is not in an incorrect state, the information can be made available to the driver in a more passive manner.”

Other approaches try to use the knowledge about the driver looking at an object outside the vehicle to combine it with geo-location and annotated maps to be able to provide answers to questions like “Is there a restaurant on this side?” Kang et al. [2015] achieved an accuracy of 65% correct answers on queried buildings.

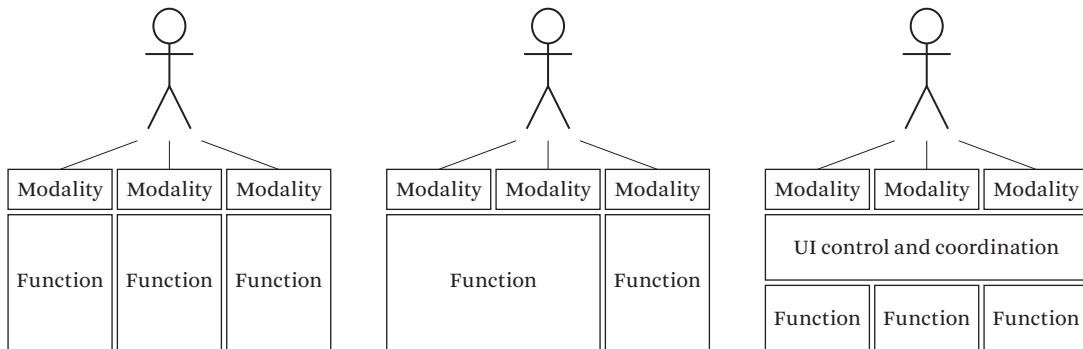
## 12.5

### Summary and Outlook

In this chapter, we argued that the complexity of interaction in the car increased significantly in the past four decades. This led to cognitive overload and distraction of the driver when operating, e.g., the infotainment system. Legal regulations like those from the NHTSA try to provide a frame for safer operation that must be followed by car manufacturers.

One option to reduce cognitive load by the use of multiple modalities was motivated by [Baddeley and Hitch 1974] decades ago. In fact, new modalities available in today’s cars are focusing on reducing cognitive load. This is the pragmatic side of the coin. On the other hand, car manufacturers introduce new interaction concepts that aim for a more intuitive interaction but are also more gimmicky, trying to fulfill hedonic motivations of the overall user experience [Hassenzahl 2007]. Sometimes, the aim to serve the latter seems to be the primary goal of manufacturers as can be seen with the addition of handwriting recognition: The pragmatic use of this modality in the car is low and the hedonic aspects of “playing around” with it were considered to be of higher importance. As a consequence, car manufacturers keep experimenting with ever-more modalities for interaction in a car. However, not all of them can easily be applied to this specific setting since they are likely be designed to work in environments where storage size and computational power do not matter that much, which also leads to increased response times as bemoaned by Nielsen.

While the output side already benefits from a complementary and concurrent employment of additional modalities, mainly to augment visual output (see Figure 12.21 middle), the input side is still predominantly restricted to sequential and alternative employment of different modalities. Therefore, more structured research for concurrent use of complementing modalities and its effect on cognitive load, as it was already described more than two decades ago [Nigay and Coutaz 1993] provides considerable space for future improvements. There are only few use cases where, e.g., beeps or vibrations augment touch interaction. More sophisti-



**Figure 12.21** Modalities are used to control different functions (left), modalities are used to replace other modalities to control the same functions (middle), and alternative and sequential use of modalities to control the available functions (right).

cated solutions that exploit multiple modalities are currently not available in the car but are investigated in research. In some cases, as for 3D augmented reality in the HUD, research is still far from systems that can actually be built. More work will be needed to provide the industry with concepts for more realistic systems. This could also be a way to meet the flaws that were identified by Nielsen and mentioned in the beginning. A holistic concept of a combined use of modalities will also reduce clumsy task flows. Concepts of multimodal dialog managements that have been around for years [Maybury and Wahlster 1998] are still missing in today's cars. A promising approach could be the usage of the W3C MMI architectural pattern<sup>16</sup> that aims at standardizing multimodal dialog architectures. This is illustrated in Figure 12.21 to the right. An important, enabling contribution which is still missing would be a proper taxonomy of modalities with regard to their applicability for various input and output tasks as well as the environmental circumstances for their gainful deployment (not unlike [Card et al. 1991] for input devices). Being able to derive a set of modalities most suited for a user interaction situation with a given input or output intent and some environmental circumstances would be most welcome and help to structure concurrent and complementary use of modalities.

In absence of this taxonomy, some modalities, especially the HUDs, are obviously promising candidates to reduce cognitive load. Unfortunately, approaches to actually measure the reduction in cognitive load are still in their infancy. Some initial models exist [Neßelrath and Feld 2013], but still lack an apparatus than can be

16. <https://www.w3.org/TR/mmi-arch/>

utilized not only in driving simulators but also in real-world scenarios. The correlation between cognitive load and driving performance is well established, however, the development of a suitable measure to quantify the cognitive load imposed by the concurrent or sequential application of the various input and output modalities is one of the more urgent tasks needed to enable UI researchers and manufacturers to apply a more rigorous scientific approach in order to rate the effects of the multimodal HMI on driving performance.

In the near future the upcoming topic of autonomous driving will also have an impact on how drivers will interact with the infotainment system. More and more sensors are expected to replace human senses. Here, we see the following categories that may play an important role.

1. Tutoring system: How to design a system that teaches the driver about the features and the limitations of the self-driving car? [Romoser \[2011\]](#)
2. Hand-over: How to design the HMI for situations where the autonomous driving algorithm detects that it can no longer guide the vehicle and that the driver has to take over? [Walch et al. \[2015\]](#)
3. External communication: How does a car that drives autonomously communicate with other traffic participants, especially pedestrians? [Keller and Gavrilis \[2014\]](#)
4. Ethical issues: How to value the lives of the traffic participants vs. the passengers if the system in case an accident cannot be avoided? [Goodall \[2014\]](#)

While the first two will be less severe as technological progress is being made, the latter are challenges that ought to be addressed right now. However, we will not delve deeper since this is out of the focus of this chapter.

The way we interact multimodally with the car is continuously evolving. Connected services and artificial intelligence get increasing relevance and shape the way we interact with them while driving. This evolution of interaction means and capabilities are still challenging to arrive at a consistent multimodal *Natural interaction* in tomorrow's vehicles. Some approaches are already available in research but need to find their way to industry.

## **Focus Questions**

- 12.1. What is meant by primary, secondary and tertiary tasks that drivers have to deal with?

- 12.2.** Which force-feedback control modes are distinguished to control knobs and what is their purpose?
- 12.3.** Which approaches are identified to overcome the limitation of missing haptic feedback for touch input in the car?
- 12.4.** What are appropriate locations of micro gestures in the car?
- 12.5.** How does voice input compare to touch interaction in terms of cognitive load?
- 12.6.** What are the different categories for secondary displays?
- 12.7.** What are the benefits of augmented reality in the car?
- 12.8.** How can gaze be employed to increase driver safety?

## References

- M. Ablassmeier, T. Poitschke, F. Wallhoff, K. Bengler, and G. Rigoll. 2007. Eye gaze studies comparing head-up and head-down displays in vehicles. In *IEEE International Conference on Multimedia and Expo*, pp. 2250–2252. Beijing, China. DOI: [10.1109/ICME.2007.4285134](https://doi.org/10.1109/ICME.2007.4285134). 508
- C. Baber and J. Wankling. August 1992. An experimental comparison of test and symbols for in-car reconfigurable displays. *Applied Ergonomics*, 23(4): 255–262. 482
- K. Bach, M. Jæger, M. Skov, and N. Thomassen. 2008. You can touch, but you can't look: interacting with in-vehicle systems. In *CHI '08 Proceedings of the Twenty Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, Vol. 7, pp. 1139–1148. DOI: [10.1145/1357054.1357233](https://doi.org/10.1145/1357054.1357233). 495
- M. Bowler. 2013. HMI and driver distraction. In 4th International Conference Automotive Cockpit HMI. Berlin. <http://emsweb.iqpc.com/MediaCenter.aspx> 488, 508
- A. D. Baddeley, and G. Hitch. 1974. Working memory. In *The Psychology of Learning and Motivation*, pp. 47–90. DOI: [10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1). 512
- D. G. Bobrow. 1964. Natural language input for a computer problem solving system. *Artificial Intelligence Project Memo* 66. Massachusetts Institute of Technology, Cambridge, MA. 504
- A. Braun, R. Wichert, A. Kuijper, and D. W. Fellner. 2015. Capacitive proximity sensing in smart environments. *Journal of Ambient Intelligence and Smart Environments*, 7(4): 483–510. 496
- F. Carrino, S. Carrino, M. Caon, L. Angelini, O. A. Khaled, and E. Mugellini. 2012. In-Vehicle Natural Interaction Based on Electromyography. In *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '12) (pp. 1–3). DOI: [10.1017/CBO9781107415324.004](https://doi.org/10.1017/CBO9781107415324.004). 496

- S.K. Card, J.D. Mackinlay, G. G. and Robertson. 1991. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems (TOIS)*, 9(2), 99–122. DOI: [10.1145/123078.128726](https://doi.org/10.1145/123078.128726). 513
- CEC 2007 Commission of the European Communities. 2006. Commission Recommendation of 22 December 2006 on safe and efficient in-vehicle information and communication systems: update of the European Statements of Principles on human machine interface. *Official Journal of the European Union*. Retrieved from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:032:0200:0241:EN:PDF> 483
- T. Döring, D. Kern, P. Marshall, M. Pfeiffer, J. Schöning, V. Gruhn, A. Schmidt. 2011. Gestural Interaction on the Steering Wheel—Reducing the Visual Demand. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems CHI '11*, pp. 483–492. DOI: [10.1145/1978942.1979010](https://doi.org/10.1145/1978942.1979010). 487
- C. Endres, T. Schwartz, and C. A. Müller. 2011. Geremin: 2D microgestures for drivers based on electric field sensing. In *Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI'11*. ACM, New York. DOI: [10.1145/1943403.1943457](https://doi.org/10.1145/1943403.1943457). 496
- M. Enriquez, O. Afonin, B. Yager, and K. Maclean. 2001. A Pneumatic Tactile Alerting System for the Driving Environment. In *2001 Workshop on Perceptive User Interfaces*, pp. 1–7. DOI: [10.1145/971478.971506](https://doi.org/10.1145/971478.971506). 491
- M. O. Ernst and H. H. Bülow. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4): 162–169. DOI: [10.1016/j.tics.2004.02.002](https://doi.org/10.1016/j.tics.2004.02.002). 493
- L. Fletcher, G. Loy, N. Barnes, and A. Zelinsky. 2005. Correlating driver gaze with the road scene for driver assistance systems. *Robotics and Autonomous Systems*, 52(1): 71–84. DOI: [10.1016/j.robot.2005.03.010](https://doi.org/10.1016/j.robot.2005.03.010). 511
- C. Forlines, D. Wigdor, C. Shen, and R. Balakrishnan. 2007. Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 647–656). ACM, New York. DOI: [10.1145/1240624.1240726](https://doi.org/10.1145/1240624.1240726). 493
- N. J. Goodall. 2014. Machine ethics and automated vehicles. In *Road Vehicle Automation*, pp. 93–102. DOI: [10.1007/978-3-319-05990-7\\_514](https://doi.org/10.1007/978-3-319-05990-7_514)
- R. Grimm, J. Davis, B. Hendrickson, E. Lemar, A. MacBeth, S. Swanson, T. Anderson, B. Bershad, G. Borriello, S. Gribble, and D. Wetherall. May 2001. Systems directions for pervasive computing. In *Hot Topics in Operating Systems*. In *Proceedings of the Eighth Workshop on*, pp. 147–151. IEEE. DOI: [10.1109/HOTOS.2001.990075](https://doi.org/10.1109/HOTOS.2001.990075). 479, 782
- L. Hackenberg, S. Bongartz, C. Härtle, and J. A. Sison. 2013. International evaluation of NLU benefits in the domain of in-vehicle speech dialog systems. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 114–120. DOI: [10.1145/2516540.2516553](https://doi.org/10.1145/2516540.2516553). 500

- C. Harvey, N. A. Stanton, C. A. Pickering, M. McDonald, and P. Zheng. 2011. To twist or poke? A method for identifying usability issues with the rotary controller and touch screen for control of in-vehicle information systems. *Ergonomics*, 54(7), 609–625. [493](#)
- M. Hassenzahl. 2007. The Hedonic/Pragmatic Model of User Experience. In *Towards a UX Manifesto. COST294-MAUSE Affiliated Workshop*, Vol. 2, pp. 10–14. DOI: [10.1183/09031936.00022308](#). [512](#)
- N. L. Haworth, T. J. Triggs, and E. M. Grey. 1988. Driver fatigue: concepts, measurement and crash countermeasures. *Technical report, Federal Office of Road Safety Contract Report 72 by Human Factors Group*, Department of Psychology, Monash University. [511](#)
- H. Hofmann. 2014. Intuitive speech interface technology for information exchange tasks. Ph.D. Thesis. Universität Ulm, Germany. [500](#), [501](#)
- J. H. Iavecchia, H. P. Iavecchiaand S. N. Roscoe. 1988. Eye accommodation to head-up virtual images. In *Human Factors*, 30(6): 703–712. [509](#)
- S. Kang B. Kim, S. Han, H. Kim. 2015. Do you see what I see: towards a gaze-based surroundings query processing system. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '15 pp. 93–100. DOI: [10.1145/2799250.2799285](#). [512](#)
- C. G. Keller, and D. M. Gavrila. 2014. Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2): 494–506. DOI: [10.1109/TITS.2013.2280766](#). [514](#)
- D. Kern, and A. Schmidt. 2009. Design space for driver-based automotive user interfaces. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 3–10. Essen, Germany. [486](#), [487](#)
- S. G. Klauer, T. A. Dingus, V. L. Neale, J. Sudweeks, and D. J. Ramsey. 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Analysis. (DOT HS Rep. 810 594).
- R. Kulmala and H. J. Mäuerer. October 2005. Final report and recommendations of the implementation road map working group. *eSafety Forum*, Brussels. [478](#), [781](#)
- F. Lauber, A. Follmann, and A. Butz. 2014. What you see is what you touch: visualizing touch screen interaction in the head-up display. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, pp. 171–180. DOI: [10.1145/2598510.2598521](#). [508](#), [509](#), [510](#)
- J. D. Lee, J. D. Hoffman, and E. Hayes. 2004. Collision warning design to mitigate driver distraction. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pp. 65–72. DOI: [10.1145/985692.985701](#). [493](#)
- J. D. Lee, K. L. Young, and M. A. Regan. 2008. Defining driver distraction. In *Driver distraction: Theory, effects, and Mitigation*, Vol. 13, pp. 31–40. CRC Press, Boca Raton, FL. [482](#)
- J. Maciej and M. Vollrath. 2009. Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, 41(5): 924–930. [498](#)

- M. T. Maybury and W. Wahlster, editors. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Publishers Inc, San Francisco, CA. [513](#)
- M. F. McTear. 2004. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer Science & Business Media. New York. [479](#), [785](#)
- B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek. 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138, 6–12. DOI: [10.3141/2138-02](https://doi.org/10.3141/2138-02). [484](#)
- P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5): 544–551. DOI: [10.1136/amiajnl-2011-000464](https://doi.org/10.1136/amiajnl-2011-000464). [504](#)
- V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman. 2005. An overview of the 100-car naturalistic study and findings (article). *National Highway Traffic Safety Administration*, Paper 05-0400. [483](#)
- R. Neßelrath and M. Feld. 2013. Towards a cognitive load ready multimodal dialogue system for in-vehicle human-machine interaction. In *Adjunct Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Eindhoven pp. 49–52. [483](#), [484](#), [485](#), [513](#)
- NHSTA. 2010. J. Melrose, R. Perroy, and S. Careas. 2012. Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices. *National Highway Traffic Safety Administration*. Docket No. NHTSA-2010-0053 Visual-Manual. [483](#), [487](#)
- NHSTA. 2013. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices, Department of Transportation, Washington, DC. [503](#)
- J. Niemann. 2013. Designing Speech Output for In-car Infotainment Applications Based on a Cognitive Model of Attention Allocation. Ph.D. Thesis. Technische Universität Berlin, Germany. [501](#)
- L. Nigay and J. Coutaz. 1993. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, pp. 172–178. New York. [512](#)
- D. A. Norman. 1986. Cognitive engineering. In *User Centered System Design: New Perspectives on Human-computer Interaction*, 3161. [478](#), [778](#)
- C. J. Normark, P. Tretten, and A. Gärling. 2009. Do redundant head-up and head-down display configurations cause distractions. In *Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment and Design*, pp. 398–404. [482](#), [508](#)
- S. Oviatt. 2003. Advances in robust multimodal interface design. *IEEE Computer Graphics and Applications*, 23(5): 62–68. DOI: [10.1109/MCG.2003.1231179](https://doi.org/10.1109/MCG.2003.1231179). [481](#)
- S. Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *ACM International Conference on Multimedia*, pp. 871–880. DOI: [10.1145/1180639.1180831](https://doi.org/10.1145/1180639.1180831). [482](#)

- F. G. Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4): 429. [478](#), [764](#)
- O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman. March 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, pp. 141–144. ACM. [484](#)
- C. L. Parker, M. L. W. O'Hanlon, A. Lovitt, and J. R. Farmer. 2012. Interaction and management of devices using gaze detection. U.S. Patent Application No. 13/474,723. [511](#)
- C. J. D. Patten. 2007. Cognitive workload and the driver: Understanding the effects of cognitive workload on driving from a human information processing perspective. Ph.D. Thesis. Psykologiska institutionen. [495](#)
- B. Pfleging, S. Schneegass, and A. Schmidt. October 2012. Multimodal interaction in the car: combining speech and gestures on the steering wheel. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 155–162. ACM. [495](#)
- C. A. Pickering, K. J. Burnham, and M. J. Richardson. 2007. A Review of automotive human machine interface technologies and techniques to reduce driver distraction. In *2nd Institution of Engineering and Technology International Conference on System Safety*, pp. 223–228. DOI: [10.1049/cp:20070468](https://doi.org/10.1049/cp:20070468). [483](#), [497](#)
- M. J. Pitts, M. A. Williams, T. Wellings, and A. Attridge. 2009. Assessing subjective response to haptic feedback in automotive touchscreens. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '09 pp. 11–18. DOI: [10.1145/1620509.1620512](https://doi.org/10.1145/1620509.1620512). [494](#)
- L. Prinzel and M. Risser. 2004. Head-Up Displays and Attention Capture. NASA/TM-2004-213000. A Tech Report. [508](#)
- S. Radomski. 2016. Formal verification of multimodal dialogs in pervasive environments. Ph.D. Thesis. Technische Universität Darmstadt, Germany. [481](#)
- A. S. M. M. Rahman, J. Saboune, and A. El Saddik. 2011. Motion-path based in car gesture control of the multimedia devices. In *Proceedings of the First ACM International Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications*, DIVANet '11 p. 69. DOI: [10.1145/2069000.2069013](https://doi.org/10.1145/2069000.2069013). [495](#)
- Q. Rao, T. Tropper, C. Grunler, M. Hammori, and S. Chakraborty. 2014. Implementation of in-vehicle augmented reality. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on Mixed and Augmented reality - Science & Technology*, pp. 3–8. München, Germany. DOI: [10.1109/ISMAR.2014.6948402](https://doi.org/10.1109/ISMAR.2014.6948402). [510](#), [511](#)
- B. Reimer, B. Mehler, J. Dobres, and J. F. Coughlin. 2013. The effects of a production level “voice-command” interface on driver behavior: summary findings on reported workload, physiology, visual attention, and driving performance. White Paper 2013-18A. MIT AgeLab. [502](#)

- B. Reimer, B. Mehler, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and A. Rumpold. 2014. Effects of an “Expert Mode” voice command system on task performance, glance behavior & driver physiology. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI ’14 pp. 1–9. DOI: [10.1145/2667317.2667320](https://doi.org/10.1145/2667317.2667320). **506**
- H. Richter and A. Wiethoff. 2011. Augmenting future in-vehicle interactions with remote tactile feedback. In *Adjunct Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI ’11, pp. 162–163. **491, 492, 494**
- A. Riener and M. Rossbory. 2011. Natural and intuitive hand gestures: a substitute for traditional vehicle control? In *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI ’11, pp. 6–7. **487**
- A. Riener, A. Ferscha, F. Bachmair, P. Hagmüller, A. Lemme, D. Muttenthaler, D. Pühringer, H. Rogner, A. Tappe, and F. Weger. 2013. Standardization of the in-car gesture interaction space. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI ’13, pp. 14–21. DOI: [10.1145/2516540.2516544](https://doi.org/10.1145/2516540.2516544). **495, 496**
- W. A. Rogers, A. D. Fisk, A. C. McLaughlin, and R. Pak. 2005. Touch a screen or turn a knob: Choosing the best device for the job. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 47(2), 271–288. **488**
- M. Romoser. 2011. An autonomous intelligent driving simulation tutor for driver training and remediation: a concept paper. In *Driving Assessment 2011: 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, (pp. 496–502). **514**
- Rydström, A., and Bengtsson, P. 2007. Haptic, visual and cross-modal perception of interface information. *Human Factors Issues in Complex System Performance*, 399–409. **493, 494**
- D. D. Salvucci. 2001. Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human Computer Studies*, 55(1). <http://doi.org/10.1006/ijhc.2001.0472> **482, 498**
- O. Shaer and E. Hornecker. 2010. Tangible user interfaces: past, present, and future directions. *Foundations and Trends in Human-Computer Interaction*, 3(1–2) 1–137. **479, 787**
- D. Schnelle. 2007. Context aware voice user interfaces for workflow support. Ph.D. Thesis. Technische Universität Darmstadt, Germany. **503**
- S. J. Summerskill, J. M. Porter, and G. E. Burnett. 2004. Feeling your way home: The use of haptic interfaces within cars to make safety pleasurable. **493**
- D. L. Strayer, J. M. Cooper, J. Turrill, J. Coleman, N. Medeiros-Ward, and F. Biondi. 2013. Measuring cognitive distraction in the automobile. AAA Foundation for Traffic Safety. **485, 498, 505**

- D. L. Strayer, J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman. 2015. Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems. AAA Foundation for Traffic Safety. [502](#)
- D. L. Strayer, J. M. Cooper, R. M. Goethe, M. M. McCarty, D. Getty, and F. Biondi. October 2017. Visual and cognitive demands of using in-vehicle infotainment systems, AAA Foundation for Traffic Safety. [502](#)
- Y. Sun, S. Wu, and I. Spence. 2015. The commingled division of visual attention. *PLoS ONE*, 10(6). DOI: [10.1371/journal.pone.0130611](https://doi.org/10.1371/journal.pone.0130611). [510](#)
- J. Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2): 257–285. DOI: [10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7). [478, 481, 764](#)
- J. Sweller. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2): 123–138. DOI: [10.1007/s10648-010-9128-5](https://doi.org/10.1007/s10648-010-9128-5).
- Tijerina, L., & Pilutti, T. 2002. U.S. Patent Application No. 10/064,979. [492](#)
- O. Tsimhoni, D. Smith, and P. Green. 2004. Address entry while driving: speech recognition versus a touch-screen keyboard. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(4): 600–610. DOI: [10.1518/hfes.46.4.600.56813](https://doi.org/10.1518/hfes.46.4.600.56813). [498](#)
- M. L. Turner and R. W. Engle. 1989. Is working memory capacity task dependent? *Journal of Memory and Language*, 28, 127–154. [499](#)
- R.-D. Vatavu. 2012. User-defined gestures for free-hand TV control. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroiTV, '12 pp. 45–48. DOI: [10.1145/2325616.2325626](https://doi.org/10.1145/2325616.2325626). [496](#)
- N. Villaroman, D. Rowe, and B. Swan. October 2011. Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor. In *Proceedings of the 2011 Conference on Information Technology Education*, pp. 227–232. ACM. [478, 780](#)
- M. Walch, K. Lange, M. Baumann, and M. Weber. 2015. Autonomous driving: investigating the feasibility of car-driver handover assistance. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 11–18. DOI: [10.1145/2799250.2799268](https://doi.org/10.1145/2799250.2799268). [514](#)
- Y. Y. Wang, D. Yu, Y. C. Ju, and A. Acero. 2008. An introduction to voice search. *IEEE Signal Processing Magazine*, 25(3): 29–38. DOI: [10.1109/MSP.2008.918411](https://doi.org/10.1109/MSP.2008.918411). [504](#)
- J. S. Wolffsohn, N. A. McBrien, G. K. Edgar, and T. Stout. 1998. The influence of cognition and age on accommodation, detection rate and response times when using a car head-up display (HUD). *Ophthalmic and Physiological Optics*, 18: 243–253. [509](#)
- C. D. Wickens. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2): 159–177. DOI: [10.1080/14639220210123806](https://doi.org/10.1080/14639220210123806). [498](#)
- T. Winograd. 1972. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language No. MAC-TR-84, MIT. [504](#)
- M. Wittmann, M. Kiss, P. Gugg, A. Steffen, M. Fink, E. Pöppel, and H. Kamiya. 2006. Effects of display position of a visual in-vehicle task on simulated driving. *Applied Ergonomics*, 37(2): 187–199. DOI: [10.1016/j.apergo.2005.06.002](https://doi.org/10.1016/j.apergo.2005.06.002). [507](#)

- K. Young and M. Regan. 2007. Driver distraction: A review of the literature. In I. J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter, and J. D. Irwin, editors, *Distracted Driving*, pp. 379–405. Australasian College of Road Safety, Sydney, NSW. 496
- R. A. Young. 2014. Self-regulation reduces crash risk from the attentional effects of cognitive load from auditory-vocal tasks. *SAE International Journal Of Transportation Safety*, 250(October). DOI: [10.4271/2014-01-0448](https://doi.org/10.4271/2014-01-0448). 499, 500
- F. Xia, L. T. Yang, L. Wang, and A. Vinel. 2012. Internet of things. *International Journal of Communication Systems*, 25(9): 1101.

# 13

# Embedded Multimodal Interfaces in Robotics: Applications, Future Trends, and Societal Implications

Elsa A. Kirchner, Stephen H. Fairclough, Frank Kirchner

## 13.1

### Introduction

In the past, robots were primarily used to perform work that was either too hard, too dangerous or simply too repetitive for humans, e.g., assembly line work, or work that could be done much faster by a robotic system, such as placement work. In the future, *human-robot interaction* will cover a much broader range of scenarios, from working interactively with humans in the context of industrial manufacturing to robotic appliances designed to care the elderly; even in applied areas, such as autonomous robots in space or operating underwater, the demand for robots to interact or to be intuitively controlled is growing. Hence, interaction will not only involve direct control of a robot or information exchange but will include direct cooperation and physical interaction between human and robot, i.e., *human-robot cooperation*. While direct cooperation has tremendous advantages it also presents a number of significant challenges that should not be underestimated. Advanced interfaces to enable human-robot cooperation will be required to meet these challenges and the needs of human-robot interaction in the future.

These interfaces must not only promote accurate and easy *explicit interaction* between humans and robots but should also enable *implicit interaction*. Explicit

interaction requires the intentional production of *active states* or commands, i.e., the production of speech or specific brain signals or gestures. On the other hand, implicit interaction operates by monitoring *passive states*, i.e., emotions, mental workload, fatigue, motivation, which arise spontaneously, or requires the interpretation of spontaneous active states passively without the *awareness* or intention of the user to improve interaction. Examples of this mechanic would be the interpretation of the user's brain signals to prepare a robot for an upcoming movement in order to enable a greater degree of coordination between human and robot or the estimation of mental workload of the user to reduce or increase interactions task-frequency (for examples see Section 13.4). Considering enhancement of awareness for explicit interaction, it can be stated that the user is responsible to increase the awareness of the robot with respect to his or her needs and the function of the interface is to translate the active state of the human into an appropriate command, which is actioned by the robotic system. In this case, the interface between user and robot is a communicator. In case of using implicit control pathways the interface must translate the active or passive state of the human into a robotic response that is both timely and intuitive from the perspective of the user. The human does not intentionally contribute to this process and may not even be aware of the underlying mechanism. The interface serves as a monitor and translator that interprets the state of the human and possible intention (see Section 13.3 for more details). Both types of interfaces, which enable implicit or *explicit control*, can be combined into a hybrid system. In case of this hybrid, interfaces between human and robot use multimodal input and generate multimodal output. Such advanced interfaces are often complex and of the type of embedded multimodal-multisensor interfaces.

While it is obvious that such interfaces serve human needs by enabling explicit and implicit interaction to optimize cooperation with respect to the needs of the human user, the interface provides two-way communication. For example, there are interfaces that enable robots to make use of human cognitive resources in those cases where they have reached the limits of autonomous behavior. Such interfaces are required for highly specialized robotic systems, which can be found in industrial environments. With respect to their specialized domain, these robots can outperform a human by contributing specialized solutions to a specific problem, but may not be able to cope with small deviations from a defined protocol. Furthermore, this type of two-way interface is required for the control of semi-autonomous robots, e.g., for teleoperation. One requirement of such interfaces is to provide feedback to the human with respect to the state of the robot. Therefore, a robotic system must either have a perception of itself, which can be transferred to the human, or

an interface to interpret the state of the robot. In this case, the interface works by either transferring the robot's state or inferring the state of the robot.

There are robotic applications that must deliver highly demanding and sensitive modes of interaction with their human users, since they have, to an extent, replaced capabilities that would usually be performed by human being, as in the care of elderly people, or robots that extend the body, i.e., provide sensory-motor functions such as exoskeletons, surgical robots, and assistive robots that aid the human user by delivering services or information. In this case, the performance demands on the human-robot interface are even higher and often transfer or interpret states bi-directionally. In summary, a combination of the aforementioned abilities of the different types of interfaces is required to develop new and advanced multimodal-multisensor interfaces that:

- provide the human operator with a greater insight into the robot's state for better control;
- provide the robot with improved insight into the intention of the human for better support or support as needed, or allow the interaction to adapt to the human's need,
- enable the robotic device to extend the human body and senses and to be used as if it were part of the human body;
- enable the robot to learn from a human in order to imitate their behavior or to learn to understand the human behavior to become a better interaction partner; and
- inherently assure the automated detection or even avoidance of malfunction and safety of interaction.

This chapter describes some measures and approaches that can fulfill the listed requirements for advanced multimodal-multisensor interfaces. While we try to give explanations and examples for all the requirements of these systems, we will focus on approaches that make use of implicit interpretation of the human state. When using implicit interpretation of the human state, the array of measures that are required to optimize human-robot interaction depends on the type of user representation that the robot requires in order to interact or to cooperate with the person. If the robot must simply avoid hitting a person or colliding with him/her, all it needs to know is where the person is located in space. No explicit interaction is required. For this situation, only a simple awareness of the user, e.g., on her or his location in space and movements, is needed. Robots that work with the elderly and must exercise soft social skills require a much higher level of awareness of

the user (and of course that means more sensors and measures). Hence, there is a relationship between the type and sophistication of the interface, its ability to interpret the human state, and the level of awareness of the user that is required for such complex interaction. Therefore, the level of awareness of the user required by the robot is hierarchical; the robot can have an awareness of the user as: (1) an object in space, (2) a co-worker or partner (what are they supposed to do, what tasks are they trying to complete), (3) an individual (gender, age, personality) and (4) a dynamic entity with respect to intentions and psychological states. Simple robots may need just (1) and (2) in order to interact or to cooperate. Robots that are designed to personalize interaction to the person would require information about stable traits of the person (3) and the means to detect dynamic changes (4). Hence, interfaces can be scaled according to the level of interaction that is required. The higher the level, the more likely it is that multimodal-multisensor interfaces are required.

The good news is that for human-robot interaction, robots are able to directly make use of a range of measures and data as part of the multimodal-multisensor interface. We will discuss the usage of *Psychophysiological measures* and how they can improve interaction and especially cooperation within this type of advanced interface. However, it is very difficult or even impossible to always interpret the state or intention of the human with one 100% accuracy. This fact is the biggest challenge in any human-machine interaction and has direct implication for the subjective perception of reliability within this interaction. If a system is deemed unreliable, it will fail to win the trust of the user. The issue of trust is particularly sensitive for multimodal-multisensor interfaces. These systems are designed to respond with a degree of autonomy, hence the user must cede a degree of control to the system. In addition, these systems monitor a range of measures related to behavior and the psychological status of the person. These data are personal and sensitive, and interaction with this type of advanced robotic system may trigger a range of societal and ethical issues around data privacy and data security (see Section 13.5). Hence, trust is multifaceted during these interactions; the user must trust in the technical proficiency of the system and be confident that their personal data is secure while they interact with the interface.

Interfaces are often not stand-alone systems in robotics. They are *embedded multimodal-multisensor interfaces* that are deeply integrated into the system's control and into the context of interaction, requiring an automated analysis of interaction context. In the future, they will develop self-adaptive properties, which require new techniques, hardware, and algorithms as discussed in this chapter. In addition, robotics presents a huge challenge for safety, especially when humans physically

interact with robots that exert high forces and accelerations together with a high net weight. Since safety of interaction is the fundamental requirement for human-robot cooperation, we will begin by discussing new and upcoming approaches from robotics to assure safety (Section 13.2). This discussion brings us directly to the definition and relevance of *embedded multimodal interfaces* in human robot interaction (Section 13.3), which often belong to the group of multimodal-multisensor interfaces. In Section 13.4 we give some application examples for embedded multimodal interfaces and explain how they can enable or improve human-robot interaction. Finally, in Section 13.5, future trends in embedded multimodal interfaces and societal implications are discussed.

For a detailed application scenario for an embedded multimodal interface in robotics, see the [Glossary](#) and [Focus Questions](#).

## **13.2 Inherently Safe Robots—a Prerequisite for Human-Robot Cooperation**

To assure safety in human-robot interaction, the most intuitive approach is to make robots inherently safe. Safety during human-robot cooperation is not only an additional benefit but often an indispensable criterion to enable sharing of common spaces or ensuring physical collaboration between fragile humans and powerful robots. Safety can be implemented on different levels during robot design to enhance reliability. Inherent safety in human-robot cooperation is achieved through a three-level process. On the lowest level (level 1), safety is ensured directly by the design of the electro-mechanical hardware. Therefore, this level is also referred to as the *safety by design* level: we can distinguish three parallel paths on this level (see Figure 13.1). First, the most straightforward path is the process of mechanical design itself, such as those classic lightweight designs that are standard in robotics. However, in recent years new smart materials can further enhance the lightweight design ethos. This process is enhanced by recent advances in 3D-printing technologies, which allow mechanical design to go in directions that were unthinkable using classic technologies, like embedding electronic circuitry and signals into the structure of the components. The integration of channels with complex 3D structure into the components, e.g., internal cabling, represents another approach. These developments are paralleled by advances in the design of robotic actuators. Using this approach compliant elements are embedded into actuators by serializing, e.g., a spring with a motor (and gear), which complicates the control of such an actuator on the algorithmic level, but enables built-in safety as external forces that act on the robot are absorbed by the spring - instead of the

### Glossary

**Active BCI** is a brain-computer interface that derives its outputs based upon a voluntary act of explicit control from the human, e.g., generates motor imagery consistent with movement of right hand to move cursor to the right.

**Active state** is a psychological state associated with a volitional act or intention generated by the person, e.g., to open the door.

**Application-specific safety level** describes the concept to include information into the robot's behavioral control that comes from sensors that are not part of the robot itself. This concept uses the information from sensors that are placed outside of the robot to monitor the environment, e.g., a workplace in a production line and which are used in more traditional applications to create strict safety boundaries around the workplace. In more advanced approaches this information is used differently; here it helps to derive contextual information that can be used to adapt the robot's behavior instead of overwriting it. For example, in a more traditional scenario a violation of the safety boundary by a human walking by would result in a full stop of the production line. The more advanced concepts would predict the humans path and instead of stopping the production line would only reduce the speed of the moving robots. This concept therefore modifies the behavior of the robot as commanded by the *high-level control* module instead of overwriting it.

**Awareness** can refer to perceiving sensory stimuli in the environment including other actors, which may be machines or peoples.

**Biocybernetic control** describes a model of *closed-loop control* (negative or positive control) wherein measures are derived from psychophysiological or neurophysiological sources and converted into control input for an adaptive computer system.

**Closed-loop control** is a control system that uses the concept of an open-loop system as its forward path but has one or more feedback loops (hence its name) or paths between its output and its input.

**Covert measure** is a measure of human behavior or performance that cannot be detected based upon human perception, e.g., heart rate, and brain activity.

**Embedded brain reading** is an approach for user state detection, which is based on the online analysis of brain activity. Brain activity is used which is spontaneously evoked during human-machine interaction. The approach is deeply embedded into the system's control, the context of interaction, and makes use of multimodal data. It is applied for implicit interaction, i.e., to non-intentionally adapt or drive explicit interaction.

**Embedded multimodal interface** is an interface that makes use of **multimodal data** from **multimodal input** and is able to generate **multimodal output**. Its main characteristic is that it is deeply incorporated into the control of the robotic system, and may be subject to complex adaptation mechanisms such as *reflexive adaptation*. While its function might be to gain explicit control of a system, it might be subject to implicit control to be adapted to the human's or system's needs.

**Glossary**

**Explicit control** represents a mode of input control where the user intentionally generates a specific behavior in order to achieve a specified goal, e.g., move a cursor upward.

**Explicit interaction** is a mode of human-computer interaction where the human user is fully cognizant of the issuing of commands and receives explicit feedback from the computer.

**High-level control** refers to the specification and feedback control of target positions in 3D space that must be reached by the end effector of a multi-joint robot based on information coming from sensors sampling the robot itself and its environment.

**Human-robot cooperation** is a subfield of human-robot interaction where a human and robot or teams of humans and robots work or act together to reach a shared goal. It often requires direct contact between human and robot or a shared workspace.

**Human-robot interaction** is any interaction between a human and a robot or teams of humans and robots including communication, control, feedback, direct contact, or information exchange.

**Hybrid BCI** describes a brain-computer interface that combines *active BCI* with either *passive* or *reactive BCI* or other measures such as eye movements or heart rate.

**Imitation learning** enables a robotic system to learn from demonstrations of nearly optimal policies executions given by a teacher (e.g., a human mentor). It is often used to initialize reinforcement learning to avoid time consuming learning from scratch.

**Implicit interaction** is a mode of human-robot interaction where the human user is not aware of the issuing of (control) commands that may be used for the control of a technical system or adaptation of an interface to the needs of the technical system or user. The user may or may not receive explicit feedback from the computer.

**Internal state of a robot** is computed on the basis of all sensor information directly or indirectly available to the robot. Directly available information is all information that comes from the robot's own sensors, while indirectly available information is all information that comes from sensors that are external to the robot but that the robot can access through communication pathways. The set of internal states of a robot is in most cases a finite set of eventually multi-dimensional vectors. Elements of this set are computed through means of clustering that range from simple thresholds to complex statistical methods.

**Low-level control** refers to the direct feedback control of movements of the motors in the joints of a multi-joint robot using sensor information coming directly from the individual motors to reach a specified position in 3D space.

**Glossary (continued)**

**Neurophysiological measures** represents the act of measurement based on physiological activity from cerebral sites in the human brain. These measures may be based upon electrical activity (electrocortical, electroencephalographical) or neurovascular changes (functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS)).

**Overt measure** is a measure of human behavior or performance that can be detected based upon human perception, e.g., voice commands, gestures. A

**Passive BCI** is a brain-computer interface that derives its outputs from arbitrary brain activity without the purpose of voluntary control.

**Passive state** is a spontaneous psychological state that arises during behavior without an intention on the part of the person, e.g., fatigue, frustration.

**Physiological computing** refers to a field of research in human-computer interaction wherein **Physiological measures** derived from the human user are used as a source of input control for a computer system or interface.

**Physiological measures** describes the act of measurement based on processes related to human physiological functions.

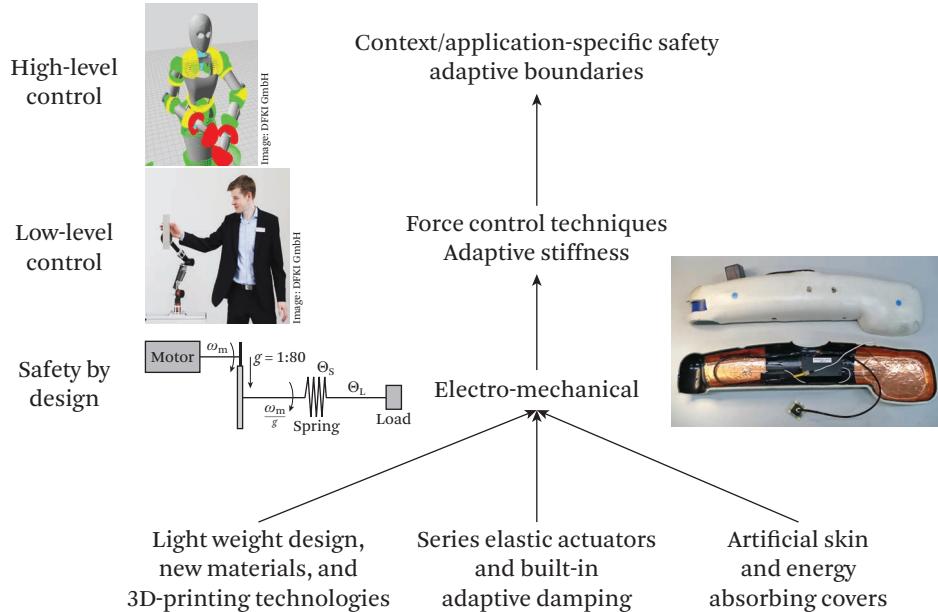
**Psychophysiological measures** describes the act of measurement and inference wherein psychological processes and concepts are inferred on the basis of physiological measurements from the autonomic nervous system.

**Reactive BCI** is a brain-computer interface that derives its outputs from brain activity arising in reaction to external stimulation, e.g., a visual stimulus or sound.

**Reflexive adaptation** refers to a second-order process of adaptation whereby the computer makes an autonomous response and subsequently monitors the response of the human user to that response in order to inform future responses.

**Safety by design** refers to the fact that next-generation technical systems for human-robot cooperation will include, e.g., a compliant element in their actuators that absorbs energy. Thereby safety is an integral part of the mechanical construction of the system.

**Temporal cascaded approach** is an approach of using multimodal data in a timely sequenced fashion where the usage and outcome of analysis of one data type influences the analysis or choice of a second- or higher-order data type.



**Figure 13.1** Safety levels for human-robot cooperation.

human body as in classical soft robots. Combining this approach with a high speed / high precision torque control technique on the algorithmic level [Bargsten and de Gea Fernández 2015] results in a robotic system that can be designed to be inherently safe (see video referred in Figure 13.2).

Finally, the level of safety by design can be realized by providing the new generation robots with a sense of touch. Touch or any kind of haptic information for robots has been largely ignored by robot designers over recent decades because the requisite technology was not available. Instead, vision was the dominant sensor for robotic perception and the primary means of avoiding contact with the user. Yet robots that directly interact with humans—e.g., while building or installing the windshield of an automobile on a shop floor of a car manufacturer—cannot completely avoid physical contact. Furthermore, physical contact may be unavoidable if a tool is handed over. This aspect of the interaction is where the third line of safety by design comes into action—sensor systems that sense humans via multiple modalities, not only by visual sensors but also by other sensors such as touch and distance. For example, artificial skins can be implemented because material technology and electronic circuitry have achieved a level of miniaturization and mechanical flexibility that allows us to design surfaces for the new generation of



**Figure 13.2** **Video:** The COMPI arm: compliant trajectory tracking via model-based control allows safe human-robot interaction. <https://youtu.be/mDODMNMC5zc>

robots that can feel [Lucarotti et al. 2013]. This is a great tool to enhance the safety of human-robot interaction, but more importantly, it allows the control programs and robotic architectures to integrate an increased range of modalities when generating a perception of the robot’s environment and the interacting human or the internal state of the robot based upon pure sensor data [Kampmann and Kirchner 2015].

The next level of the hierarchy (level 2: low-level control; see Figure 13.1) can be related to the robot control regime that plays an important role in human-robot interaction. The systems within the new generation of robots are growing in complexity at a rate that is comparable to that of chip complexity growing according to Moore’s Law [IntelCooperation 2005]. The low-level control (level 2 of our hierarchy) is realized in state-of-the-art robots by an approach to control actuators using torque and force rather than position [de Gea Fernández and Kirchner 2015] as the main control signal. The advantage of this control regime with respect to safety is obvious and refers to the fact that a position-controlled robot simply moves to a predefined position regardless of objects blocking its path, while a force-controlled robot also moves toward the desired position in 3D space but only as long as no external forces are encountered—forces like a collision with an object. The robots under force control regime would immediately (fractions of a second) reduce the torque on the appropriate joint resulting in an immediate

stop. As soon as the external forces are gone the robot will continue on its path. In some more advanced cases the robot would in fact search for an alternative path or trajectory.

The last level in the hierarchy (level 3), also referred to as high-level control, is implemented based on sensors, internal of the robot, that are used to describe the state of the external environment of the robot [Lüth et al. 2015], e.g., cameras, lasers, or Time-of-Flight cameras. This information provides the basis for planning of manipulator arm trajectories and robot navigation paths and includes aspects of safety by avoiding to hit any obstacles (as well as humans) and describes a standard in robot control. However, these sensors can be combined with sensors that are external to the robot and that are very specific to the concrete application, e.g., a shopfloor production assembly line. There are many examples like: external overhead cameras [SafetyEye 2014] or laser range finders that are usually implemented to create safety boundaries around the robot. In traditional robot applications a violation of these borders just results in an alarm and a shutdown of the production line. In more advanced applications this information is used to adapt the behavior of the robot in case the boundaries mentioned above are violated by an object or a human [de Gea Fernández et al. 2017].

Safety boundaries are not static in modern robotics, but can be adaptive and vary with the context of the robot's task and application [Vogel et al. 2013], e.g. the robot would not go to a full stop but rather slow to a predefined speed in a human-robot cooperation scenario [de Gea Fernández et al. 2017]. Because the kind of adaptation of the robot is dependent on the context of the task, applying this adaptive type of safety level can also be seen as a context or application-specific safety level (see also Haddadin [2015]).

In summary, there are approaches to enhance safety that are inherent in the design of a robotic system. These approaches provide the robot with a good perception of the environment but do not necessarily require a concrete understanding of the human state or intention, or even recognition of the human apart from other objects in the environment. However, those internal states of robots that exist to enhance safety can also be the basis for the creation of more complex forms of awareness to support the interaction with the user. Hence, advanced interfaces that make use of multimodal data to enable explicit and implicit interaction with humans (see Section 13.3) must not only focus on establishing a representation of the user state but must also encompass a description of the status of the robot. There are simple mechanisms to present the state of the robot to a human operator, e.g., written or colored light feedback [de Gea Fernández et al. 2017], video data feedback from the point of view of the robot (i.e., its internal cameras), 3D reconstruction of

the robot in its environment, or force feedback which conveys an impression of the robot's tactile perception during its interaction with the environment.

These modes of feedback are able to communicate information about the internal state of the robot. For certain tasks, such as teleoperation, continuous and clear feedback from the robot provides the user with the means to achieve easy control of the robot. In Section 13.4 we give some more examples for different applications, and in Section 13.3 we focus on the categorization of different approaches for utilizing human states for human-robot interaction and its relevance for the development of advanced embedded multimodal interfaces.

## 13.3

### Definition and Relevance of Embedded Multimodal Interfaces

The basic goal of interfaces is to provide the robotic entity with a quantification of the state of the human user [Schuller \[2018\]](#) for an overview on multimodal user state recognition) and to provide the human with feedback regarding the state of a robotic system. The purpose of the interface is to generate bi-directional awareness and to enable bi-directional interaction and/or the support of the interacting human or robot. Whereas perceiving the environment via multiple modalities is very natural for more complex biological systems, the sensory capabilities of many technical systems are often limited to one modality. This modality is usually used in one way, i.e., to intentionally transfer commands or to actively perceive objects that are relevant for the system's action. In the past, this restriction of modality limited the possibilities for generating a representative level of (bi-directional) awareness. However, as pointed out in the previous section, this situation is now changing in the field of robotics. Technical systems can be equipped with different sensing modalities, which can be utilized for different purposes, with respect to both multimodal input and output [\[Kirchner et al. 2015\]](#). This technical progress has been created by the increased availability and ease of usage of sensor technology, enabling robotic systems to receive a variety of data about their environment and users [\[Kampmann and Kirchner 2014\]](#). This innovation requires the development of advanced multimodal-multisensor interfaces.

Let us give some examples with regard to the possible approaches that improve the interaction between human and robot using multimodal interaction. A camera can monitor the spatial position and body posture of the person and can capture any movement in space. But modern camera technology is also capable of monitoring information about facial expression and heart rate via a webcam [\[Monkaresi et al. 2014\]](#). In a similar way, a conventional microphone can record sounds and utterances from the user, but is also capable of capturing those emotional responses that

are inherently part of vocal expression [Bachorowksi and Owren 2010]. Both can thus be used to record psychophysiological measures of the human. Psychophysiological techniques grant technology access to signals from the autonomic nervous system via wearable sensors, which allows the robot to make inferences about the psychological state of the user. For example, changes in heart rate or galvanic skin response represent the level of psychological activation experienced by the individual; increased levels of frustration or anxiety or excitement are associated with higher psychological activation. The availability of wearable devices to measure electrocortical [Nijboer et al. 2015] and neurovascular activity [Piper et al. 2014] from the cortex would allow a robotic system to draw inferences about high-level cognitive states experienced by the user, such as intentionality, mental workload and skill acquisition [Bozinovski and Bozinovski 2015, Canning and Scheutz 2013, Kirchner et al. 2016a].

When developing a taxonomy for multimodal interfaces, it is convenient to classify techniques to monitor the status of the user into techniques that use *overt measures* and *covert measures*. The former refers to methods that record and infer on the basis of what could be seen or heard by a hypothetical (human) observer. These overt methods are designed to record movements, changes in facial expression, and vocal utterances, or the same approach can be used to capture behavioral indices such as performance or task activity. Covert methods represent those measures from the user that are imperceptible to the hypothetical observer, such as psychophysiological and *neurophysiological measures*. Certain data types, such as electrocortical activity, are invisible to the human eye, others may be perceived visually (e.g., pupil dilation) but cannot be accurately assessed in real-time by a human observer. Furthermore, a robotic system can use overt and covert measures to assess two broad categories of user state: (a) active states that represent intentionality, preparation for action and movement; and (b) passive states, such as emotions, mental workload, fatigue, and motivation, which arise spontaneously as a consequence of human-robot interaction. Table 13.1 provides two examples that capture the distinctions between overt/covert measures and active/passive state categories.

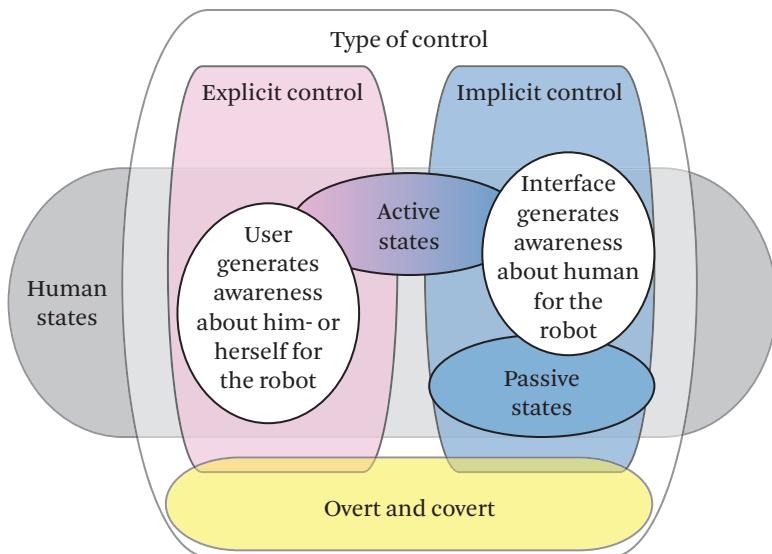
On the other hand, when classifying multimodal interfaces from a usage perspective, it becomes obvious that there is a substantial overlap between explicit or implicit interaction (see Figure 13.3). For example, when developing a speech interface for explicit control, both covert and overt measures can be used. The natural choice is to monitor the overt auditory output, but covert muscle activity, i.e., the human electromyogram (EMG), recorded by electrode arrays, can be used for such an interface [Wand et al. 2013]. The same is true for implicit control. Overt changes

**Table 13.1** Examples of active and passive states monitored using overt and covert measures.

	Active	Passive
OVERT	user moves right hand	facial expression indicative of surprise
COVERT	increased activity in somatosensory cortex during preparation to move hand	elevation of heart rate and skin conductance level

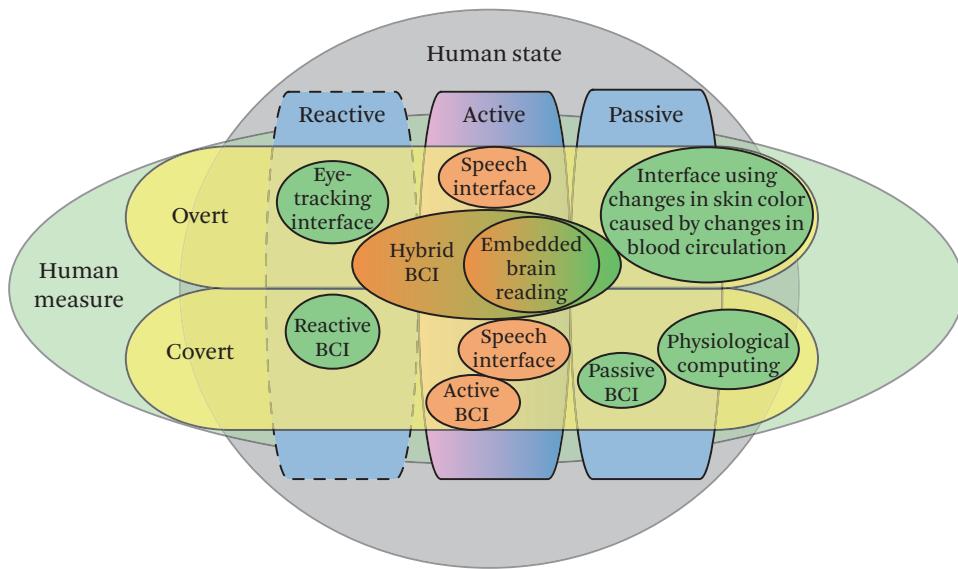
in facial expression can be used to adapt an interface with respect to the emotional state of a user. Covert changes in skin conductance can be used to capture emotional activation. Thus, covert and overt measures are applicable for explicit and implicit control purposes. The same principle applies to the detection of those active states exhibited by the user. An active state can of course be used for explicit control. However, an active state, such as the preparation of a movement, can also be used for implicit control, i.e., to adapt a robotic system for likely upcoming movements to reduce interaction forces (see Sections 13.4 and 13.6). These examples emphasize that it is important to clearly state whether a taxonomy of interfaces is based on the techniques used to monitor the status of the user or the functionality of the multimodal interface. It is important that there is an awareness of both possibilities to avoid misunderstandings regarding the purpose of the interface.

If we consider how these techniques can be used to monitor the status of the user, a question arises concerning how various methodologies can be combined to create a dynamic and complex representation of the user state in order to enable implicit or explicit control. Brain-computer interfaces (BCI) [Wolpaw et al. 2002, Brunner et al. 2014] that often make use of the electroencephalogram (EEG) provide an interesting case for consideration. BCI technology is generally understood in its active form wherein neurophysiological correlates of voluntary control are used as an explicit control input to a robotic system (see *active BCI* in Figure 13.4). For example, BCI systems can use signals derived from motor imagery to intentionally direct the movements of a humanoid robot [Yongwook et al. 2012] or exoskeleton [Barsotti et al. 2015]. However, there are at least two other types of BCI that can be applied to robotic systems [Zander and Kothe 2011]. Reactive BCI describes a system where changes in brain activity in response to an external stimulus drive the output of the system. This type of BCI is activated by the neurophysiological response to a sensory event, e.g., an evoked-cortical potential (ERP) or steady-state visually evoked



**Figure 13.3** Type of control and the relation to human states and human measures.

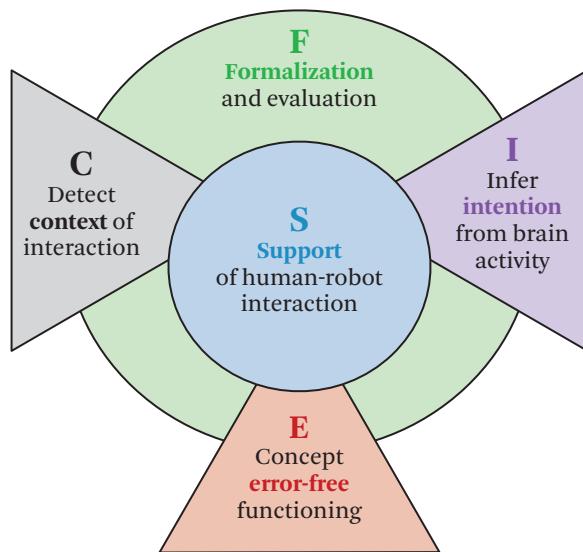
potential (SSVEP), rather than the intention to act [Zander et al. 2014]. A reactive non-BCI interface which makes use of overt measures is, for example, a reactive eye-tracking interface (see Figure 13.4), which activates a specific action whenever the user is looking at a specific object, part of a robot, or area on a screen. The passive category of BCI (see *passive BCI* in Figure 13.4) describes a system where outputs are derived from changes in brain activity related to spontaneous changes in psychological states, such as: mental workload, frustration, anxiety, fatigue, etc. This ‘passive’ type of BCI is identical to the concept of *biocybernetic control* from *Physiological computing* [Fairclough 2009]. It is possible to create hybrid forms of BCI where active, reactive, and passive forms are used in conjunction to enhance the speed and fidelity of control [Müller-Putz et al. 2015]. For example, passive states of fatigue and frustration may affect the ability of the user to produce motor imagery, which translates into impaired control of an active BCI [Myrdan and Chau 2015]. Cotrina et al. [2014] presented a *hybrid BCI* wherein passive state measures of electroencephalographic frontal asymmetry, a measure that is associated with emotion and motivational disposition, were used to refine the reactive response from an active BCI based upon SSVEP. Hence, active and passive state measures of the user are situated within a data space in order to improve fidelity of active con-



**Figure 13.4** Examples of interfaces and their relation to human states and human measures.

trol over a robotic system. This data space can be further extended by adding overt measures. For example, [Kim et al. \[2014\]](#) designed a hybrid BCI that combined covert electroencephalographic data with overt eye movement data to navigate a quadcopter in 3D space. Hybrid BCIs can therefore be multimodal interfaces and can be used for explicit control of a technical system. They may also have integrated implicit control designed to optimize the multimodal interfaces and improve the degree of user control over the technical system. Also, if we consider the earlier example of the adaptation of an exoskeleton's control for teleoperation, this can be seen as a hybrid BCI: The user's overt arm movements while covert electroencephalographic data is used to non-intentionally and implicitly adapt and improve the explicit interaction between the human and the exoskeleton.

When considering the use of embedded multimodal interfaces, the meaning of the word "embedded" should be explained. The measurement of psychological concepts associated with the user state is generally enhanced by consideration of task context, e.g., the type of task being performed, task criticality, difficulty, duration, etc. The monitoring capability of multimodal interfaces encompasses task models and related variables in order to monitor the psychological state of the user within a specific task context. Thus, interfaces that make use of psychological concepts must be embedded into the task context. Therefore, the automated analysis



**Figure 13.5** Concept for embedded brain reading for the support of human-machine interaction based on the context of interaction and inferred intentions. (Figure courtesy of Kirchner et al. [2015])

of the interaction context is of tremendous relevance for embedded multimodal interfaces [Kirchner and Drechsler 2013]. On the other hand, interfaces that improve human-robot interaction must be embedded into the control of the robotic system in order to achieve and sustain safe operation [Kirchner and Drechsler 2013]. This does not necessarily mean that all data processing must be performed “on board” or that no external sources of information or processing power can be used (such as cloud-based solutions). It does imply that all processing that can only be done at the place of generation should be embedded into the system.

Figure 13.5 provides an example of embedded brain reading, a sub-type of embedded multimodal interfaces. Embedded brain reading includes the analysis of brain activity such as the EEG recorded from the surface of the head. To enable the interpretation of brain activity, embedding the analysis into the context of interaction and thus task state, human state, or spatial state is highly relevant. Without this processing module, it would nearly be impossible to interpret the user’s brain activity to infer his or her intention. While context is a far-reaching term, it is used here very general. Context refers to the state of interaction in the broader sense (e.g., space, task, human state, environmental state, system state, etc.). The required interfaces are often not stand-alone systems, as a mouse or

a keyboard are, but they are part of the robotic system that make use of input from internal and external sensor systems as well as sensors that are worn by the interacting human [Hung et al. 2015] designed to capture physiological measures from the body and neurophysiological measures from the brain. For example, today, exoskeletons are commonly equipped with gravity compensation [Lewis et al. 2003]. This is an algorithm which allows control of the exoskeleton so that the user does not feel the weight of the exoskeleton. This feature is most relevant for the control of distal body parts, i.e., the human arm, since the system's weight would otherwise be too high to allow extended usage, such as during teleoperation [Mallwitz et al. 2015] or for exoskeleton-assisted rehabilitation [Kirchner et al. 2013a, Kirchner et al. 2016c]. Forces can be redirected to, e.g., the hip or in case of a whole-body exoskeleton also to the ground on which the user is standing on. This control does not require knowledge about the intention of the user. However, an additional approach can be applied that enables the exoskeleton to also carry the weight of the arm. For example, gears can be locked to keep the arm in a certain position [Folgheraiter et al. 2012]. To release the locked position the system must know whether the user wants to move again. This knowledge can be gained from the analysis of the user's brain activity. However, to infer a movement intention is only relevant while the exoskeleton is keeping the arm fixed in a certain position. Thus, the exoskeleton "knows" from its own control state when to consider brain activity analysis to detect a change in the state of the human (see further explanation in Sections 13.4 and 13.6).

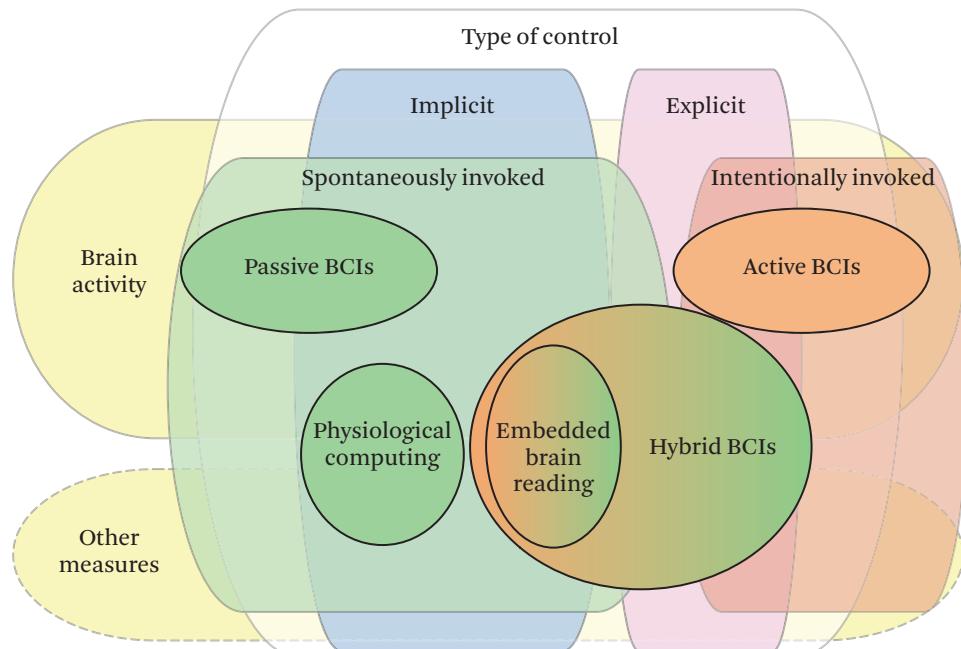
When using complex data such as covert brain activity to infer intentions it must be considered that the outcome of analysis can be incorrect. Therefore, embedded brain reading only considers approaches that are fault tolerant, i.e., will not lead to a malfunction of the whole embedded multimodal interface (see Folgheraiter et al. [2012], Kirchner et al. [2014]). Furthermore, methods are applied that allow formalization and evaluation of the implementations not only to estimate and measure quantitative and qualitative improvement [Folgheraiter et al. 2012] but to also verify correctness [Kirchner and Drechsler 2013].

In general, embedded brain reading describes the use of active and passive human states for explicit or implicit interaction, i.e., to non-intentionally adapt a robot or its interface with respect to an active or passive state to usually improve explicit control. The same approach can also be applied for explicit control purposes only. For example, in rehabilitation robotics, embedded brain reading can be applied to "drive" an exoskeleton to support or execute the intended movements of the patient. Spontaneously generated brain activity, which depends on the recording method (extracranial or intracranial), may not always be sufficient

to control the exoskeleton in 3D. Other data such as muscle activity or eye movement data may need to be combined for an effective explicit control. Moreover, in the given example the exoskeleton's control can further be adapted to improve interaction. For example, the strength of support by the exoskeleton needed by the patient can be adapted based on an "assist-as-needed" approach [Kirchner et al. 2016b]. Such an adaptation can be achieved by directly measuring the force the patient can still exert or the strength of electromyographic signals of the supported limb or body part which can be recorded during interaction. The main goal is to combine multimodal data, such that the intended interaction or behavior can be supported best [Folgheraiter et al. 2012, Kirchner et al. 2013a, Kirchner et al. 2013b, Kirchner et al. 2014]. However, embedded brain reading can also be applied to infer the user's passive neurophysiological state, such as their current workload or task load (see also [Volume 2, Chapter 10]), to adapt an interface for explicit robot control in such a way that the user is neither stressed nor bored [Kirchner et al. 2010, Kirchner et al. 2013b, Wöhrle and Kirchner 2014, Kirchner et al. 2016a] which would have negative impact on both the quality and quantity of interaction.

For embedded brain reading only brain activity is used which is spontaneously evoked (see Figure 13.6). Further, the approach is designed to interpret brain activity dynamically during interaction. Relevant data may also include spontaneous changes in brain activity in response to an external stimulus as used in reactive BCIs. To use intentionally evoked brain signals as it is often the case for many active BCIs, where, for example, the imagination of right- and left-hand movements can be used to spell a word [Blankertz et al. 2006], requires the attention of the user. Such attentional effort would require too many resources, which is one reason why classical BCIs are often considered to be inadequate for robot control in complex applications, such as, for example, in space applications. Figures 13.4 and 13.6 illustrate the interrelationship between different interfaces and embedded brain reading.

In summary, the basic goal of embedded multimodal interfaces is to provide the robotic entity with a quantification of the user state that enables easy explicit control of the robot and allows for implicit control of the robot or its interface. Such interfaces also enable the system to make use of the cognitive capabilities of the human or to enable the robot to learn from the human [Kirchner et al. 2015] (see Section 13.4). The availability of overt and covert measures to capture active and passive states (Table 13.1) allows the multimodal system to construct a dynamic representation of the user that is both sophisticated and scientifically valid as discussed in this chapter.

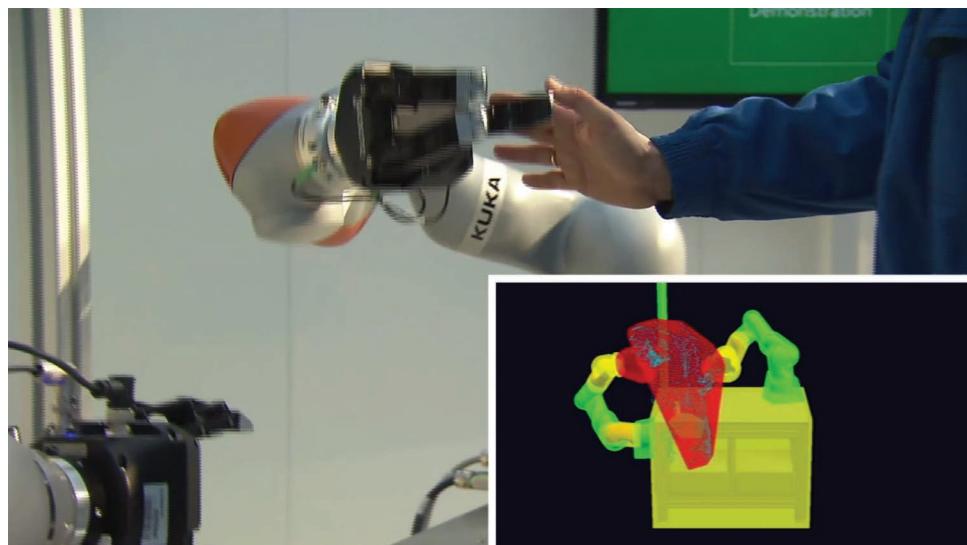


**Figure 13.6** Usage of spontaneously and intentionally evoked brain activity.

## 13.4

### Embedded Multimodal Interfaces in Robotic Applications

This section will explain the advantages of embedded multimodal interfaces and how they would function in the context of robotic applications. First, we give examples of multimodal-multisensor interfaces from two perspectives, i.e., robotics and physiological computing, and explain where and what the purposes of such interfaces usually are in both fields. Later, we give examples that focus on specific applications in robotics. We explain how both approaches, driven by the robotic control view and driven from the perspective of human state analysis, can be combined in a temporally cascaded fashion to: (1) make use of overt and covert measures to ease explicit control of robotics systems or to enable implicit control that adapts an interface or robot to improve human-robot cooperation within the same application. Further, we provide examples of how embedded multimodal interfaces will (2) improve bi-lateral awareness using covert human measures and multimodal data of the robot. Moreover, we will (3) explain how the usage of multimodal overt and covert measures allows us to increase the level of awareness.



**Figure 13.7** **Video:** iMRK: an embedded multimodal interface for human-robot interaction.  
<https://www.youtube.com/watch?v=VoU3NbTyFtU>

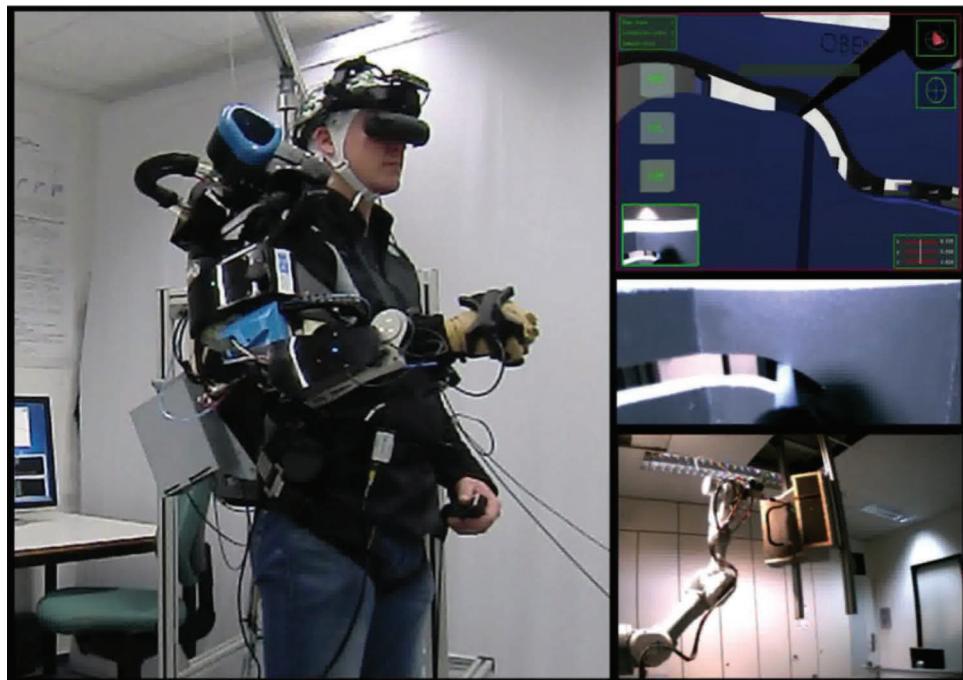
One highly important reason for applying embedded multimodal interfaces is to enhance safety and reliability. For example, a human must be detected when he or she has penetrated the work space of an industrial robot. Different sensors on the robot or in the environment, such as lasers, 3D depth-image camera, or motion sensors (that are attached to the human or human's clothes) can be combined to detect the location of the person and predict their path [de Gea Fernández et al. 2017]. Furthermore, the weaknesses of one sensor, i.e., a restricted range, can be compensated by another to improve the recognition of a human entering the work space. A combination of sensors enables accurate predictions to interpret the situation and the user's intention. The same sensors can also be used for explicit interaction, e.g., to intentionally command the robot to stop moving or to hand over a workpiece (see the video referred to in Figure 13.7). In most cases, approaches in robotic control are designed to enhance safety and reliability, and make use of overt measures to analyze the state of the user.

There are advantages to measure a psychological concept using two or more different indices from the perspective of physiological computing. Certain psychological states may be described as many-to-one [Hettinger et al. 2003], i.e., a number of measures are required in order to represent a single psychological concept. This approach also represents a form of convergent validity wherein multiple measures

are collected simultaneously in order to derive a composite score based on the degree of correlation or coherence between different measures. Therefore, the capacity of multimodal interfaces to encompass different measures allows the robotic system to monitor the user in a way that is scientifically valid. For example, Bekele and Sarker [Cacioppo et al. 2000] constructed an adaptive mode of human-robot interaction where task difficulty was dynamically adjusted in order to keep the user engaged with the task. This system combined physiological measures, i.e., covert measures, from the cardiovascular system, electromyography, and skin conductance to capture the level of task engagement exhibited by the user. The same group applied a similar approach to measuring emotional responses to create robotic interventions for children on the autistic spectrum [Liu et al. 2008] (see also Volume 3, Chapter 13.).

Making use of both covert and overt measures goes beyond classical control approaches in robotics. With the help of the application shown in the video referred to in Figure 13.8 we explain how such an approach, based on overt and covert measures, enables improvement of human-robot interaction, i.e., explicit control by non-intentionally adapting the robotic system with respect to the active human state and by applying implicit control within the same application. In this application, a user is teleoperating a robotic system by means of an active exoskeleton [Folgheraiter et al. 2012]. An active exoskeleton is a robotic system that is worn by the human and thus is in direct physical contact with the human. It is both an interface as well as a complex robotic actuator. To enhance transparency for the wearer the interaction between user and exoskeleton in this example is supported by embedded brain reading [Kirchner and Drechsler 2013, Kirchner et al. 2013b], an embedded multimodal interface that is making use of covert electroencephalographic data recorded from the scalp. Based on the analysis of the user's EEG and his or her behavior, transitions between (tele-)operation modes (see Figure 13.9) are supported (see Section 13.6 for details).

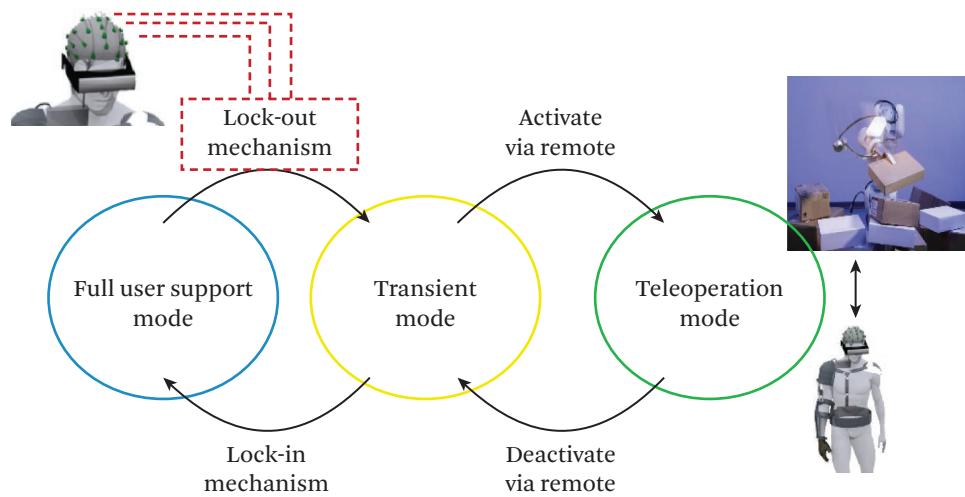
But why is covert brain signal data *not* directly used for explicit control, i.e., explicit change of operation mode? Even when using advanced signal processing and machine learning methods to infer movement intention from brain signals, state detection is inaccurate due to the complexity of the EEG signal and due to the fact that similar brain signals are generated when a human is only imaging a movement or does indeed prepare the movement to be executed. Therefore, the outcome of EEG analysis is in the given example not used to intentionally control the change between the modes. It is instead used to enhance sensitivity of the sensors that detect the movement onset, with the result that interaction forces are reduced and the operator can clearly feel the enhancement in transparency of the exoskeleton (see Section 13.6 for more details on approach and evaluation).



**Figure 13.8** **Video:** VI-Bot: an embedded multimodal interface for robot control via teleoperation with an exoskeleton and a virtual environment. [https://www.youtube.com/watch?v=3RhcgvRz\\_O8](https://www.youtube.com/watch?v=3RhcgvRz_O8)

Moreover, the implemented embedded multimodal interface does not only enhance transparency but is highly reliable and safe for use. The user is always able to unlock the exoskeleton by pressing against the sensors even in a case where movement planning was missed. Furthermore, adaptation of the exoskeleton is changed every 50 ms which made it unlikely that a false positive will lead to an unwanted lock-out. During all our online tests with real online movement prediction, the user never experienced a false lock-out using this system. By combining two modalities, i.e., covert brain data and overt movement data, in a strict temporal order the interface represents an example of a temporally cascaded multimodal interfaces, which will be explained later.

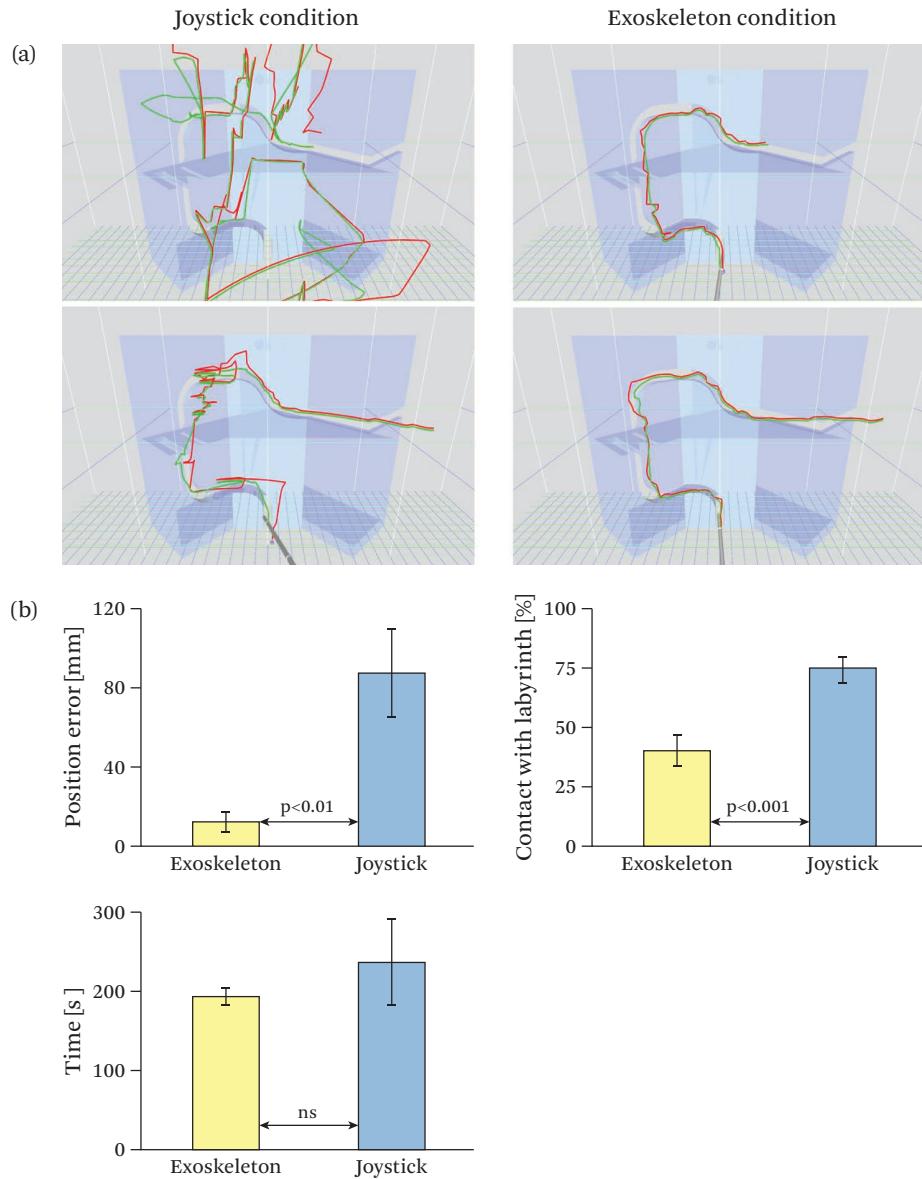
Besides enhancing safety, reliability, and bilateral interaction, an embedded multimodal interfaces enhance bilateral awareness. Different modalities can be used to provide the user with a greater insight into the state of a robotic system, i.e., to improve awareness of the robot's state. For example, in the scenario which was developed for a teleoperation application (shown in the video referred to in



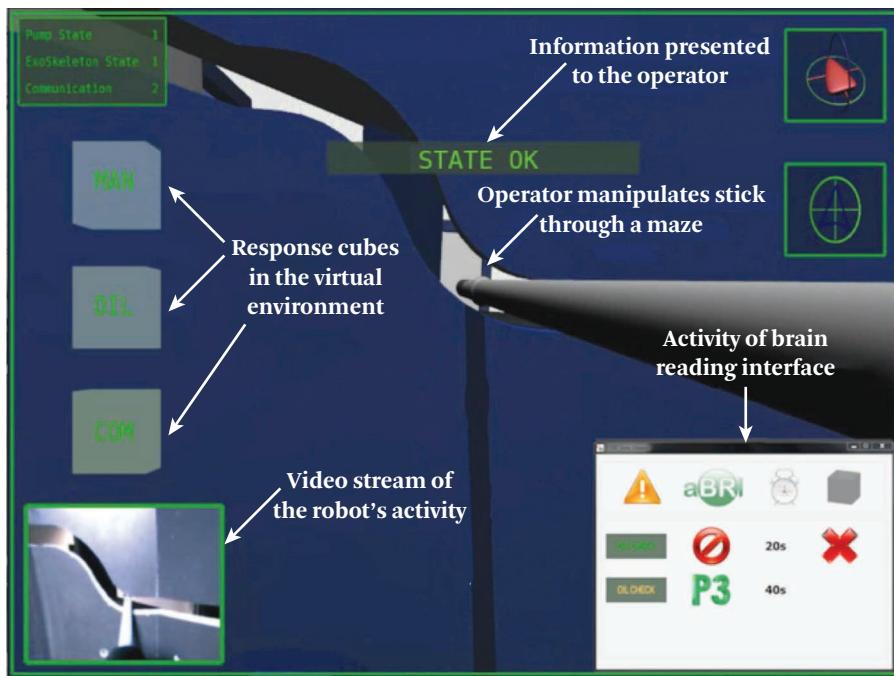
**Figure 13.9** Exoskeleton states: teleoperation mode, transient mode, full user support mode. Lock-out mechanism is activated by overt measures of an active human state (arm movements detected by the exoskeleton). The activation of this mechanism is adapted by covert measures of an inferred active state change (from no movement to movement) by analysis of the human's EEG, i.e., detection of activity that correlates with movement preparation. (Figure courtesy of [Folgheraiter et al. 2012])

Figure 13.8) the user is not only supported by the exoskeleton for explicit control. He or she is further virtually immersed into the situation of the robot by using 3D simulation and a head-mounted display while being able to make use of video material. Moreover, the user receives tactile force feedback from the robotic system via the exoskeleton. The combination of virtual immersion, mapping between the human and the robot's movement, and force feedback allows the user to become immersed into the situation and virtually feel what the robot feels. By means of the embedded multimodal interface, the user becomes strongly aware of the robot's states and changes in state, which eases explicit control and reduces interaction errors, such as failures in path following (see Figure 13.10).

Furthermore, the user is monitored whether he or she is indeed aware of the state of the exoskeleton and the display of important messages that are visually presented to the user. This facility is achieved by adapting the embedded multimodal interfaces based on the predictions that are made about the success of the user in recognizing these relevant messages. The adaptation of the display of messages is again based on online brain-signal analysis. Signals in the EEG are detected



**Figure 13.10** Comparison of accuracy in path following between joystick and exoskeleton control in the VI-Bot scenario. (A) Example of two subjects steering the robotic arm through a 3D maze. Green line: path corrected in the virtual environment. Red line: theoretical path without correction. (B) Behavioral analysis of nine subjects. Accumulated position errors (left), percentage where path contacted the wall of the maze (middle), and measured time for a complete sweep through the labyrinth (right). (Figure courtesy of Straube et al. [2011])



**Figure 13.11** **Video:** VI-Bot - virtual immersion for holistic feedback control of semi-autonomous robots. Shown is the implicit control approach implemented to adapt the embedded multimodal interface to enhance the awareness of the user for relevant information on the robotic system based on EEG signals. <https://youtu.be/8WEVZz6bpJU>

that allow the system to infer whether the user recognized the presented messages. If those signals are detected, information is not repeated for a longer time, since it is expected that the user will respond to the message. In case the relevant brain activity cannot be detected, the message is repeated instantly and at the same time highlighted to make the user more aware of the relevant information (see video referred to in Figure 13.11). The chosen approach is an implicit control of the interface [Kirchner and Drechsler 2013, Wöhrle and Kirchner 2014].

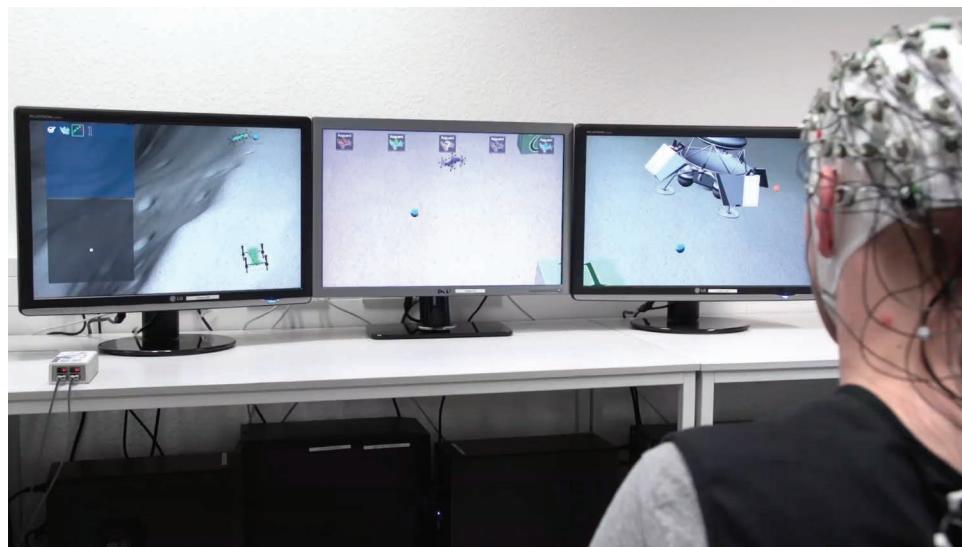
Besides enhancing ease of interaction and reducing errors, many robotic applications, especially those requiring the simultaneous control of a team of robots, have to handle the limited cognitive resources of a human that lead to cognitive overload. Embedded multimodal interfaces represent a good solution to handle this scenario by adapting the interaction with respect to the cognitive load of the human, as demonstrated in the field of physiological computing (see Zhou et al.



**Figure 13.12** Immersive virtual 3D multi-robot control using a CAVE. (Figure courtesy of [Kirchner and Drechsler 2013])

[2018] for indicators of cognitive load). To achieve this, the awareness by the interface of the human's state (vs. the awareness by the human of the robot's state as in the example given before) is enhanced by means of the embedded multimodal interfaces, as will be explained next.

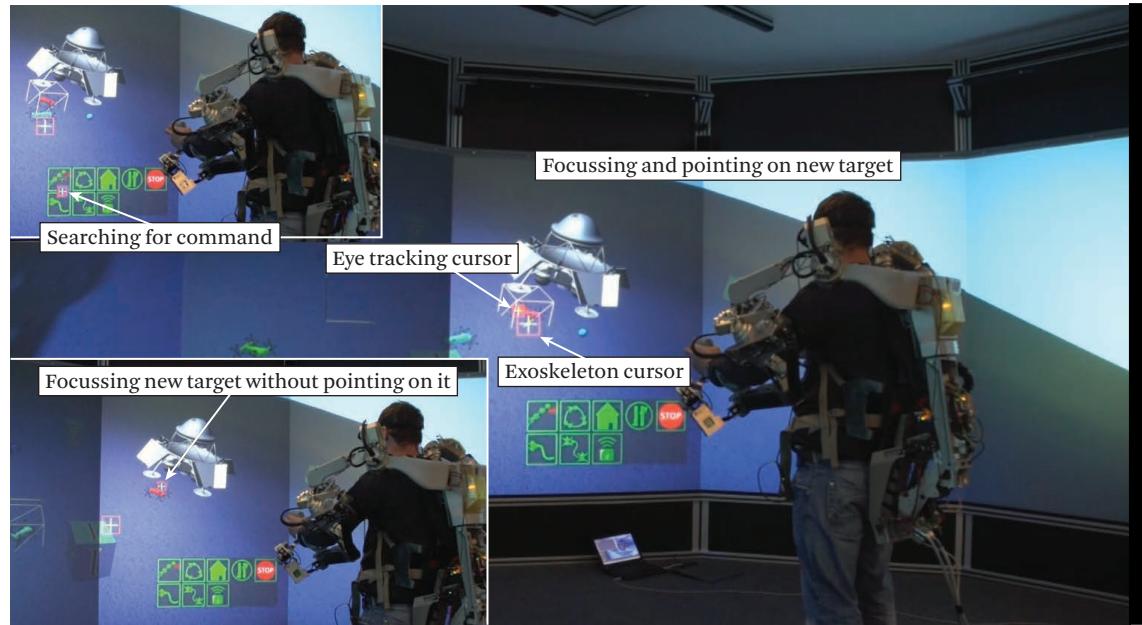
In the video referred to in Figure 13.13 an application of multi-robot control is displayed. The control interface is a very complex one. It is based on a virtual environment using the software Machina Arte Robotum Simulans (MARS) [Rommerman et al. 2009, MARS 2015], which can be run as a 3D environment in, e.g., a CAVE (see Figure 13.12), as a 2D environment on a standard personal computer or on a multi-screen system (see the video referred to in Figure 13.13). Information about the robot is mainly presented in visual 2D or 3D mode. However, when using an exoskeleton as input device, force feedback can be applied to the operator. Thus, the exoskeleton is not only implemented to control the display but also to interface with a robotic arm, similar to the VI-Bot interface (see the video referred to in Figure 13.8). Moreover, the interface is implemented so the user can interact by using different alternative modes or different modes in combination. For example, eye tracking is implemented to enable the operator to select interaction icons instead of using a 3D mouse, a wand, or an exoskeleton to allow the operator to use the referenced input device and modality; see Figure 13.14. In the future, gestures will additionally be implemented to navigate through the scenario, to change the virtual camera position, or to select interaction modes in order to send control commands.



**Figure 13.13** **Video:** Virtual multi-robot control in 2D using a multi-PC system supported by embedded brain reading. [https://www.youtube.com/watch?v=zeFp\\_JBSBxA](https://www.youtube.com/watch?v=zeFp_JBSBxA)

While explicit control is optimized to ease the interaction in the VI-Bot scenario, the cognitive resources of the operator may be insufficient when the task load is too high. Therefore, implicit control of the interface is enabled by an embedded brain reading approach that adapts the interface with respect to the individual task-load of the user [Kirchner et al. 2016a]. This aspect is highly relevant since ease of control is one component that improves interaction quality; the skills and mental workload of the operator are other aspects that have to be considered in the context of limited resource theories (see Volume 1 Chapter 1). The embedded multimodal interfaces can be adjusted online with respect to the current task-load and general training status of the operator, resulting in different task frequencies, i.e., interstimulus intervals (ISI) for individual operators.

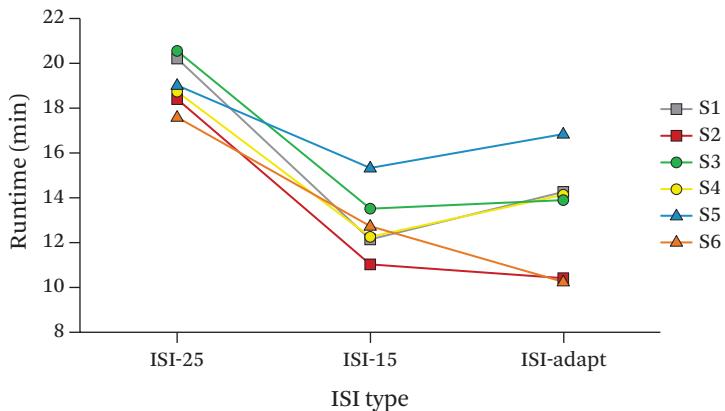
In Figure 13.15 it is shown that the adaptation in run 5 and 6 supported by embedded brain reading results in individual differences in total runtime, i.e., the time needed to fulfill 30 given tasks. In runs 1 and 2 and runs 3 and 4 the ISI was fixed to a long (run 1 and 2: ISI-25 condition) or shorter time (run 3 or 4: ISI-15 condition). Under both conditions the total runtime was very similar between participants. While the performance in runs 1–4 without adaptation seems very similar over the group, some participants reported that they were bored under ISI-25 or even IS-15 condition. Subjects 2 and 6 could achieve much shorter runtimes when supported



**Figure 13.14** Multimodal explicit control: operator can chose to interact with interaction icons of the interface and the robotic systems via the exoskeleton or the eye tracker or a combination of both.

by embedded brain reading. To reduce runtime from the beginning would not have been a good solution, since some individuals would have become stressed, like participant 5. Thus, implicit control is required to adapt an interface to the overall needs (depending on her or his skill level) and the current needs (depending on task load) of the human operator while avoiding excessive cognitive load or boredom.

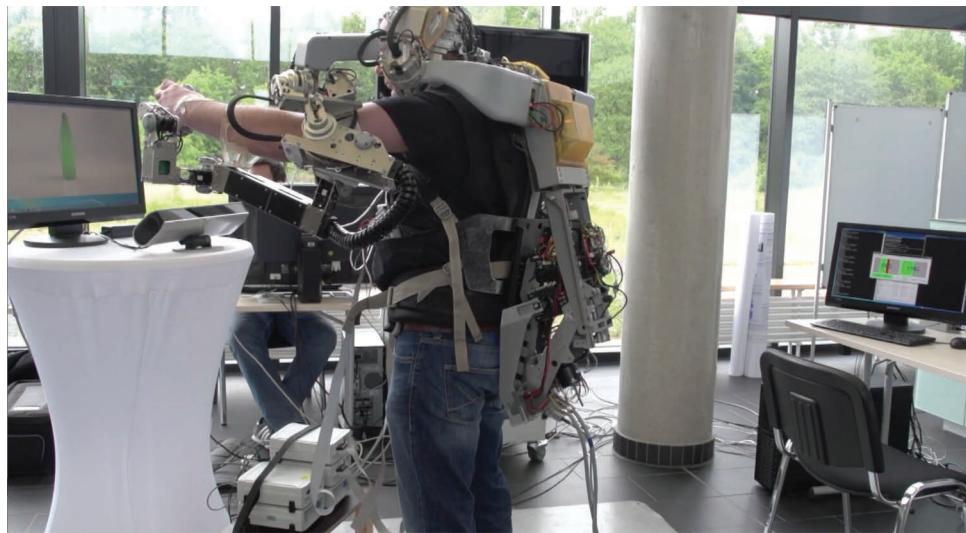
The usage of multimodal and multisensor data in embedded multimodal interfaces have many positive effects on human-robot interaction. Whereas we have already discussed that a temporal cascaded approach allows a smoother interaction, the combination of multimodal data can further increase the dimension of insight possible into the intention of the user more than a single modality could. Some examples should be given next that accord with the Gestalt theory of perception (see Volume 1, Chapter 1), that the whole can be more than the sum of the individual parts. An example is the combination of different covert physiological measures and overt measures from the human, like eye movement data, EEG, and EMG data for the control of a robotic rehabilitation system [Kirchner et al. 2013a]



**Figure 13.15** The means of runtime for a fixed ISI of 25 S (ISI-25), a fixed ISI of 15 S (ISI-15), and ISIs adapted by embedded brain reading (ISI-adapt) are depicted. (Figure courtesy of Kirchner et al. [2016a])

(for more information on multimodal behavioral signal processing systems see Volume 2, Chapter 10 and Volume 2, Chapter 12).

Using each individual signal not only reduces the reliability but would further generate less "meaning", e.g., to look at an object does not tell us whether interaction is wanted. EMG activity alone does not tell us whether interaction is desired to begin, e.g., to move the arm to the object, or whether the subject was just bumped by someone and used the arm to balance or has experienced a spasm. In addition, EEG cannot clearly tell us whether the subject is planning to actually execute the interaction or is only imagining the interaction, since internal visualization and planning of motor execution result in very similar brain patterns. If these individual signals are combined, the outcome becomes more reliable, especially if the expected temporal order of signals is consistent, i.e., a temporal cascaded approach is followed. Moreover, only if at the same time or afterward brain activity related to movement planning is detected in the EEG, the user is probably not only looking at the object but at least is thinking about starting an interaction. Finally, by adding the EMG signal, it becomes quite clear that at the moment when the subject looks at the object, plans a movement and tries to activate the muscles, that he or she indeed wants to start a movement. Therefore, the combination of all three signals tells us more than the individual signals alone and results in very reliable interpretations of the human's intention.



**Figure 13.16** **Video:** Demonstration of biosignal usage to control a robotic system: the exoskeleton is moving the right or left arm in case that the bottle is in focus and biosignal analysis detects the intention of the operator to move from EEG and EMG signals.  
<https://www.youtube.com/watch?v=BRpbZFOXdRk>

Deciphering the intention of the human user is highly relevant for rehabilitation robotics. In the video referred to in Figure 13.16 it is shown how the combination of EEG, EMG, and eye tracking data can be used to control an assistive device, i.e., an exoskeleton. While the most straightforward approach is to combine all signals in an “and” fashion, i.e., to only drive the exoskeleton in case that all three signals are detected, EEG and EMG data can, to some degree, substitute for each other (see decision-level fusion in Volume 2 Chapter 12, Section 12.8). Therefore an “or” combination is also possible. Depending on the application or state of rehabilitation it might be more useful to use the “and” or the “or” combination, i.e., different temporal cascaded approaches.

For example, during early rehabilitation, the patient should strongly be motivated to train her or his impaired limb, however muscle and brain signals might be weak. It might be better to use the “or” combination here, i.e., to drive the exoskeleton in case there is either the expected EEG or EMG signal detected. Later in rehabilitation the patient might be more annoyed by unwanted movements or the patient should be more strongly involved in the interaction. Here, the “and”

combination might be more suitable. The kind of combination of different signals has therefore an impact on the behavior of the interface and on the reliability of intention recognition (see Figure 13.17 Kirchner et al. [2014]). Hence, the temporally cascaded approach that is followed does strongly determine the quality of awareness of the human's state that a robotic system can derive.

In summary, embedded multimodal interfaces enable bilateral interaction. They make use of overt and covert measures to detect active and passive human states to make the robotic system aware of the state of the human. Furthermore, they use multimodal- multisensor data to make the human aware of the robot's state. This bilateral awareness strongly improves interaction, reducing cognitive load, interaction forces, and interaction errors. Based on a carefully chosen temporally cascaded approach, they allow that multimodal data can be combined, such that the whole can be more than the sum of the individual parts (see Panagakis et al. [2018]). The later requires a deep integration or embedding of the multimodal interface into the system control.

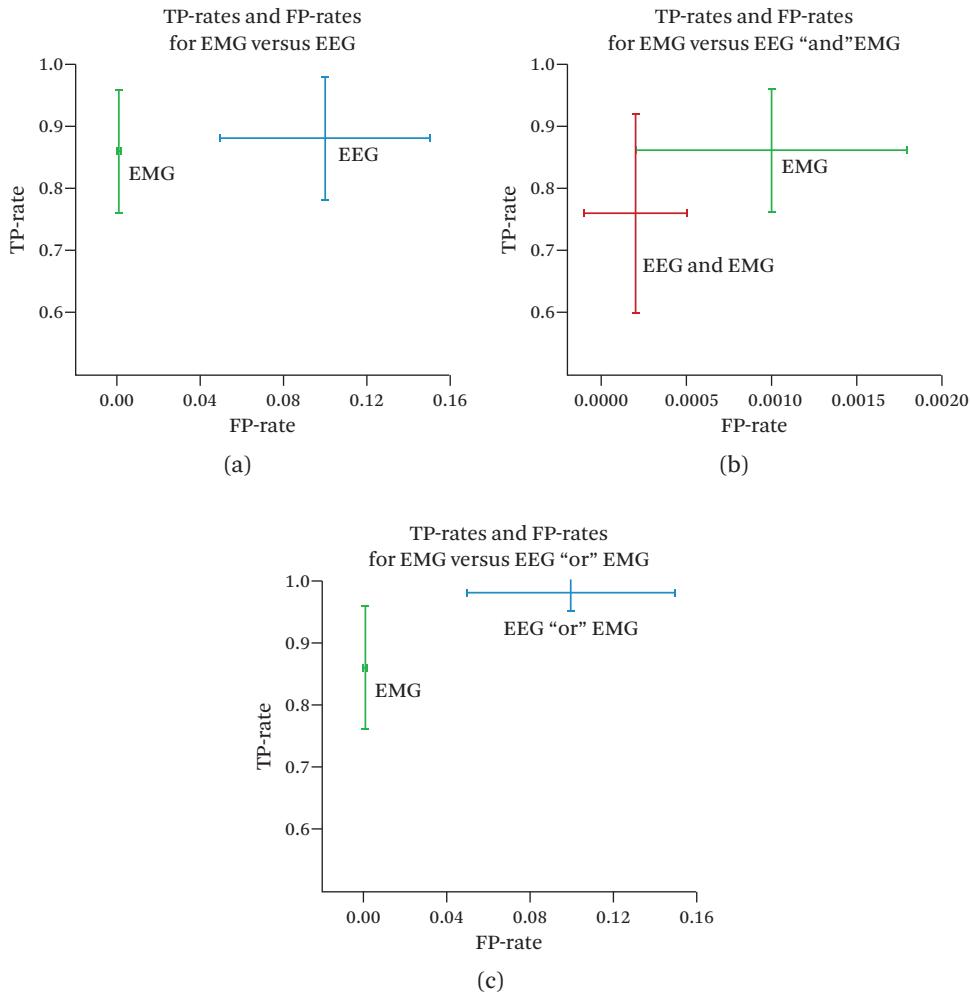
## 13.5

### **Future Trends: Self-Adapting Embedded Multimodal Interfaces and Societal Implications**

In this chapter, we first introduce three main trends in embedded multimodal interfaces which will enable future applications in human-robot interaction.

1. The functionality of embedded multimodal interfaces must be highly adaptable in an online fashion. They should adapt to changes in the context of interaction, changes in the state of the robot or the human, or changes in availability of certain sensors or sensor data.
2. To enable this level of adaptability future interfaces must be deeply embedded into the control of a robot in order to develop specific hardware and software solutions.
3. There is a trend to embed the interface into applications and systems, in order to make them smaller and more energy-efficient and to allow both long-term usage and a high degree of flexibility.

While many developments are new research territory for human-robot interaction, we will show with examples that research from other areas can be used and solutions can be adapted. Using these approaches, a more direct and efficient interaction between human and robot will be enabled. However, these changes and future developments will have societal implications.



**Figure 13.17** Effect of different combinations of EEG and EMG data on true and false positive rates during classification. (A) Prediction results in TP- and FP-rate for EEG (blue) and EMG (green) analysis. (B) Prediction results in TP- and FP-rate for EEG “AND” EMG (red) and EMG (green) analysis. (C) Prediction results in TP- and FP-rate for EEG “OR” EMG (black) and EMG (green) analysis. (Figure adapted from Kirchner et al. [2014])

### 13.5.1 Inherent Self-Adaptation and Deep Integration

Multimodal-multisensor interfaces that are embedded into application procedures, robotic systems, application environments, into the users’ clothes, or that are implemented as wearable devices will be perceived as normal in the future. Con-

sidering continuously changing interaction environments and interaction tasks, it is obvious that multimodal-multisensor interfaces must be able to adapt to changes by themselves in the future. Today's robotic systems and their interfaces are already of such complexity that any hard-coded change will require a high number of specialists to perform the required software and hardware adaptations. For the common user, it will become impossible to handle this kind of technical detail. However, the need for adaptation not only for robotic systems but also for their interfaces will increase.

### 13.5.1.1 **Closed-Loop Design for Self-Adaptation to the State of the Human**

An embedded multimodal interface allows the robotic system to adapt to the movement intentions and states of human operators. This system works via *closed-loop control* logic whereby multimodal signals from the user are monitored, analyzed, classified and converted into appropriate outputs or adaptive responses from the robot. This closed-loop design requires a set of rules whereby a target state triggers an adaptive response. However, this may not be an exclusive relationship and a range of potential responses are available once a specific state has been identified. For example, if the robotic system detects high mental workload during a process control task, it could slow the pace of operations (see Section 13.4) or suggest a break. The rules that translate multimodal detection into an adaptive response at the interface draw from a repertoire of possibilities, all of which are equally likely to create the desired effect on user behavior. This scenario poses the question: How does the multimodal interface select the most appropriate response from an existing repertoire of possible responses?

A closed-loop system for either positive or negative control is often characterized with reference to a single discrete cycle of monitoring and adaptation. In this case, a single cycle may describe how the detection of high mental workload is translated into a reduction of task pacing in order to aid the human operator. This is a first-order process of adaptation wherein the loop detects and responds to a target state in the short term. Once this adaptation is activated, it is possible for the multimodal system to detect whether its own adaptive response had the desired effect on the human operator. If slowing the pace of the task has reduced the mental workload of the operator (assessed via neurophysiological monitoring), this adaptive response is deemed to be successful. The other possibility is that the adaptive response failed to reduce the mental workload of the human operator, in which case the multimodal interface must enter a second cycle of monitoring and adaptation in order to select another response, e.g., suggest a break. This latter process is called second-order adaptation or *reflexive adaptation* [Serbedzija and Fairclough 2012]

because an adaptive response is based upon the closed-loop monitoring of the consequences of its own intervention on the state of the user.

A second-order process of reflexive adaptation can facilitate machine learning over a sustained period of interaction with an individual operator. In order for the multimodal system to adjust to the individual, it must accumulate a database that identifies those adaptive responses found to be effective for a particular user. Second-order adaptation describes a generative process where the repertoire of adaptive responses is ‘pruned’ or customized on the basis of repeated interaction with a specific user. This evolving cycle has been described as a process of mutual adaptation with three main phases [Fairclough 2015]. The initial encounter between the multimodal system and user is characterized by improvisation. The system responds to the user in a generic fashion using default adaptations with no prior knowledge of individual preferences. Adaptation may be perceived by the user to be less than optimal during this early phase. As the user spends more time interacting with the system, second-order adaptation improves the acceptability of responses from the perspective of the user. This second phase of reciprocal coupling is characterized by tailoring the adaptive repertoire of the robotic system to the individual. If we look further ahead in time, in terms of months and years, it is reasonable to expect that any stable model of preferences acquired during reciprocal coupling will have limited longevity—as the user acquires greater skill or habituates to popular adaptive responses or experiences cognitive changes due to aging. The third phase of co-evolution describes a process of updating a stable model of user preferences over a longer period of time.

### **13.5.1.2 Self-Adaptation to the Context of Interaction**

Besides adapting an interface to the user, adapting robotic systems and their interfaces to the context of interaction will become an important direction for the future of robotics. To allow this development, the context of interaction must be recognizable from the perspective of a robotic entity. In an industrial context where operational sequences are prescribed and must accurately be fulfilled, it seems that the recognition of the context of interaction should be relatively straightforward and may even be predefined. However, flexible support must consider deviations from the procedure and personal preferences of the human user. For example, personal preferences can be handled by systems that learn during interaction from physiological data (such as error-related potentials in the human EEG) what the user means for example by certain individually chosen gestures, i.e., it learns the mapping between gesture and action [Kim et al. 2017]. In non-predefined interaction scenarios, the recognition of context becomes even more relevant. To recognize the

context of interaction, different data sources can be considered if available data about procedures and preferences of the human can be used. For example, knowledge about preferences of patients might even be more relevant than the analysis of physiological measures to optimally support them by a robotic system [Novak et al. 2013]. However, physiological measures as explained in the example in the video referred to in Figure 13.16 can also be used to detect the context of interaction. Moreover, the supporting system itself is also able to detect the context of interaction, as explained in the example presented in the video referred to in Figure 13.8. On the other hand, interaction usually requires physical activity of the human. Thus, complex movement data is a highly relevant source of information to deduce the context of interaction. While today human movement behavior is often analyzed to develop approaches that enable robotic systems to learn to imitate human behavior [Metzen et al. 2013, Mülling et al. 2013, Pastor et al. 2009], movement data can also be used to recognize the context of interaction [Senger and Kirchner 2016], especially body posture can be very informative about the behavioral context of interaction. In a simple case, the direction of movement with respect to a robotic system can tell the robot whether a human wants to interact or not [de Gea Fernández et al. 2017].

### 13.5.1.3 Software Frameworks and Hardware Solutions for Deep System Integration

To use the power of embedded approaches, future multimodal interfaces must potentially make use of any available data, like sensor data of the robotic system or from environmental supervision, data about procedures, and even (physiological) data about the interacting human. This brings along some challenges in data storage, processing, selection, and handling. Thus, new software and hardware solutions, such as the open source software framework pySPACE [Krell et al. 2013] or the specialized software framework reSPACE [Wöhrle and Kirchner 2015] which can run on embedded systems, must be developed that allow for flexible usage of multimodal data. Optimization during runtime will require reconfiguration of hardware at runtime to allow optimal, i.e., time, resource, and energy-efficient interpretation of potentially changing data sets. These approaches have to consider which data is most relevant to interpret the current situation, to infer the intention of a human in a certain situation, or to estimate the best possible support of a human by a robotic system at hand.

Supervised and unsupervised learning methods must be applied on top of sophisticated signal processing that reduces data and filters relevant information. All this must be achieved on embedded hardware that allows fast but also resource-saving analysis. Different approaches must be combined. For example, the usage

of processing units based on Field Programmable Gate Array (FPGA) was shown to support powerful and fast processing with low energy consumption on small-sized, embeddable devices [Wöhrle et al. 2014]. However, these approaches are still limited. Thus, a combination of small embedded devices and powerful central processing units must be considered and promoted. While the embedded system will perform data analysis within the interface, the robot or the wearable device, the central processing unit or units will be able to perform more complex calculations required to optimize data selection or combination, processing, and, hence, adaptation of the current processing flows. Thus, hybrid hardware/software solutions will enable self-adapting embedded multimodal interfaces.

In summary, future robotic systems must not only behave autonomously but must deeply understand humans to better support them and to allow flexible interaction and cooperation. While this seems to require extra effort at first glance, on closer examination two things become obvious. (1) Approaches that are currently developed to enable robots to behave better and to perform complex task, such as approaches that enable imitation learning, i.e., learning from human demonstrations [Schaal 1997, Argall et al. 2009], are also relevant to improve interaction. (2) The amount of multimodal data from multiple sources will increase. Thus, for both approaches new software and hardware solutions must be developed in order to profit from each other.

In order for next-generation robotic systems and their embedded multimodal interfaces to accommodate the requirements dictated by the human-robot cooperation it will be necessary to develop standardized robot control frameworks and architectures that not only allow easy integration of internal parts of the robot such as motors, cameras etc., but also adapt the systems control to human interaction. The frameworks must be designed such that they are flexible toward changes in multimodal input and multimodal output during runtime, e.g., changes in sensor input or changes in interface input. Examples are the Robot Construction Kit (ROCK) and DROCK [DRock 2015]) software frameworks. These model-based approaches allow the designer of a robotic system to define the system from a library of well-defined and mathematically modeled components. This approach works from the hardware as well as from the software perspective. Therefore, software-based concepts like adaptation and learning can be integrated with standard planning and control approaches via machine learning to enable usage by non-specialists.

It is obvious that both main research directions, i.e., autonomous artificial intelligent robotic systems and human-machine interaction become inseparable to develop future robotic systems that are able to optimally support humans and

to interact with them intuitively. Robotics is a very good example that shows how different fields of research must not only work together to their mutual advantage.

### 13.5.2 Current and Arising Societal Implications

As stated before an embedded multimodal interface is capable of enhancing human-robot cooperation by: (a) increasing machine awareness of the user via monitoring and (b) personalizing the behavior of the robot to the preferences of the individual via the closed-loop process of reflexive adaptation. This emergent approach is designed to evolve the technical sophistication of how people interact with robots. However, these technological advancements are associated with a number of societal implications.

It is important to understand that closed-loop systems are driven by goal-directed logic. The closed-loop within the multimodal interface is programmed with a specific directive, e.g., to prevent mental overload, to improve performance efficiency, to preserve the safety of the operator. Its repertoire of adaptive responses are simply the means by which the system achieves its specified goal. Unlike the inert and passive technology of today, this symmetrical interaction is characterized by a degree of agency on the part of the machine and the requirement for a human operator to cede a degree of control to the system. Given this, it is important to define the agenda of the machine to be effective, reliable, and not to lead to unforeseen circumstances [Kirchner and Drechsler 2013].

#### 13.5.2.1 Establishing Trust

The challenge for multimodal interfaces is how to make the robotic system an effective “team-player” from the perspective of a human user [Klein et al. 2004]. In order for a robotic system to work with human users, it is important for embedded multimodal interfaces to establish a degree of ‘trust’ with their users. Miller [2005] argued that technology could earn the trust of the user by transparency, i.e., the laws of cause and effect encapsulated within the closed-loop are clearly understood by the user. This transparency can be enhanced by clear feedback to the user during interaction and predictable behavior on the part of the robotic system. The development of trust between robot and user requires time and can only be achieved through repeated interaction over a long period.

Multimodal interfaces utilize increased data processing capacity to: (1) monitor the behavior/physiology of the user; (2) make inferences about the psychological status of the user based on monitoring; and (3) translate those inferences into timely and intuitive responses at the interface. In order to monitor-infer-adapt within a working control loop, multimodal interfaces must operate as surveillance

systems, gathering data on individual users in order to respond proactively in an intelligent fashion. In addition, the multimodal interface requires a degree of autonomy to adapt to changes in user state without any requirement for explicit commands. This combination of intensive user monitoring with autonomous function is the price to be paid for the advanced level of functionality characterized by multimodal interfaces.

Societal issues of trust and system autonomy are both significant and inherently interconnected for the introduction of multimodal interfaces, particularly those designed to capture non-intentional responses as part of passive or reactive systems. The first concern is the degree of confidence that the user has in the technical prowess of the system. In other words, can the system collect data with sufficient fidelity? Is it capable to make a sensitive and accurate discrimination between different psychological states? Can the system successfully translate these multimodal data into sensitive and intelligent adaptation at the interface? If the user can answer those questions in the positive, he is likely to trust the system and will be comfortable ceding control to autonomous functions. If not, the user will either desire a return to manual control (if that is possible) or work unhappily and suspiciously with a technology that he views as erratic and unpredictable; in either case, the proposed advantages of multimodal interfaces will be lost.

A second issue concerns the ‘values’ that are inherent in the control directives of a ‘machine with an agenda’ [Fairclough 2015]. In order for the multimodal interface to respond to the detection of specific psychological states, it must translate the detection of a target state into an appropriate response at the interface. This process of translation can be straightforward. If the user is working on a safety-critical task, the adaptive logic of the multimodal interface should promote sustained engagement with the task in order to achieve error-free performance and maximize safety. If the user becomes bored or complacent, the multimodal interface will evoke a strategy to restore task engagement, e.g., to transfer tasks from autonomous to manual control in order to re-engage the user with task requirements. This enhanced autonomy, which is characteristic of multimodal interfaces, permits the system to adapt in order to operate upon the user—to effectively manage the psychological state of the person. Naturally it is important for the user to trust the system if he is to be completely comfortable with this type of advanced interaction. It is also possible that the goals and desires of the human user may diverge from the control directives of the multimodal interface: the user may desire to take a rest break, may feel unwell, or may resent working with a system that seems to expect him to do all the work. There is potential for an interaction with the multimodal interface to descend into a battle of “wills” where the human is forced to subjugate

his wishes or desires in the face of a technical entity, which is both implacable and incapable of behaving with sufficient flexibility. This example demonstrates how implicit control based on monitoring combined with autonomous function can ‘snowball’ into a subversion of human goals and desires. This is why trust is such an important societal factor for acceptance of multimodal interfaces, humans must: (1) have faith in the technical proficiency of the system in order to comfortably relinquish some control over the interaction; and (2) interact with the system in the knowledge that autonomous decision-making will not subvert their autonomy and rights as human beings.

### 13.5.2.2 The Relevance of Consent and Data Privacy

Consent is a second important societal factor to be considered when the behavior and physiology of the person is monitored by technology. The data that streams from the user to the multimodal interface must be considered to be personal data. Issues surrounding data privacy and data ownership may be crucial influences on the extent to which users will accept the introduction of multimodal interfaces. It has been argued that openness with respect to data acquisition, storage and sharing is fundamental to the relationship between the user and the system, i.e., reciprocal accountability [[Brin 1999](#)]. In this case, the user allows personal data to be collected in full knowledge on how it will be used, stored, and protected by the system. There is evidence that users would prefer to have a contractual arrangement whereby data is only obtained, stored and shared with full written consent [[Reynolds and Picard 2005](#)]. At the time of writing, research governance when performing experiments with human participants often requires written consent and compliance with data protection laws before personal data can be collected. But the degree of control that users can legally exercise over the ways in which personal data is stored and used is fundamentally determined by the extent to which users are deemed to own their own data [[Fairclough 2014](#)]. If the user is granted full ownership, they can control what is stored and who can access these data. Full ownership would allow a user to remove data from the system if they wished to do so, the user could even charge for access to their data.

Personal ownership of data and informed consent are important steps to protect the individual equipped with multimodal sensors. However, it is equally important that personal data is stored and managed in a way that continues to safeguard the rights of the individual. The management of personal data by external agencies sits at the heart of the General Data Protection Regulation (GDPR) that came into force throughout Europe in May 2018. This legislation grant greater protection and rights to the individual whose personal data are held by ‘controllers’ or ‘pro-

cessors', whether they be individuals, companies, or organizations. The issue of consent is central to GDPR and entities that hold personal data are subject to fines if a breach of confidentiality is detrimental to the individual. While GDPR is a positive development with respect to accountability, the primary risk to privacy from multimodal monitoring is a process of inference fueled by data aggregation from multiple sources [Friedland and Tschantz 2018]. The process of data aggregation is an important technique for accurate assessment of the operator state, for example, the detection of high mental workload is improved by cross-referencing task monitoring (e.g., activity or phase of operation) with neurophysiological data, such as EEG. The former provides a meaningful context for the latter. Similarly, emotional responses can be characterized by a combination of facial expression, autonomic psychophysiology, body posture, and vocal expression. It is important for users of multimodal sensor systems to understand which data sources are active and how they aggregate in order to deliver an inference about operator state, especially if aggregation occurs across databases that are controlled by different 'controllers' or 'processors.' In the previous example, these types of data can also serve secondary purposes, such as identifying individuals via unique features such as facial expression or voice, which leads to a scenario where individuals can effectively 'profiled' and compared with respect to specific operator states, e.g., John displayed high mental workload twelve times during the task compared to Jane who only entered a state of high workload twice.

The issue of data ownership points to another dimension of trust in multimodal interfaces—the issue of data privacy. Certain types of multimodal interfaces rely on the monitoring and measurement of psychological states, such as mental workload or frustration or fatigue (see also Volume 3, Chapter 13), but which other parties are allowed to access these data? And perhaps more importantly, can the individual user be identified on the basis of these data? A company may wish to record data from all user sessions, particularly during safety-critical activities, for purposes of accident investigation. In this case, part of the conditions of employment would require a user to share personal data and to be identified with that data. The societal issues associated with system use are more profound when a company wishes to access data from multimodal interfaces for purposes of performance management. For example, to assess the level of concentration exhibited by an employee during their duties or to capture episodes of frustration or to gauge alertness at the beginning of the work session. In this case, personal data is collected and interpreted from the individual in a way that could actively disadvantage that person. It is unlikely that the user would trust the system in this scenario, not because of what the technology is designed to do, but because of the way data is harvested

and used by the system administration. It is also possible to imagine a more positive scenario where employee monitoring is performed to identify instances of high occupational stress in order to bring about changes to working conditions that ultimately benefit the employee. If we extend this scenario to service industries, such as multimodal interfaces in the context of internet search, a company can argue that harvesting data tied to an individual can be used to improve the quality of service offered to that user. The key issue in this case is whether the user must surrender their ownership of personal data in order to use that service, especially a service as pervasive and indispensable as internet search. A secondary problem that relates to the recent GDPR legislation concerns the sharing of personal data by a service provider with other entities, i.e., if the user surrenders ownership to use a service, do they also surrender the right to control the distribution of their personal data?

The issue of data privacy, anonymization, and sharing is particularly pertinent for systems that collect data from the brain and body as part of the human-computer interaction (see [Friedland and Tschantz \[2018\]](#) for further discussion). It is possible to extract information about the health of the person based on these data. As two examples, the presence of epilepsy can be detected from an EEG record collected on a long-term basis and it would also be possible to identify markers of cardiovascular disease from the regular acquisition of electrocardiographic data. It would be controversial for an employer to collect data from a specific individual for the purposes of mental workload monitoring, for example, and use the same data set as part of a health assessment. The sharing of sensitive, personal data of this type with other legal entities, such as health insurance companies, is also problematic from the perspective of the user. If the individual cannot own their data, they cannot control how it will be stored, shared, and analyzed, hence, some kind of informed consent possibly in the form of a contractual agreement along the lines of GDPR is likely to be a prerequisite for users of multimodal interfaces.

The ethical issues around data ownership have a number of practical implications for the design of multimodal interfaces. An individual could log on to the system anonymously and wipe the data record after use. This level of data protection is not possible when the system must store data that is linked to an individual user, for example, when the multimodal interface is designed to personalize adaptation to that person during sustained and repeated system use (see above). In this case, data must be stored and storage should be secure, e.g., password protected and encrypted. One solution is local storage, using a physical media such as a USB pen drive, that is owned by the user and used as a repository for all data collection during the interaction. However, due to the size of the data file over a period of time and the need for a back-up, it is unlikely that a physical device would serve as

a practical solution in isolation and remote storage elsewhere would require protection and anonymization. There are a number of studies (e.g., Barra et al. [2016]) where data from the brain and body has been used to identify the individual and the use of a biometric key to unlock data from the same source offers an intriguing solution for the future.

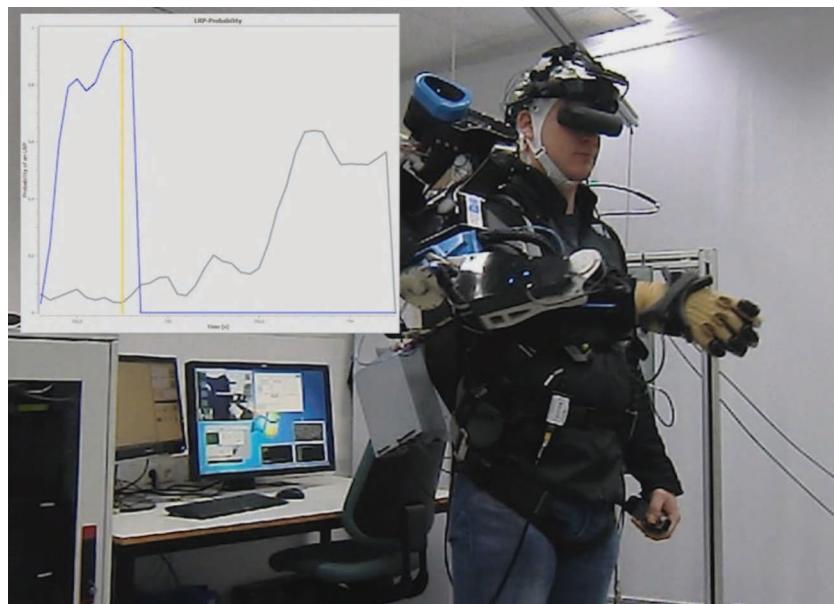
Embedded multimodal interfaces offer the possibility of enormous technical advances but progress in this direction does introduce novel problems of user trust, system autonomy, and data ownership.

## **13.6 Supplementary Digital Materials: Exoskeleton's Mode Change Supported by Embedded Brain Reading—Approach and Evaluation**

The VI-Bot exoskeleton (see the video referred to in Figure 13.8) has three operation modes: full user support mode, teleoperation mode, and transient mode (see Figure 13.9). During teleoperation mode, the movements of the human are mapped to the robotic system and the user receives force feedback from the robotic system. During full user support mode, the exoskeleton keeps the user's arm in a fixed position. During transient mode, the user can move freely without controlling the robotic system.

Changes between full user support mode and transient mode are intentionally controlled by overt arm movements, i.e., an active human state is detected by the exoskeleton. Changes between transient mode and teleoperation mode and back are intentionally commanded by overt hand gestures. Again, an active human state is detected by the interface. Being in the transient mode, there is a mode change possible back to full user support mode. This change is non-intentionally controlled whenever the user is *not* moving his or her arm for a certain time period. The user is usually not pausing to intentionally control the mode change but is stopping during robot control for different reason, e.g., to change position to get a better view of the robot's arm or to think about a solution for a difficult situation or to handle additional requests like communication with a second person, while the exoskeleton is supporting the user by keeping the arm in a fixed position such that the operator does not have to hold his arm by himself. The latter becomes relevant in case of very fine manipulations that are interrupted by longer breaks. In such situations, the user wants to avoid any big movements and wants to keep his arm in a specific position.

Thus, while mode changes between transient mode and teleoperation mode are intentionally controlled by the human, mode changes between transient mode and



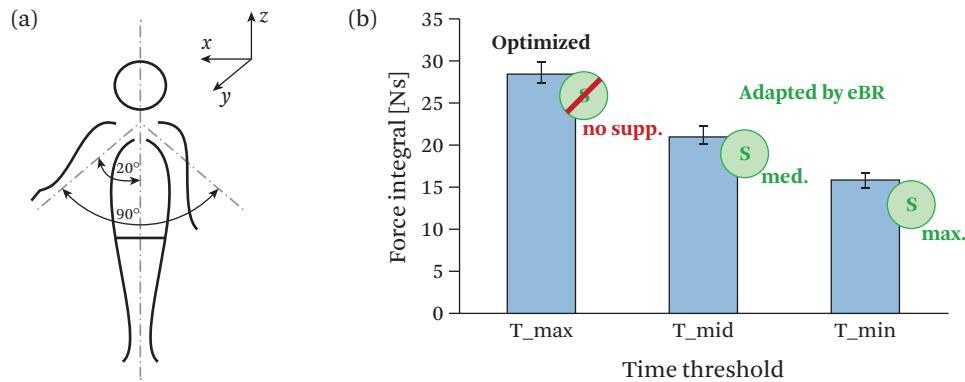
**Figure 13.18** **Video:** VI-Bot implicit control to adapt the exoskeleton's control to the user's inferred needs: support of interaction by movement prediction based on EEG signals.  
<https://youtu.be/fiI4MKPTFg0>

full user support mode are non-intentionally elicited (see above). Further, changes between full user support mode back to transient mode which are intentionally elicited by the human are additionally non-intentionally adapted by the interface itself by detecting specific active states of the human from covert measures, i.e., brain signals recorded from the surface of the human's head [Figure 13.18]. As mentioned above, when the system is in the full user support mode, the human can intentionally control the interface to change back to transient mode simply by starting to move his arm again. This overt behavior is measured by the exoskeleton to detect the active state of the human, i.e., arm movement.

To intentionally activate a mode change from full user support mode to the transient mode the user must press against force sensors of the exoskeleton for a certain time that is long enough to avoid false commands by, e.g., muscle twitches or movement of the upper body. However, even if the strength of the forces and the duration of pressure application to the exoskeleton are individually optimized for each user to cope with different body measures and muscle strength, the user will feel that the exoskeleton will not directly respond to her or his movements. Therefore, the ex-

plicit control from the full user support mode back to the transient mode is adapted by another implicit control loop that makes use of covert measures, i.e., brain activity, that enable the interface to detect another active state of the human, i.e., preparation of an arm movement. This cascaded approach of combining explicit control and implicit control for adaptation enhances transparency, since the movement planning phase can be detected from EEG before the movement is executed. This information can be used to prepare the exoskeleton for the explicit control by a later arm movement by enhancing the sensitivity of the sensors that detect the movement onset. As for the implicit control example, the adaptation requires the system to infer a human state or upcoming human state. No explicit command is required from the human.

To optimize the general sensitivity of the exoskeleton for each subject, a time threshold was defined individually, i.e., a minimum time he or she had to press against the exoskeleton with a minimum force to release it from full user support mode. Both minimum time threshold and force threshold were chosen by performing a calibration procedure in which the subjects had to keep their forearm completely extended and the shoulder flexed forward in order to bring the third joint of the exoskeleton in an angular range between 15 – 20°. The user was then asked to repeat an oscillatory movement of about 90° with a regular speed (see Figure 13.19A). The requested type of movement and speed was shown to the user beforehand in a movie that showed the exact movement sequence and timing. At the beginning of each session, the force threshold was set to a maximum value  $F_{th}^{Max}$  and the time threshold to a minimum value ( $T_{th}^{Min}$ ) in order to assure that no movement would cause the system to lock-out. The user was asked to perform the rotational movement ten times. If the lock-out did not occur,  $F_{th}$  was decreased for 10% and the experiment was repeated. This went on until the user locks out due to the rotational movement. At this point, the force threshold was kept constant while the time threshold was increased until the perturbation movement was no longer able to cause the lock-out event. The discovered thresholds  $F_{th}^{Min}$  and  $T_{th}^{Max}$  were the values that were individually estimated for each subject. These values represented the time and force of pressure that must have been exceeded to lock out the system from full user support mode to transient mode. Thus, the time  $T_{th}^{Max}$  was chosen which was the shortest one that did not cause an unwanted lock-out at a minimum interaction force  $F_{th}^{Min}$ . This individually defined time threshold ( $T_{th}^{Max}$ , see  $T\_max$  in Figure 13.19b) could further be reduced based on the outcome of the embedded brain reading interface that predicted movement preparation, while the required force ( $F_{th}^{Min}$ ) was not adapted. The user had to press shorter (shorter than  $T_{th}^{Max}$ , see  $T\_max$  in Figure 13.19b) against the sensors the more likely movement planning



**Figure 13.19** Reduction of interaction force between exoskeleton and human by adapting the exoskeleton’s control by means of embedded brain reading. (A) positioning of forearm and movements of the upper body during the calibration session are depicted. (B) Mean values for applied force over time (force integral) calculated for ten measured movements across five subjects under each time threshold condition are depicted.  $T$ : time threshold, i.e., time to press against the sensors that must be exceeded to unlock the exoskeleton from full user support mode.  $T_{\text{max}}$ : individually estimated for each subjects based on calibration measurements with no support by embedded brain reading (no supp.);  $T_{\text{min}} = 10 \text{ ms}$  at  $S_{\text{max}}$ : maximum adaptation (prediction score equivalent with 100% correct movement planning detection);  $T_{\text{mid}}$ : medium time threshold value (between  $T_{\text{max}}$  and  $T_{\text{min}}$ ) at medium adaptation  $S_{\text{med}}$  by embedded brain reading (prediction score equivalent with 75% correct movement planning detection). (Figure modified after Folgheraiter et al. [2012])

was predicted based on EEG analysis. This adaptation of the exoskeleton required less force over time from the human for explicit control of the exoskeleton, i.e., to make the exoskeleton change from full user support mode to transient mode (see Figure 13.19). To calculate the effort required to pass from full user support mode to transient mode, we integrated the force according to Equation 13.1:

$$I = \int_{T_0}^{T_{\text{Lout}}} F_{\text{int}}(t) dt. \quad (13.1)$$

To analyze how much the required minimum interaction force  $I$  could be reduced by the approach, an experiment was performed in which a prediction score equivalent to 75% correct movement planning detection and a prediction score equivalent to 100% correct movement planning detection was randomly chosen to simulate the adaptation by embedded brain reading. Both conditions were inter-

leaved with a no adaptation condition ( $T_{th}^{Max}$ ; see  $T\_max$  in Figure 13.19b) chosen to adapt the exoskeleton control. In case of 100% correct movement planning the individually estimated time threshold  $T_{th}^{Max}$  was reduced to  $T_{th}^{Min}$ , i.e., 10 ms (see  $T\_min$  in Figure 13.19b) which is the minimum time that the exoskeleton needs to react to a signal. This minimum response time is caused by the exoskeleton's 100Hz control cycle. In case of 75% correct movement planning, the individually estimated time threshold  $T_{th}^{Max}$  was reduced to a time value between  $T_{th}^{Max}$  and  $T_{th}^{Min}$ , i.e.,  $T_{th}^{Mid}$  (see  $T\_mid$  in Figure 13.19b). Thus,  $T_{th}^{Min}$  was equivalent with maximal adaptation by embedded brain reading,  $T_{th}^{Max}$  with no adaptation by embedded brain reading, and  $T_{th}^{Mid}$  with a medium adaptation by embedded brain reading. Interaction forces under all three conditions ( $T_{th}^{Max}$ ,  $T_{th}^{Mid}$ , and  $T_{th}^{Min}$ ) were measured by the force sensors which were embedded into the exoskeleton. Mean values were calculated for each prediction score value for ten measured movements across five subjects. It could clearly be shown that the interaction force applied over time was reduced (see value for force integral under all three condition), i.e., for more than one third under maximum adaptation compared to no adaptation. Subjects reported that they could clearly feel the differences in transparency of the exoskeleton even in case of medium adaptation, i.e. in case of  $T_{th}^{Mid}$  ( $T\_mid$  in Figure 13.19b).

## Focus questions

- 13.1. What is the difference between explicit and implicit control approaches?
- 13.2. Which three safety levels can be defined for robots?
- 13.3. How can Moore's Law be related to robotics?
- 13.4. When classifying interfaces which two types of behaviors and two types of methodologies can be used?
- 13.5. What is the difference between overt and covert measures of the user?
- 13.6. How does a passive BCI differ from an active BCI?
- 13.7. What is a hybrid BCI?
- 13.8. What is used in physiological computing and to what can it be compared?
- 13.9. How can a speech interface be implemented from the point of view of human measures?
- 13.10. What does "embedded" in "embedded multimodal interface" mean?
- 13.11. What is embedded brain reading?

- 13.12.** How is the VI-Bot application non-intentionally adapted based on covert measures? Explain both approaches.
- 13.13.** Why can the whole be more than the sum of the individual parts when combining different measures in embedded multimodal interfaces?
- 13.14.** What are the two main trends for future embedded multimodal interfaces?
- 13.15.** What is the relevance of the closed-loop design for personalization and machine intelligence?
- 13.16.** How can we design an embedded multimodal interface while preserving the privacy of the individual?

## References

- B. D. Argall, S. Chernova, M. Veloso, and B. Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5): 469–483. DOI: [10.1016/j.robot.2008.10.024](https://doi.org/10.1016/j.robot.2008.10.024). 559
- J. Bachorowksi and M. J. Owren. 2010. Vocal expression of emotion. In M. Lewis et al., editor, *Handbook of Emotions*, pp. 196–210. Guilford Press. 535
- V. Bargsten and J. de Gea Fernández. 2015. COMPI: Development of a 6-DOF Compliant Robot Arm for Human-Robot Cooperation. In *Proceedings of the 8th International Workshop on Human-Friendly Robotics, (HFR-2015)*, Technische Universität München (TUM), Munich, Germany. 531
- S. Barra, A. Casanova, M. Fraschini, and M. Nappi. 2016. Fusion of physiological measures for multimodal biometric systems. *Multimedia Tools and Applications*, pp. 1–13. [10.1007/s11042-016-3796-1](https://doi.org/10.1007/s11042-016-3796-1). 565
- M. Barsotti, D. Leonardis, C. Loconsole, M. Solazzi, E. Sotgiu, C. Procopio, C. Chisari, M. Bergamasco, and A. Frisoli. 2015. A full upper limb robotic exoskeleton for reaching and grasping rehabilitation triggered by mi-bci. In *IEEE International Conference on Rehabilitation Robotics (ICORR)*, pp. 49–56. DOI: [10.1109/ICORR.2015.7281174](https://doi.org/10.1109/ICORR.2015.7281174). 536
- B. Blankertz, G. Dornhege, M. Krauledat, M. Schröder, J. Williamson, R. Murray-Smith, and K.-R. Müller. 2006. The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell. In *Verlag der Technischen Universität Graz*, pp. 108–109. 541
- S. Bozinovski and A. Bozinovski. 2015. Mental states, eeg manifestations, and mentally emulated digital circuits for brain-robot interaction. *IEEE Transactions on Autonomous Mental Development*, 7: 39–51. DOI: [10.1109/TAMD.2014.2387271](https://doi.org/10.1109/TAMD.2014.2387271). 535
- D. Brin. 1999. *The Transparent Society: Will Technology Force Us to Choose between Privacy and Freedom*. Perseus Books. 562
- C. Brunner, B. Blankertz, F. Cincotti, A. Kubler, D. Mattia, F. Miralles, A. Nijholt, B. Otal, P. Salomon, and G. R. Müller-Putz. 2014. BNCI horizon 2020—towards a roadmap

- for brain/neural computer interaction. In *Brain-Computer Interfaces 2*, pp. 1–10. DOI: [10.1007/978-3-319-07437-5\\_45](https://doi.org/10.1007/978-3-319-07437-5_45). 536
- J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson. 2000. Psychophysiological science. In J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, editors, *Handbook of Psychophysiology*, pp. 3–26. Cambridge University Press, Cambridge UK. 544
- C. Canning and M. Scheutz. 2013. Functional near-infrared spectroscopy in human-robot interaction. *Journal of Human-Robot Interaction*, 2(3): 62–84. DOI: [10.5898/JHRI.2.3.Canning](https://doi.org/10.5898/JHRI.2.3.Canning). 535
- A. Cotrina, A. Benevides, A. Ferreira, T. Bastos, J. Castillo, M. L. Menezes, and C. Pereira. 2014. Towards an architecture of a hybrid BCI based on SSVEP-BCI and passive-BCI. In *2014 36th Annual International Conference of the IEEE, pp. 1342–1345. Engineering in Medicine and Biology Society (EMBC)*. DOI: [10.1109/EMBC.2014.6943847](https://doi.org/10.1109/EMBC.2014.6943847). 537
- J. de Gea Fernández and F. Kirchner. 2015. Predictive compliance for interaction control of robot manipulators. In *Proceedings of the International Conference on Intelligent Robots and Systems, (IROS-2011)*, pp. 4134–4140. Technische Universität München (TUM), San Francisco, CA. DOI: [10.1109/IROS.2011.6094476](https://doi.org/10.1109/IROS.2011.6094476). 532
- J. de Gea Fernández, D. Mronga, M. Günther, T. Knobloch, M. Wirkus, M. Schröer, M. Trampler, S. Stiene, E. A. Kirchner, V. Bargsten, T. Bänziger, J. Teiwes, T. Krüger, and F. Kirchner. 2017. Multimodal sensor-based whole-body control for human-robot collaboration in industrial settings. *Robotics and Autonomous Systems*, 94: 102–119. <http://www.sciencedirect.com/science/article/pii/S0921889016305127>. DOI: [10.1016/j.robot.2017.04.007](https://doi.org/10.1016/j.robot.2017.04.007). 533, 543, 558
- D-rock. 2015. <http://robotik.dfki-bremen.de/en/research/projects/d-rock.html>. 559
- S. H. Fairclough. 2009. Fundamentals of physiological computing. *Interacting With Computers*, 21: 133–145. DOI: [10.1016/j.intcom.2008.10.011](https://doi.org/10.1016/j.intcom.2008.10.011). 537
- S. H. Fairclough. 2014. Physiological data should remain confidential. *Nature*, 505:263. DOI: [10.1038/505263a](https://doi.org/10.1038/505263a). 562
- S. H. Fairclough. October 2015. A closed-loop perspective on symbiotic human-computer interaction. In B. Blankertz, G. Jacucci, L. amberini, A. Spagnolli, and J. Freeman, editors, *Symbiotic Interaction: 4th International Workshop, Symbiotic 2015*, Berlin, Germany, pp. 57–67. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-24917-9\\_6](https://doi.org/10.1007/978-3-319-24917-9_6). 557, 561
- M. Folgheraiter, M. Jordan, S. Straube, A. Seeland, S. K. Kim, and E. A. Kirchner. 2012. Measuring the improvement of the interaction comfort of a wearable exoskeleton. *International Journal of Social Robotics*, 4(3): 285–302. DOI: [10.1007/s12369-012-0147-x](https://doi.org/10.1007/s12369-012-0147-x). 540, 541, 544, 546, 568
- G. Friedland and M. Tschantz. 2018. Privacy concerns of multimodal sensor systems. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool Publishers, San Rafael, CA, 563, 564

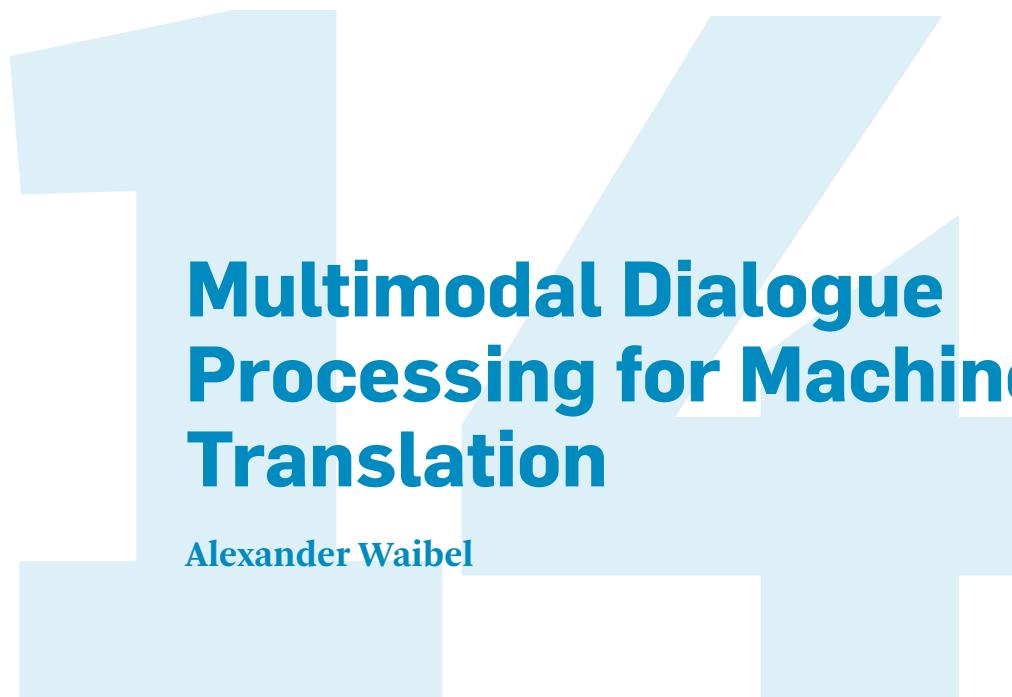
- S. Haddadin. 2015. *Intuitive Interaction with Robots—Technical Approaches and Challenges*, pp. 249–271. Springer Verlag GmbH Heidelberg. 533
- L. J. Hettinger, P. Branco, L. M. Encarnaco, and P. Bonato. 2003. Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. *Theoretical Issues in Ergonomic Science*, 4: 220–237. DOI: [10.1080/1463922021000020918](https://doi.org/10.1080/1463922021000020918). 543
- K. Hung, C. C. Lee, and S.-O. Choy. 2015. Ubiquitous health monitoring: Integration of wearable sensors, novel sensing techniques and body sensor networks. In S. Adibi, editor, *Mobile Health: A Technology Road Map*, pp. 319–342. Springer International Publishing, Cham. DOI: [10.1007/978-3-319-12817-7\\_15](https://doi.org/10.1007/978-3-319-12817-7_15). 540
- IntelCooperation. 2005. Excerpts from a conversation with Gordon Moore: Moore's law. [http://large.stanford.edu/courses/2012/ph250/lee1/docs/Excerpts\\_A\\_Conversation\\_with\\_Gordon\\_Moore.pdf](http://large.stanford.edu/courses/2012/ph250/lee1/docs/Excerpts_A_Conversation_with_Gordon_Moore.pdf). [Online; accessed September 15, 2016]. 532
- P. Kampmann and F. Kirchner. 2014. Integration of fiber-optic sensor arrays into a multimodal tactile sensor processing system for robotic end-effectors. *Sensors—Open Access Journal*, Special Issue Tactile Sensors and Sensing Systems, 14(4): 6854–6876. DOI: [10.3390/s140406854](https://doi.org/10.3390/s140406854). 534
- P. Kampmann and F. Kirchner. May 2015. Towards a fine-manipulation system with tactile feedback for deep-sea environments. *Robotics and Autonomous Systems*, 67(C): 115–121. DOI: [10.1016/j.robot.2014.09.033](https://doi.org/10.1016/j.robot.2014.09.033). 532
- B. H. Kim, M. Kim, and S. Jo. 2014. Quadcopter flight control using a low-cost hybrid interface with eeg-based classification and eye tracking. *Computers in Biology and Medicine*, 51: 82–92. DOI: [10.1016/j.compbiomed.2014.04.020](https://doi.org/10.1016/j.compbiomed.2014.04.020). 538
- S. K. Kim, E. A. Kirchner, A. Stefes, and F. Kirchner. 2017. Intrinsic interactive reinforcement learning—using error-related potentials for real world human-robot interaction. *Scientific Reports*, 7: 17562, 2017. <https://doi.org/10.1038/s41598-017-17682-7>. 557
- E. A. Kirchner and R. Drechsler. 2013. A formal model for embedded brain reading. *Industrial Robot: An International Journal*, 40(6): 530–540. DOI: [10.1108/IR-01-2013-318](https://doi.org/10.1108/IR-01-2013-318). 539, 540, 544, 548, 549, 560
- E. A. Kirchner, H. Wöhrle, C. Bergatt, S. K. Kim, J. H. Metzen, D. Feess, and F. Kirchner. 2010. Towards operator monitoring via brain reading—an EEG-based approach for space applications. In *Proceedings of the 10th International Symposium Artificial Intelligence, Robotics and Automation in Space*, pp. 448–455, Sapporo. 541
- E. A. Kirchner, J. C. Albiez, A. Seeland, M. Jordan, and F. Kirchner. February 2013a. Towards assistive robotics for home rehabilitation. In M. F. Chimeno, J. Solé-Casals, A. Fred, and H. Gamboa, editors, *Proceedings of the 6th International Conference on Biomedical Electronics and Devices (BIODEVICES-13)*, pp. 168–177, Barcelona, Spain. SciTePress. 540, 541, 551
- E. A. Kirchner, S. K. Kim, S. Straube, A. Seeland, H. Wöhrle, M. M. Krell, M. Tabie, and M. Fahle. 2013b. On the applicability of brain reading for predictive human-machine interfaces in robotics. *PLoS ONE*, 8(12): e81732, 12. DOI: [10.1371/journal.pone.0081732](https://doi.org/10.1371/journal.pone.0081732). 541, 544

- E. A. Kirchner, M. Tabie, and A. Seeland. 2014. Multimodal movement prediction - towards an individual assistance of patients. *PLoS ONE*, 9(1): e85060, 01. [10.1371/journal.pone.0085060](https://doi.org/10.1371/journal.pone.0085060). 540, 541, 554, 555
- E. A. Kirchner, J. de Gea Fernández, P. Kampmann, M. Schröer, J. H. Metzen, and F. Kirchner. 2015. *Intuitive Interaction with Robots - Technical Approaches and Challenges*, pp. 224–248. Springer Verlag GmbH Heidelberg. DOI: [10.1007/978-3-658-09994-7\\_8](https://doi.org/10.1007/978-3-658-09994-7_8). 534, 539, 541
- E. A. Kirchner, S. K. Kim, M. Tabie, H. Wöhrle, M. Maurus, and F. Kirchner. 2016a. An intelligent man-machine interface - multi-robot control adapted for task engagement based on single-trial detectability of p300. *Frontiers in Human Neuroscience*, 10:291. ISSN 1662-5161. [10.3389/fnhum.2016.00291](https://doi.org/10.3389/fnhum.2016.00291). 535, 541, 550, 552
- E. A. Kirchner, N. Will, M. Simnofske, L. M. Vaca Benitez, B. Bongardt, M. M. Krell, S. Kumar, M. Mallwitz, A. Seeland, M. Tabie, H. Wöhrle, M. Yüksel, A. Heß, R. Buschfort, and F. Kirchner. 2016b. Recupera-Reha exoskeleton technology with integrated biosignal analysis for sensorimotor rehabilitation. In *Proceedings of the Second Conference on smartASSIST 2016*. 541
- E. A. Kirchner, N. Will, M. Simnofske, L. M. Vaca Benitez, B. Bongardt, M. M. Krell, S. Kumar, M. Mallwitz, A. Seeland, M. Tabie, H. Wöhrle, M. Yüksel, A. Heß, R. Buschfort, and F. Kirchner. December 2016c. Recupera-reha: Exoskeleton technology with integrated biosignal analysis for sensorimotor rehabilitation. In 2. *Transdisziplinäre Konferenz "Technische Unterstützungssysteme, die Menschen wirklich wollen."* *Transdisziplinäre Konferenz "Technische Unterstützungssysteme, die Menschen wirklich wollen,"*, pp. 504–517. Elsevier, Hamburg, Germany. [https://www.dfg.de/web/forschung/publikationen/renameFileForDownload?filename=20161121\\_Recupera-Reha\\_ExoskeletonTechnology with Integrated.pdf&file\\_id=uploads\\_2992](https://www.dfg.de/web/forschung/publikationen/renameFileForDownload?filename=20161121_Recupera-Reha_ExoskeletonTechnology with Integrated.pdf&file_id=uploads_2992). 540
- G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich. 2004. Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intelligent Systems*, 19: 91–95. 560
- M. M. Krell, S. Straube, A. Seeland, H. Wöhrle, J. Teiwes, J. H. Metzen, E. A. Kirchner, and F. Kirchner. December 2013. pySPACE - a signal processing and classification environment in Python. *Frontiers in Neuroinformatics*, 7(40). <https://github.com/pyspace>. [10.3389/fninf.2013.00040](https://doi.org/10.3389/fninf.2013.00040). 558
- F. L. Lewis, D. M. Dawson, and C. T. Abdallah. 2003. *Robot Manipulator Control: Theory and Practice*. CRC Press. 540
- C. Liu, K. Conn, N. Sarkar, and W. Stone. 2008. Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder. *International Journal of Human-Computer Studies*, 66: 662–677. 544
- C. Lucarotti, C. M. Oddo, N. Vitiello, and M. C. Carrozza. 2013. Synthetic and bio-artificial tactile sensing: A review. *Sensors*, 13(2): 1435. URL <http://www.mdpi.com/1424-8220/13/2/1435>. [10.3390/s130201435](https://doi.org/10.3390/s130201435). 532

- C. Lüth, U. Frese, H. Täubig, D. Walter, and D. Hausmann. 2015. Sams sicherheitskomponente für autonome mobile serviceroboter. In *Proceedings ROBOTIK 2008*, Munich, Germany. VDI-Verlag. 533
- M. Mallwitz, N. Will, J. Teiwes, and E. A. Kirchner. 2015. The capio active upper body exoskeleton and its application for teleoperation. In *Proceedings of the 13th Symposium on Advanced Space Technologies in Robotics and Automation*. ESA. 540
- MARS. 2015. Mars - a cross-platform simulation and visualization tool. <http://rock-simulation.github.io/mars/>. Last accessed on April 11, 2015. 549
- J. H. Metzen, A. Fabisch, L. Senger, J. Gea Fernández, and E. A. Kirchner. 2013. Towards learning of generic skills for robotic manipulation. *KI - Künstliche Intelligenz*, 28(1): 15–20. DOI: [10.1007/s13218-013-0280-1](https://doi.org/10.1007/s13218-013-0280-1). 558
- C. A. Miller. 2005. Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods. In *First International Conference on Augmented Cognition*, Las Vegas, NV. 560
- H. Monkaresi, S. Hussain, and R. Calvo. 2014. Using Remote Heart Rate Measurement for Affect Detection. *Florida Artificial Intelligence Research Society Conference*. 534
- G. Müller-Putz, R. Leeb, M. Tangermann, J. Hohne, A. Kubler, F. Cincotti, D. Mattia, R. Rupp, K. R. Muller, and J. Del R Millan. 2015. Towards noninvasive hybrid brain-computer interfaces: Framework, practice, clinical application, and beyond. In *Proceedings of the IEEE*, 103: 926–943. DOI: [10.1109/JPROC.2015.2411333](https://doi.org/10.1109/JPROC.2015.2411333). 537
- K. Mülling, J. Kober, O. Koemer, and J. Peters. 2013. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32: 263–279. DOI: [10.1177/0278364912472380](https://doi.org/10.1177/0278364912472380). 558
- A. Myrden and T. Chau. 2015. Effects of user mental state on EEG-BCI performance. *Frontiers in Human Neuroscience*, 9:308. DOI: [10.3389/fnhum.2015.00308](https://doi.org/10.3389/fnhum.2015.00308). 537
- F. Nijboer, B. van de Laar, S. Gerritsen, A. Nijholt, and M. Poel. 2015. Usability of three electroencephalogram headsets for brain-computer interfaces: A within subject comparison. *Interacting with Computers*, 27: 500–511. DOI: [10.1093/iwc/iwv023](https://doi.org/10.1093/iwc/iwv023). 535
- D. Novak, X. Omlin, R. Leins, and R. Riener. 2013. Predicting targets of human reaching motions using different sensing technologies. *IEEE Transactions on Biomedical Engineering*, 60(9): 2645–2654. <http://www.biomedsearch.com/nih/Predicting-targets-human-reaching-motions/23674417.html>. DOI: [10.1109/TBME.2013.2262455](https://doi.org/10.1109/TBME.2013.2262455). 558
- Y. Panagakis, O. Rudovic, and M. Pantic. 2018. Learning for multimodal and context-sensitive interfaces. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 554
- P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal. 2009. Learning and generalization of motor skills by learning from demonstration. In *2009 IEEE International Conference on Robotics and Automation*, pp. 763–768. IEEE. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5152385>. DOI: [10.1109/ROBOT.2009.5152385](https://doi.org/10.1109/ROBOT.2009.5152385). 558

- S. K. Piper, A. Krueger, S. P. Koch, J. Mehnert, C. Habermehl, J. Steinbrink, H. Obrig, and C. H. Schmitz. 2014. A wearable multi-channel fnirs system for brain imaging in freely moving subjects. *Neuroimage*, 85(1): 64–71. DOI: [10.1016/j.neuroimage.2013.06.062](https://doi.org/10.1016/j.neuroimage.2013.06.062). 535
- C. Reynolds and R. W. Picard. 2005. Affective sensors, privacy and ethical contracts. In *CHI'2005*, Vienna, Austria. DOI: [10.1145/985921.985999](https://doi.org/10.1145/985921.985999). 562
- ROCK. 2015. Rock - robot construction kit. <http://robotik.dfki-bremen.de/en/research/softwaretools/rock-1.html>. Last visited on 28.02.2016.
- M. Rommerman, D. Kühn, and F. Kirchner. 2009. Robot design for space missions using evolutionary computation. In *Evolutionary Computation, 2009. CEC '09. IEEE Congress on*, pp. 2098–2105. [10.1109/CEC.2009.4983200](https://doi.org/10.1109/CEC.2009.4983200). 549
- SafetyEye. 2014. SafetyEye from PILZ. [http://brochures.pilz.nl/bro\\_pdf/SafetyEYE\\_2014.pdf](http://brochures.pilz.nl/bro_pdf/SafetyEYE_2014.pdf). Accessed August 16, 2016. 533
- S. Schaal. 1997. Learning from demonstration. In *Advances in Neural Information Processing Systems 9*, pp. 12–20. MIT Press. 559
- B. Schuller. 2018. Multimodal user state and trait recognition: An overview. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 534
- L. Senger and E. A. Kirchner. 2016. Automatic Detection and Recognition of Human Movement Patterns in Manipulation Tasks. In *Proceedings of the International Conference on Physiological Computing Systems (PhyCS 2016)*, Lisbon, Portugal. SCITEPRESS Digital Library. 558
- N. Serbedzija and S. H. Fairclough. 2012. Reflective pervasive systems. *ACM Transactions on Autonomous and Adaptive Systems*, 7. 556
- S. Straube, M. Rohn, M. Roemermann, C. Bergatt, M. Jordan, and E. Kirchner. 2011. On the closure of perceptual gaps in man-machine interaction: virtual immersion, psychophysics and electrophysiology. *Perception*, 40 ECV Abstract Supplement:177. 547
- C. Vogel, C. Walter, and N. Elkemann. 2013. A projection-based sensor system for safe physical human-robot collaboration. In *IROS*, pp. 5359–5364. IEEE. 533
- M. Wand, C. Schulte, M. Janke, and T. Schultz. 2013. Array-based electromyographic silent speech interface. In *6th International Conference on Bio-inspired Systems and Signal Processing*. [http://www.csl.uni-bremen.de/cms/images/documents/publications/BS13\\_WandSchulteJankeSchultz\\_ArrayBasedEMGSSI.pdf](http://www.csl.uni-bremen.de/cms/images/documents/publications/BS13_WandSchulteJankeSchultz_ArrayBasedEMGSSI.pdf). BIOSIGNALS 2013. 535
- H. Wöhrle and E. A. Kirchner. 2014. Online classifier adaptation for the detection of p300 target recognition processes in a complex teleoperation scenario. In H. P. da Silva, A. Holzinger, S. Fairclough, and D. Majoe, editors, *Physiological Computing Systems*, volume 8908 of *Lecture Notes in Computer Science*, pp. 105–118. Springer Berlin Heidelberg. 541, 548

- H. Wöhrle and F. Kirchner. 2015. Reconfigurable hardware-based acceleration for machine learning and signal processing. In R. Drechsler and U. Kühne, editors, *Formal Modeling and Verification of Cyber-Physical Systems: 1st International Summer School on Methods and Tools for the Design of Digital Systems*, Lecture Notes in Computer Science, chapter Springer Fachmedien Wiesbaden. Springer. [558](#)
- H. Wöhrle, J. Teiwes, M. Tabie, A. Seeland, E. A. Kirchner, and F. Kirchner. 2014. Prediction of Movements by Online Analysis of Electroencephalogram with Dataflow Accelerators. In *Proceedings of the International Congress on Neurotechnology, Electronics and Informatics (NEUROTECHNIX 2014)*, pp. 31–37, Rome, Italy. ScitePress. [559](#)
- J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6): 767–791. [536](#)
- C. Yongwook, J. Jaeseung, and J. Sungho. 2012. Toward brain-actuated humanoid robots: Asynchronous direct control using an EEG-based BCI. *IEEE Transactions on Robotics*, 28: 1131–1144. [536](#)
- T. O. Zander and C. Kothe. 2011. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*, 8:1–5. [536](#)
- T. O. Zander, J. Bronstroo, R. Lorenz, and L. R. Krol. 2014. Towards BCI-based implicit control in human-computer interaction. In S. H. Fairclough and K. Gilleade, editors, *Advances in Physiological Computing*, pp. 67–90. Springer-Verlag. [537](#)
- J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad. 2018. Multimodal behavioural and physiological signals as indicators of cognitive load. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. [548](#), [549](#)



# Multimodal Dialogue Processing for Machine Translation

Alexander Waibel

## 14.1

### Introduction

Humans converse with each other to communicate and to develop ideas interactively in the presence of imprecise and under-specified information. In an increasingly multicultural world, such communication of ideas necessitates communication across language boundaries. With more than 7,000 languages spoken on our planet, however, such boundaries cannot be overcome by language learning or human translation effort alone and require technical solutions that can help mediate between humans and machines. To be effective, such mediation cannot be accomplished by text translation alone, as human communication expresses itself in several modalities. Speech, discourse, dialogue, handwritten or texted text, road signs, gestures, eye gaze, and facial expressions all participate in human communication and complement text as an expression of thoughts and ideas, so that our messages must be transmitted multimodally across language barriers as well.

Among those modalities, speech may perhaps be the most important (next to text), because we express ourselves in multiple languages and that requires us to translate language in its spoken as well as textual form. Technologies that aim to take on such cross-lingual interpretation duties of speech are commonly known as speech-to-speech translators. In the following, we will begin with a discussion of speech-translators and their underlying technology. We will then show how their design and realization must be closely matched to their intended use case and how they must be fieldable and adaptive to respond to the needs of their deployment.

### Glossary

- Automatic speech recognition:** the signal spoken in language is recorded by microphone, processed, and converted to text (speech to text).
- Code switching:** mixing words from different languages, declination rules and compounding.
- Consecutive interpretation** typically interprets a few sentences, one at a time, before giving the dialogue partner a chance to respond.
- Cross-lingual subtitling:** a mixture of consecutive and simultaneous interpretation where interpretation is performed on media content and delivered textually as subtitles.
- Earplugs and pixel-buds:** a set of ear-plugs provides input and output for a speaker attempting to dialogue with others.
- Electromyography:** electrodiagnostic medicine technique for recording the electrical activity produced by muscles.
- JANUS system was the first speech translation system presented to the public in the USA and Europe in 1991.
- Linguistic scalability/portability.** Implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet.
- Neural machine translation:** greater abstraction and greater ease of integration is obtainable through neural translation approaches, where internal (“hidden”) abstractions are generated as a side-effect of training many layers of neural structures.
- Out-of-vocabulary words (OOVs):** when words are missing in the pronunciation dictionary of a recognizer, leading to one or more substitution errors. Named entities and specialty terms are particularly prone to this type of problem.
- Simultaneous interpretation** attempts to recognize and translate spoken language in parallel to the input speech without making the speaker pause.
- Speech synthesis:** text in the target language is output in spoken language (text to speech).
- Speech translation goggles translate output from a simultaneous (lecture) translation system delivered textually via heads-up display goggles.
- Statistical machine translation:** greater speed of learning and better performance and generalization to broader topics, but still requires collections of large parallel corpora.
- Targeted audio:** synthetic speech output in a speech translation system delivered selectively by directional loudspeakers.
- Text-to-speech synthesis (TTS):** TTS makes translated sentences audible in the target language and thus permits full speech-to-speech dialogues between two participants.

We will then also consider additional modalities and flexibilities between them in view of developing such seamless and language-transparent communication tools.

In its most direct form, a speech-to-speech translator could be constructed by combining a speech recognition engine (speech-to-text (STT)) with a machine translation (MT) engine, so as to translate a spoken sentence from language A to language B. If a response from a speaker in language B is to be translated into language A, we will also need recognition and translation engines in the reverse direction. Decomposing the problem in this fashion, however, vastly oversimplifies the problem of cross-lingual communication.

If we recall that the goal of cross-lingual communication and dialogue is to effectively *communicate* ideas, several orthogonal dimensions emerge that we must carefully consider to achieve a thoughtful and effective design.

**1. Spoken Language.** The first set of such issues pertains to the problems associated with translation of *spoken language*.

- *Errors.* Speech recognizers make errors and translation engines must be robust against such errors or attempt to correct for them.
- *Spoken language.* Speech is disfluent and hardly corresponds to syntactically well-formed text. Machine translation must therefore be adapted and trained for spoken language instead of text.
- *Punctuation, casing, and disfluencies.* Human speech misses punctuation markers and casing, which otherwise provide important clues for translation. Instead, speech contains an array of potentially confusing disfluencies (hesitations (ah, um, uh, er, etc.), false-starts, and fragments).
- *Prosody.* Speech (unlike text) encodes additional information by way of pitch, intensity and rhythm, which transmit meta-level signals, such as emotion, gender of the speaker, emphasis, discourse information, social cues, degree of formality, etc.

**2. Interaction Style.** The second dimension pertains to the type and style of translation that depends on the situation and use case.

- *Consecutive interpretation* typically interprets a few sentences, one at a time, before giving the dialogue partner a chance to respond, again with a short utterance of one or a small number of sentences. Processing can be more accurate and communication more effective in a face-to-face dialogue situation, as both participants are always aware of the mediation provided by translation and are thus generally more

cooperative. Also, interactive error handling can be employed. Consecutive interpretation, however, introduces a delay that slows down communication. Typical use cases are given by pocket translators, or bi-directional dialogue translators.

- Consecutive interpretation *in combination with dialogue processing* aims to emulate the ability of human interpreters and to carry out monolingual dialogues in addition to interpreting between the languages.<sup>1</sup> In this way, certain transactions can be handled in a more compressed manner monolingually and some are communicated via interpretation [Oviatt and Cohen 1992]. A system design involves maintaining two linked dialogues with an interpreter, one in each language. The interpreter is a dialogue participant, who translates some of what is said, but might also answer questions directly (i.e., without translation), since they may already have been told the answer. In this fashion, repeated requests or clarifications can be handled by monolingual dialogue, and do not require the full round-trip to the other language.
  - *Simultaneous interpretation* attempts to recognize and translate spoken language in parallel to the input speech without making the speaker pause. This mode of interpretation can be faster, and generates less interference to the speaker. It is more challenging, however, as speakers tend to be less aware of the interpretation efforts, and cannot participate in resolving errors. It must also trade-off context (and thus accuracy) against the latency between the spoken and translated words. Typical use cases are the interpretation of lectures or speeches.
  - *Cross-lingual subtitling* is a mixture of the above where interpretation is performed on media content and delivered textually as subtitles. The input speech is typically less disfluent (prepared speech) than lectures or speeches; latency may be less of a concern and in some instances error processing may be possible.
3. **The third dimension is concerned with the form of delivery.** As we aim for *effective communication*, we cannot limit ourselves to recognizing and translating spoken sentences. If the goal is to get one's point across with minimal interference and minimal delay, we must also be concerned with a

---

1. Such as, for example, human interpreters on AT&T Language Line (see [Oviatt and Cohen 1992])

most effective human interface design and multimodal strategies. Thus, we must also consider the following.

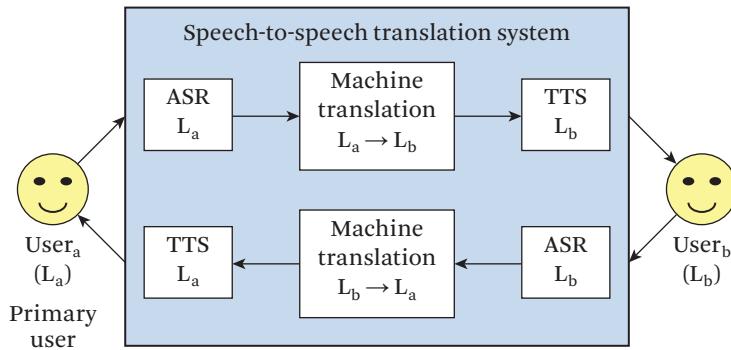
- *Multimodal input and output.* To optimize efficiency of *communication*, it is often more effective to switch or combine multiple modalities, such as speaking, texting, typing, images, handwriting, gesturing, pointing. Output can also be produced alternatively by synthetic speech or as text depending on situation and delivered on smartphones, tablets, in heads-up display goggles or by targeted audio speakers.
- *Error handling and multimodal error repair.* Speech recognition and machine translation will always produce errors and so it is essential for effective communication to detect and correct errors in the most effective manner. Errors can, for example, be flagged visually on a screen or articulated verbally and corrected by dialogue or multimodal repair.
- *Field-adaptable and extendable systems.* Languages and vocabularies change, and interpreting dialogue systems must evolve alongside such changing languages and vocabularies and adapt to any given dialogue scenario. The situations are too numerous to predefine vocabularies and language use once and for all *a priori*. Effective systems must provide mechanisms that allow (non-expert) users to perform such adaptations in the field and during use.

In this chapter, we begin with an introduction to the technology of interpreting systems. We then review use-cases and deployed systems in use today. Finally, we discuss the science and art of multimodal interface design that make such systems effective in the field.

## 14.2

### Technology

The components technologies of speech-to-speech translators and their performance are subject of much research in computer speech and machine translation communities. While different use cases (as discussed in the previous section) require different configurations (see Section 14.4), let us first consider a typical two-way speech-to-speech interpretation system (see Figure 14.1). For a human being, speaking in one language to understand another human being speaking in another language (depending on use-case, up to), three partial tasks have to be solved (possibly in two or more language directions).



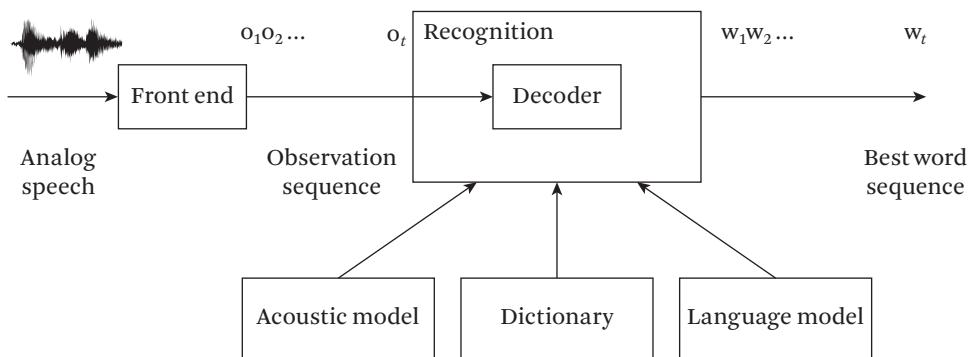
**Figure 14.1** Translation of spoken language (speech to speech translation)—overview.

1. **Automatic speech recognition.** Here, the signal spoken in language ( $L_a$ ) is recorded by microphone, processed, and converted to text (speech to text);
2. **Machine translation** Here, text in language  $L_a$  is translated into text in the other language ( $L_b$ ) (text to text).
3. **Speech synthesis ( $L_b$ ).** Here, text in the target language  $L_b$  is output in spoken language (text to speech). For a dialogue between persons speaking two languages, this process also has to be possible in the other direction (from  $L_b$  to  $L_a$ ) and, hence, requires analogous subsystems in the other language. A final integration of these subsystems with a comfortable user interface then has to be operable easily in real communication situations.

Each of these partial tasks represents an area of research, which over the years has been harder to solve than might be expected by the casual observer due to the complexity and ambiguity of human language. For this reason, they have been studied by scientists for several decades and are still challenging in spite of the considerable progress achieved. The most important lessons learned are that (1) because of inherent ambiguity and errors, we can never make hard decisions, but only soft probabilistic ones for every source of knowledge in human language, and (2) because of their complexity, we cannot encode these statements and their interactions manually, but must learn them from data.

### 14.2.1 Automatic Speech Recognition (ASR)

For the unaware observer, the problem of speech recognition may not appear very difficult at first, as we human beings manage it well and easily. However, several ambiguities occur in spoken language already: The English acoustic sequence of



**Figure 14.2** A typical speech decoder (speech to text).

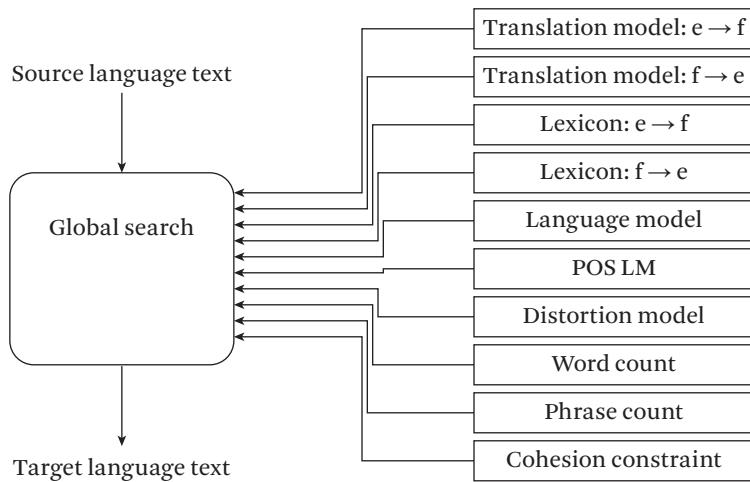
sounds “yu-thu-nā-zhu” may mean both “Euthanasia” and “youth in Asia.” Sentences like “This machine can recognize speech” are pronounced in the same way as “This machine can wreck a nice beach.” Speech recognition requires an interpretation as to which of several similar alternatives is the more meaningful or more probable one in a given context. In modern speech recognition systems, this is achieved by a combination of acoustic models that assign a probability to every sound, a pronouncing dictionary (that assigns a pronunciation to every word), and a language model that evaluates the probability of every possible word sequence “w<sub>1</sub>, w<sub>2</sub>, . . . ” of the sentence. Figure 14.2 shows such a typical decoder. Evaluation of these models during recognition and settings of the best parameters of these models, however, cannot be determined manually, but require automatic search and optimization algorithms.

Parameters of acoustic and linguistic models are determined with the help of machine learning algorithms using huge databases of speech samples, whose transcriptions are known.

Algorithms work with statistical optimization methods or neural networks and learn the best match between signals and symbols (context-dependent phonemes and words) based on known exemplary data. Today’s systems use neural networks in each of these models with several millions of neural links optimized by the learning algorithm.

### 14.2.2 Machine Translation (MT)

First attempts to translate texts by machines (MT = machine translation) were made as early as during the second World War, but early systems attempted to encode all requisite knowledge by rules and failed due to the ambiguity of language and

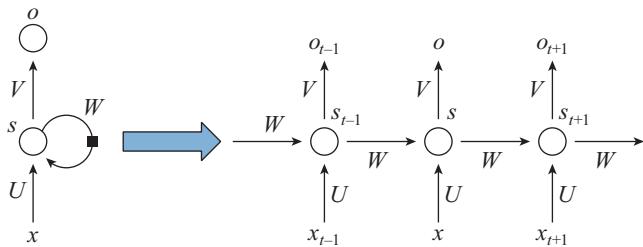


**Figure 14.3** Statistical machine translation (text to text).

the complexity of required associated context knowledge. Nearly every word (skate, row, mouth) has several meanings and, hence, translations can only be interpreted correctly in context. MT folklore recounts that the sentence from the bible “The spirit is willing but the flesh is weak” was supposedly translated into Russian by an early machine translator as “The vodka is good but the flesh is rotten.”<sup>2</sup> Also, language structure is frequently ambiguous. For instance, what does the pronoun “it” refer to in “If the baby doesn’t like the milk, boil it”? Most likely the author meant boiling the milk (not the baby!) and hence the pronoun should be translated into German as “sie” and not “es.”

The attempt to manually encode all required syntactic, semantic, and lexical knowledge with the help of rules would generally not scale (beyond well-defined contained domains). With the arrival of faster and more powerful computing platforms and larger data-resources on the internet, rule-based approaches eventually gave way to automatic learning systems. Modern MT system now use system architectures that optimally trained statistical knowledge sources (see Figure 14.3), or arrangements of recurrent neural network encoder/decoder structures [Kalchbrenner and Blunsom 2013, Sutskever et al. 2014, Bahdanau et al. 2015].

2. The example is due to an early article on MT from the *New Yorker* but it is uncertain if this confusion ever actually occurred in an actual machine translation system.

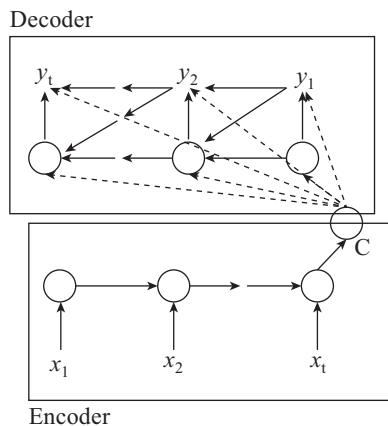


**Figure 14.4** Recurrent neural nets, unfolded in time.

**Statistical Machine Translation** offers greater speed of learning and better performance and generalization to broader topics, but still requires collections of large parallel corpora. However, they still have to be trained one language-pair at a time and cannot easily abstract across languages or include more varied information sources (e.g., prosody, meta-information, etc.) without ever more complicated combinations of individual models.

**Neural Machine Translation.** Greater abstraction and greater ease of integration is obtainable through neural translation approaches, where internal (“hidden”) abstractions are generated as a side-effect of training many layers of neural structures. Generally, they are today implemented as recurrent networks that encode sentences by presenting words (or some compact representation of them) in sequence, and then decoding them in sequence in the other language. A recurrent neural network (RNN) and its sequential unfolding is shown in Figure 14.4. As before outputs  $O$  are generated from inputs  $X$ , but also influenced by the state of the net. In Figure 14.4 we see such a recurrent net unfolded in time. Here a neural net produces a sequence of outputs. The output at timestep  $t$  ( $o_t$ ) is based on the input  $x_t$ , but also from the state of the net at the previous timestep  $s_{t-1}$ . With words represented as vectors as input  $x_t$ , a recurrent neural network can remember sequential information in this recurrent state. Once an entire sentence is *encoded* in this manner, the remaining context vector can then be used to *decode* a sentence in another language, as shown in Figure 14.5. Recurrent encoder-decoder models as shown in Figure 14.5 were attempted for neural dialogue modeling and machine translation as early as the late 1980s [Miikkulainen and Dyer 1989, Jain et al. 1989] and early 1990s [Wang and Waibel 1991], and more recently [Cho et al. 2014a, Sutskever et al. 2014].

Early RNN-based encoder-decoder models had limited success for MT, however, because the recurrences in an RNN tend to remember only recent information



**Figure 14.5** Recurrent encoder-decoder for neural MT.

(words) and forget the earlier context. This is a problem, particularly, when translation requires long-distance reordering between words. Several modifications of the models were proposed to prevent such forgetting in RNNs. The addition of an “attention” mechanism, finally, was shown to be effective to overcome this limitation. It permits different words from the input word sequence to be weighted appropriately (“attention”) for each word in the output sequence (see Figure 14.5b) [Bahdanau et al. 2015]. The attention mechanism was found to yield considerable improvements in MT performance, particularly for language pairs that involve long-distance reorderings (e.g., German). Most recently, it was found that this attention mechanism is indeed sufficient for high performance even without a recurrent state model [Vaswani et al. 2017].

Using attention to represent long-distance relationships and context in language, neural machine translation (NMT) networks now achieve such dramatic improvements over statistical methods that they have all but replaced statistical machine translation (SMT) as the method of choice in MT. Moreover, in addition to superior performance, the practical advantage of NMT is that it is possible to train networks over multiple languages at the same time, so that a single network architecture can serve multiple languages and language *pairs* at the same time, by way of learning some kind of internal semantic representation for all! [Ha et al. 2016, Johnson et al. 2017]. The practical impact (in addition to the better performance) cannot be overstated. At 7,000+ languages in the world, simplifying extensions to new languages and language pairs is key to scaling machine translation, globally. Moreover, as neural abstractions are learned just from data, the input/output to

such networks also does not need to be words alone. They can be acoustic features, meta-level information, or even images. Indeed, cross-modal translation from video to text and vice versa is now a subject of intense research. Automatic descriptions of visual scenes or video generation from text are likely applications.

#### **14.2.3 Speech Synthesis (Text to Speech (TTS))**

If a speech-translator is to output speech in another language, the third component is created by text-to-speech synthesis (TTS). TTS makes translated sentences audible in the target language and thus permits full speech-to-speech dialogues between two participants. In comparison to automatic speech recognition (ASR) and MT, TTS synthesis is generally considered to be a simpler problem, as only one signal has to be produced from a textual sentence to be understandable and it is not necessary to handle the great breadth of ambiguities of the other components. Still, open and important issues exist; however, that continues to be a subject of research. These include improving the language portability of TTS subcomponents through machine learning, to reduce the cost and effort to build TTS systems for more languages. Voice conversion (to adapt the output voice to an input speaker) is also a topic of interest. And critically, better prosodic control of output is needed, so that more suitable emotional emphasis, tone of voice, dialogue context, social setting, level of formality, gender, social roles of speaker and addressee, and other such factors that affect a conversation can be better situated and synthesis thus delivered.

#### **14.2.4 Machine Learning, Statistics, and Neural Networks**

All three components of a speech translation system are now powered by systems that are built by exploiting machine learning, both to deal with ambiguity as well as to learn automatically from data as opposed to a developer's writing rules following introspection. Crucial to their success: the dramatic growth in available data resources (mostly over and through the internet) and available computing resources. These resources have led to a rethinking and replacement of the dominant learning paradigm from statistical modeling back to neural network models, i.e., the models that had already been explored in the 1980s. Neural models that are almost identical to those developed during the late 1980s [[Waibel et al. 1989](#), [Waibel 1989](#), [Bourlard and Wellekens 1989](#)], now show their advantage fully as they are trained over several orders of magnitude larger databases and they now deliver up to 30% relative performance gains in speech recognition and MT performance. At the time of this writing, neural "deep learning" models are rapidly replacing statistical models as the dominant approach for speech recognition, MT, and speech synthesis

[Zenkel et al. 2017, Zweig et al. 2016, Miao et al. 2015, Sennrich et al. 2016, Cho et al. 2016, Neubig 2016]. Systems that include multimodal signals, generalize across many languages, and systems that could train directly end-to-end, from speech to speech, may be possible and become reality soon.

## 14.3

### **Evolution of System Prototypes and Deployments**

The development of automatic spoken language translation systems started in the early 1990s when ASR, MT, and TTS systems first reached a minimum degree of maturity required to attempt a first integration [Waibel et al. 1991, Morimoto et al. 1993, Wahlster 1993]. In the course of the following two decades, major limitations in technology were overcome in a number of research and development phases. Today, speech translators have entered commercial and public usage and can be used by everyone. In the following, we review the different technological milestones, phases of maturity, and the use cases and key deployments they enabled (see Table 14.1 for an overview of system qualifications).

#### **14.3.1 First Demonstrations**

The JANUS system was the first speech translation system presented to the public in the U.S. and Europe in 1991 (see Figure 14.6). JANUS was developed for German, Japanese, and English by Universität Karlsruhe in Germany and Carnegie Mellon University in Pittsburgh, PA, USA. It was a result of cooperation with the ATR Interpreting Telephony Laboratories in Japan, which developed similar systems for the Japanese language in parallel. The systems together were presented in the first translating video conferences [Waibel et al. 1991, Handelsblatt 1991, Morimoto et al. 1993].

These systems represented first steps, managed an initially small vocabulary (< 1000 words), required a relatively restricted syntax, and covered a limited domain (e.g., registration for a conference). They were too large and slow to really be of assistance in field situations, e.g., to a traveler. Similar demonstration systems were presented by other research groups—AT&T [Roe 1992] and NEC [Hatazaki et al. 1992].

#### **14.3.2 Research Systems and Prototypes**

For these systems to be used in practice, other important phases of development followed to successively master difficult problems:

**Spontaneous Speech, Domain-limited Research Systems.** To implement practical systems, the assumption of syntactic correctness has to be eased or eliminated. People

**Table 14.1** Development phases of speech translation systems.

	Years	Vocabulary	Speaking Style	Domain	Speed	Platform	Example Systems
First Dialogue Demonstration Systems	1989–1993	Restricted	Constrained	Limited	2–10× RT	Workstation	JANUS-1 (ATR, CMU), C-STAR-I, NEC, ATT
One-Way Phrasebooks	1997–Present	Restricted, Modifiable	Constrained	Limited	1–3× RT	Handheld	Phraselator, Ectaco
Spontaneous Two-way Systems	1993–Present	Unrestricted	Spontaneous	Limited	1–5× RT	PC/Handheld Devices	JANUS-III, C-STAR, Verbmobil, Nespole, Babylon, Transac
Translation of Broadcast News, Political Speeches	2003–Present	Unrestricted	Read/Prepared Speech	Open	Offline	PC's, PC-Clusters	NSF-STRIDUST, EC TC-STAR, DARPA GALE,
Simultaneous Translation of Lectures	2005–Present	Unrestricted	Spontaneous	Open	Realtime	PC, Laptop	KIT/CMU-Lecture Translator
Commercial Consecutive Translators on a Mobile Phone	2009–Present	Unrestricted	Spontaneous	Open	Online and Offline	Smartphone	Jibbigo, Google, Microsoft,
Simultaneous Interpretation Services	2012–Present	Unrestricted	Spontaneous	Open	Realtime, Online	Server, Cloud-Based	KIT-Lecture Translator, EU-BRIDGE, Microsoft
Consecutive Interpreting Telephony Services	2015–present	Unrestricted	Spontaneous	Open	Realtime	Server, Cloud-Based	Microsoft Skype



**Figure 14.6** First speech translation prototypes in video conferences (1991).

rarely speak syntactically correct and complete sentences. They rather speak fragmentary segments with stammering, repetitions, filler words, and hesitations (er, hum, aeh, etc.). These fragments first have to be identified correctly and then filtered out or corrected by processing before translation takes place. First, spontaneous speech translation systems were developed from 1993–2000 [Morimoto et al. 1993, Takezawa et al. 1998]. These systems were still slow and required extensive hardware. Their domain continued to be too limited to extract the fragments relevant to translation by modeling the semantics. JANUS-III, C-STAR Systems, VERBMOBIL, and other projects made considerable progress, but still remained unusable in practice [Lavie et al. 1997]. Domain limitation and vocabulary restrictions had to be overcome first and systems had to be accelerated and readied for mobile use. In due course, manually programmed approaches (possible in limited domains) were replaced by automatically learned, statistic subsystems that scaled better to larger domains, and improved robustness and accuracy [Brown et al. 1993, Och and Ney 2004, Wang and Waibel 1997, Koehn et al. 2007]. Smartphones and cloud computing offered platforms that could perform these tasks in real-time and were accessible by a broad audience of users.

Two types of applications, serving different use cases began to emerge.

- The first is given by mobile devices that provide consecutive interpretation in human-human interactive dialogues. Here, speakers converse through an interpretation system that translates sentences consecutively. A speaker says one or more sentences in one language, followed by the system's translation. Then the other party responds in another language followed by translation back to the first speaker's language. Consecutive translation slows the flow of a conversation (because speakers have to wait for a translation to complete), but they make interpretation controllable and observable, and (in case of errors) speakers can intervene to make themselves understood. For most applications (tourism, medical uses, humanitarian aid, etc.) a vocabulary of about 40,000 words is sufficient to cover most conversational needs. But systems have to run on small mobile devices, which requires either fast cloud-based operation via telephone networks or compact-efficient implementations on the device.
- The second is given by simultaneous interpretation for stationary use: in many deployments of speech translation, a dialogue between two conversation partners is actually not needed but rather a fast interpretation of a stream of speech (or monologue) is desired. For example, TV broadcasts, internet videos, lectures, speeches, and addresses all require no response. In most of these deployments mobility is less of a concern, as the actual processing can be performed in the cloud on powerful servers. Simultaneous interpretation, however, is complicated by a broader range of vocabularies and special terms, and by the absence of obvious sentence markers. The system itself has to determine the beginnings and end of translatable units or sentences, and—in the case of simultaneous interpretation—must deliver translation output with little delay, before a speaker is finished speaking. Segmentation into units or translatable fragments have to be performed automatically and punctuation (full stops, commas, question marks) inserted automatically based on partial context (The Economist 2006). Statistical and neural models perform these predictions, and display interfaces must manage updates when further context requires revision.

### 14.3.3 Translation of Deployments and Services

Early research systems (1990–2005) solved technical problems and paved the way for the sales and real use of speech translation systems in society.

#### 14.3.3.1 Mobile Consecutive Interpretation Systems

Interpretation systems were first tested in the field during humanitarian and logistic exercises of the U.S. government. Although network-based solutions were proposed, fieldable speech-translators usually required off-line operation, as network access could not be assured (or might be prohibitively expensive) in most humanitarian, logistic, and—indeed—tourist/travel deployments. The resulting systems resorted to laptops, and later PDAs and smartphones with all their speech translation software running on device. Computation was kept within manageable bounds, initially, by limiting the domain of speech-to-speech translation systems to transactional tasks of limited scope (e.g., hotel reservations, scheduling, health-care interviews) or, alternatively, by the use of simple phrase books that would be accessed by voice. Either solution required only smaller vocabularies and could anticipate a more limited language use and thus restrict computation and memory requirements. [Eck et al. 2010, Stüker et al. 2006, Voxtec none, Ectaco 1989]. Early models of such systems offered commercially by VOXTEC and MOBILE TECHNOLOGIES are shown on the left of Figure 14.5. Due to the limitations in vocabulary and hardware, and -in the case of phrase-books- due to the inflexibility of expression, such early systems could achieve adoption only in special situations, were restricted phraseology and limited tasks are acceptable. For the wider use of speech translators by the wider public during travel and communication, further advances were necessary.

With the emergence of smartphones, both the general availability of a suitable platform as well as the necessary computational performance reached the critical capacities that made speech recognition and translation of open unlimited (> 40,000 words) vocabularies embedded on a device in near real time possible.

In 2009, Mobile Technologies (a startup of Carnegie Mellon researchers) launched, Jibbigo, the first domain-unlimited speech-to-speech translation system fully embedded on a phone in 2009 [Eck et al. 2010]. The system found quick adoption and distribution through the simple sales mechanisms of the Apple iTunes app stores and with the growing use of smartphones worldwide, Jibbigo quickly expanded to 15 languages and reached worldwide distribution. Other similar products followed suit, such as systems by Google and Microsoft. While Jibbigo offered a downloadable off-line solution (for a fee—a network-based solution was also available for free), many other entries were and still are exclusively network based. Although network-based solutions can access more powerful computational resources and connect with related internet resources, off-line systems require no roaming fees nor existing infrastructure and are thus preferable in many



**Figure 14.7** First commercial systems: (A) Phraselator, (B) iPaq PDA-based speech translator (2005), (C) Jibbigo, the world's first speech-to-speech translator on a phone (2009). (Phraselator™ by VOXTEC LLC and Speech Translator™ by Mobile Technologies LLC)

humanitarian and travel situations. Jibbigo has thus been used in a number of humanitarian missions and government deployments, where an existing network infrastructure cannot be relied upon (Figure 14.8 A-D, show healthcare initiatives in Thailand, Cambodia, and Honduras for translation between English-speaking physicians and patients speaking other languages).

Typical system configurations may run on iPhones, Android smartphones, or on tablet computers. Tablets were found to be particularly well-suited for face-to-face interaction between partners sitting opposite to each other in medical missions. After five years of development in field situations, the systems were evaluated to perform well in humanitarian missions (MEDCAP—Medical Civil Action Program, Thailand, in 2013 [[Hourin et al. 2013](#)]). It was found that 95% of the interactions during the registration of patients, the conversation could be managed with the sole aid of the automatic tablet interpreter (Jibbigo).

Google and Microsoft followed suit (2013) with translation capability of their own that could be downloaded for off-line use, while broadening the number of languages on offer, making smartphone translators a common tool for today's travelers.



**Figure 14.8** Medical operations in Thailand, Cambodia, and Honduras: (A) translingual dialogues between American physicians and patients in Thailand; (B) medical care with help of the JIBBIGO-speech to speech translator in Thailand; (C) medical operations in Cambodia; and (D) humanitarian operations with Jibbigo in Honduras.

#### 14.3.3.2 Consecutive Interpreting Telephony

Mobile speech translators on smartphones or tablets offer effective and flexible consecutive translation in face-to-face field situations. Of course, consecutive translation can also be used for remote communication over telecommunication networks. Indeed, the earliest prototypes and research projects had envisioned translated video chat services as their use case. For example, the ATR-Interpreting Telephony Laboratories in Osaka, Japan, were already established in 1986 to investigate this possibility, and subsequent public demos together with partners in the U.S. (CMU) and Germany (Siemens, Karlsruhe) demonstrated such video chat sessions as early as 1991. Commercialization of consecutive translation services (human and automatic) followed in the decade since. In the U.S., AT&T established a human interpreting service (AT&T Language Line)<sup>3</sup> to fill consecutive interpreting needs over telephone lines, followed by software-driven services. Consecutive translation

3. <http://www.languageline.com>.

(human or mechanical) as a fee-based service has only been moderately successful, however, and so it was frequently packaged as a feature for video chat service providers, where translation provides broader network reach and contributes to consistent service expansion (a language on/off-ramp of sorts) for operators of communication services. In this manner, an early commercial video chat room enhanced by speech translation was introduced in 2010 by Hewlett-Packard's in its MyRoom Video Chat product and other communication services followed suit.

The broadest and largest telephone and video communication provider today is Skype (by Microsoft), a free voice-over-IP telephony service. In its continuing drive to expand its network, Skype now offers language interpretation through “Skype Translator” (see Figure 14.9), an automatic speech-to-speech interpretation service for human dialogues. Due to its massive and growing user base, Skype Translator represents one of the largest deployments of speech translation. The system accepts speech from speakers in two languages, interprets their messages, and outputs results as speech or text in the other language. A “TrueText” facility cleans up the disfluencies of spontaneous speech and turns them into more readable text. Synthetic output after translation is also overlaid on top of the original speech (at reduced volume) much like voice-overs in TV reporting. The approach helps reduce delays in the consecutive translation of dialogues (Skype calls this approach “ducking”). Skype (as well as other) research teams also experimented with robotic mediators, but Skype found this approach—or at least its implementation—somewhat awkward. Given the text messaging features of Skype, cross-lingual communication is also improved multimodally by allowing the participants to resort to complementary communication modalities, including speech, text, and video. Given the large number of users, a Skype translator<sup>4</sup> can then learn and improve from continued use [Lewis 2015].

#### **14.3.3.3 Simultaneous Interpretation**

In a multi-lingual environment, dialogue between conversation partners speaking different languages is not the only challenge. When thinking of TV news, films, presentations, lectures, speeches, road signs, transparencies for lectures, and short messages, we see many other challenges, where translingual technologies are required.

An important area of application is the interpretation of lectures. In spite of excellent scientific equipment and funding, German universities, for example, are often disadvantaged in international competition for talents, simply because many

---

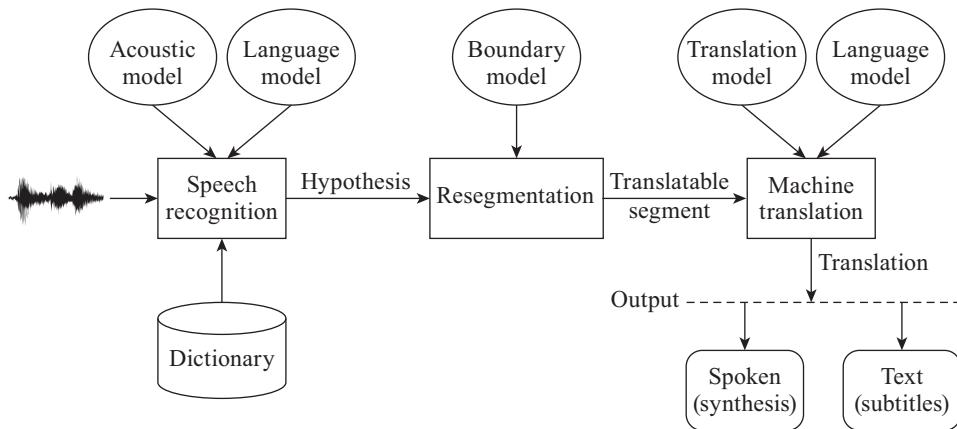
4. <https://www.skype.com/en/features/skype-translator>.



**Figure 14.9** The Skype translator: consecutive interpretation on a telephone network.

foreign students or scientific employees and academics do not want to learn another new language (especially such a difficult language as German). How are German universities or German companies to react? Is a German university supposed to have all courses and lectures presented in English? The author of this article does not consider this desirable or practicable. A hybrid solution with the help of modern language technologies that supports linguistic and cultural diversity and tolerance (and does not suppress one or the other direction) appears to be far more promising, as it fosters and improves internationalization and international understanding.

At Karlsruhe Institute of Technology (KIT), such a system is being used for students in the main auditorium [Cho et al. 2013]. Speech translation continues to be the subject of research, as not all problems are solved. But thanks to continuing benchmarking and competitive evaluations (the IWSLT campaign, EU funded programs like EU-BRIDGE, etc.) consistent improvements and advances can be observed. While output is far from perfect (and falls short of expert human interpreters) for a listener at a University or conference, who does not speak the language of the lecturer, an imperfect computer-based interpreter is better than nothing.



**Figure 14.10** Speech translation of lectures.

The first such system was proposed by researchers at CMU, KIT and Technologies in 2005 (see Figure 14.10)[Fügen et al. 2007]. It manages such a simultaneous interpretation use-case uni-directionally as a monologue to be translated into a target language. Such a system does not have to be run on a mobile device, but may be operated on servers in a cloud-based manner and accessed via the internet. Contrary to a translating system for dialogues, a lecture translator requires just a speech recognition component and a translation machine, if only subtitles are desired. Speech synthesis can take place afterward, but it is optional. In addition, a segmentation component is required to decide explicitly or implicitly when the end of a sentence or at least of a translatable fragment is reached in the stream of words. Segmentation can also be performed incrementally with multiple segmentation hypotheses to be explored in parallel, during execution. Vocabularies containing many technical terms and jargon, foreign words, and expressions, formulas and acronyms present an additional range of problems for lectures.

A lecture translator may be operated in two modes: as a simultaneous interpretation system *during* a lecture and also *afterward* over recorded archival lectures in a post-processing mode for retrieval and review. Simultaneous interpretation is required when a listener wishes to follow along while present in a lecture, and both the input language (transcript) and/or translations can be presented. Simultaneous use requires real-time recognition and translation (i.e., the system has to keep up with the speech). Latency (i.e., the time lag between the spoken word and the translated word) also has to be minimized. Otherwise, the listener will lose track

of the lecture and of what is happening in the lecture hall. For some languages (especially German, as it turns out), these requirements are a challenge, when verbs or important parts of the verb appear at the end of a sentence (or sometimes even later), thus introducing substantial uncertainty when decisions have to be made before a sentence is completed. The verb “vor-schlagen” means “to propose,” and “schlagen” (without the prefix “vor”) means “to hit.” But in a sentence such as “Ich schlage Ihnen nach eingehender Prüfung Ihres Antrags, der uns gestern . . . eine neue Vorgehensweise . . . vor” (translation: I propose to you after considerable review of your proposal a new approach . . .), German syntax strips the leading prefix of “vor-schlagen” off and moves it to the end of the sentence, after potentially many words and minutes of speech later. Appropriate interpretation of German in a low latency mode thus keeps us guessing how the story might end and forces an early translation decision before all the information is in.

In many application scenarios of academic teaching and multimedia broadcasting, offline processing of speech and translation are acceptable and sometimes desirable. Offline operation does not necessarily require real-time capability, although an excessively long processing time may become a relevant cost factor. Furthermore, the system can make a better transcription and translation when taking into account a longer context. A lecture translator, for instance, may be run online in the lecture hall and the output may be reprocessed in offline mode later for storing an improved version for listeners in the archive.

Such a lecture translation system was taken into operation at KIT in 2012 as an internet service in several of its main lecture halls (Figure 14.11) [Cho et al. 2013, Greve-Dierfeld 2012]. Students, who wish to have translation support, connect their phones, tablets or PCs to a course-website via a normal internet web browser and are provided with a simultaneous transcription of the text in German (useful in case of hearing problems) and a translation into English. Output languages include French, Arabic, Spanish, and further languages are under development.

Even though the system is already deployed and in actual use, many linguistic and machine learning problems remain. They continue to be subject of ongoing research. In addition to the problems of word order and verbs discussed above, the following difficulties are encountered (particularly in the German language).

- **Compound words.** German words like “Fehlerstromschutzschalterprüfung” (examination of the protective electric currency malfunction switch) switch first have to be decomposed before they can be translated into English. Algorithms for compound word decomposition have to be developed. Due to the ambiguities of language, however, this, too, is not necessarily easy. While de-



**Figure 14.11** The lecture translator in use in the main auditorium of KIT.

composition into *Fehler-Strom-Schutz-Schalter-Prüfung* in our example may be straight forward, decomposition of “dramatisch” (dramatic) into “Drama-Tisch” (drama table) or of “Asiatisch” into “Asia-Tisch” (asia table) are inappropriate in the context or even change the intended meaning [[Koehn and Knight 2003](#)].

- **“Agreement”.** Suffixes in the German language have to be consistent and agree with the nouns: “in der wichtigen, interessanten, didaktisch gut vorbereiteten, heute und gestern wiederholt stattfindenden Vorlesung” [in the important, interesting, well prepared today, . . . lecture]. The suffixes of each adjective depends on the gender and case of the final noun.
- **Technical terms, jargon and unknown words.** This is a major problem, in particular when processing lectures at a university, because every lecture has its own technical terms and linguistic features. What are “Cepstral-Koeffizienten” (cepstral coefficients), “Walzrollenlager” (roller bearings), and “Würfelkalküle” (cube calculi), and how do we translate them? In order to avoid major dictionary maintenance efforts usable systems must seek out the necessary information from other complementary resources across multiple modalities by themselves. As one effective solution to this problem, automatic algorithms can be devised that search the video and presentation materials of a lecturer for clues to the most likely interpretation of a speaker’s

speech during a lecture (see Waibel, US patents 2013–2018). Technical terms are automatically extracted from the slides and related terms identified on the internet. Unknown words are then added to the recognition vocabulary and translations derived from internet sources, such as Wikipedia [[Niehues and Waibel 2011](#)]. In addition to finding unknown words, performance is improved by cross-referencing spoken language with words and concepts from the corresponding slides.

A second alternative to the problem of unknown words (beyond technical terms these typically also include names, foreign words and abbreviations) is to include human assistance, either by professional editors or the crowd-sourced spontaneous edits from student users. This is done online during a lecture or after the fact in the archive, and the ground truth obtained in this manner, provides further opportunities for machine learning to improve overall system performance over time. Of course, all these methods build on successful interaction with a human user and thus depend greatly on a well-designed multimodal user interface, designed to establishing context and obtain corrections naturally, seamlessly and unobtrusively.

- Code switching. Often, lectures and speeches contain quotations and phrases from other languages. Especially computer science lectures are peppered with English terms that are typically not translated into German. Germans talk about the “iPhone,” “iPad,” “cloud-basiertem Webcastzugriff” (cloud-based webcast access), or “Files,” that are “downgeloaded” thus mixing English words with German declination rules and compounding!
- Pronouns. What do pronouns refer to? Here, problems occur rather frequently. The spoken word version of “Wir freuen uns, Sie heute hier begrüßen zu dürfen” may be translated as “We are happy to welcome her here” or “We are happy to welcome you here” (in writing, Germans use a capital and a small “s” to distinguish both versions).
- Readability. When people speak, they do not speak punctuation marks or the ends or starts of paragraphs contained in readable text. Hence, full stops, commas, question marks, paragraphs, and sometimes even titles have to be generated and inserted automatically [[Cho et al. 2014](#)].
- Spontaneous speech. Different speakers speak more or less syntactically. Hesitations, stuttering, repetitions, and discontinuations of speech aggravate readability and make translation difficult. A spoken sentence of a lecture transcribed by a perfect speech recognition system would contain all such

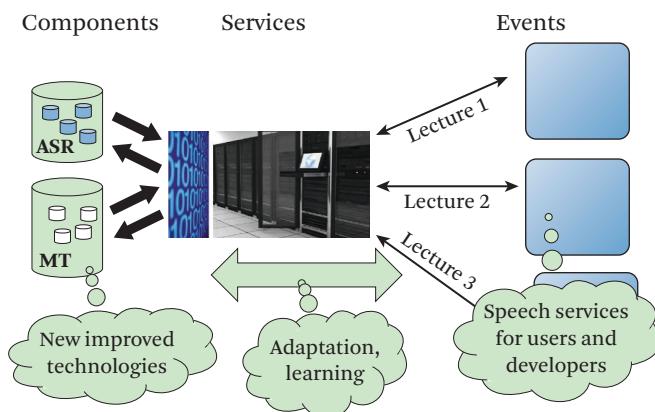
disfluencies and have no punctuation marks. We must therefore process the raw output from speech recognition first linguistically in order to make it readable in the source language. It can then be translated into readable text in the target language [Cho et al. 2014b].

- Microphones and noise. The Karlsruhe lecture translator presently is configured to accept input from a dynamic noise-canceling microphone that the lecturer wears. This is acceptable during lecturing, as lecturers carry microphones in auditoriums anyway, but for seminars and meetings this may be a distraction. Unfortunately, signals from distant microphones or table top microphones are distorted by reverberation, noise, and the potentially overlapping speech from several speakers, which leads to considerable losses in recognition performance.
- Linguistic scalability/portability. How can we implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet? To achieve this, development costs of a translation system would have to be reduced considerably. Language-independent technologies, adaptation, inference, abstraction, better use of monolingual resources, and crowd sourcing (to better harvest the multilingual knowledge of mankind) are promising approaches.

The architecture of the Lecture Translator for practical use was introduced in 2005 as a research prototype between CMU, KIT, and Mobile Technologies, and went into service in 2012 at KIT. Infrastructure techniques and support were developed and merged with other sites under the EU-BRIDGE integrated project of the European Union (see Figure 14.12). Now, several lectures and multiple sites are supported in a cloud-based manner at the same time. In recent developments, Microsoft also launched a similar lecture interpreting service. Its features and capabilities are similar to the one described above. It provides integration with PowerPoint from the Microsoft Office suite to permit subtitling during presentations. Using the merged PowerPoint slides it also merges information from the slides' content in similar ways as described above (Waibel 2013–2018).

By means of this server architecture, translation services can be used in several auditoriums and other application scenarios (not only during lectures at university).

Apart from use at universities (where usually no translation support is provided), automatic systems can also be applied to support experts, for example human interpreters at parliaments. In 2012, 2013, and 2014, the lecture translation system



**Figure 14.12** EU-BRIDGE: The automatic interpreter as a cloud-based service.



**Figure 14.13** Automatically translated speech at the European Parliament.

was presented at the European Parliament, at several Rectors' Conferences (see Figure 14.13), and was featured in training courses for interpreters (see Figure 14.14). The goal of such discussion is to develop possible symbiotic human-machine arrangements that support human interpreters in their efforts to deliver high quality interpretation efficiently. A first test in interpreting booths of the European Parliament was carried out successfully in late 2014 under the EC Integrated Project "EU-BRIDGE" in Strasbourg. [Koehn et al. 2015]



**Figure 14.14** First test of an automatic interpreter during voting at the European Parliament.

Because the European Parliament operates one of the world's broadest government interpretation efforts working continuously with more languages ( $23 \times 23$  language directions!) than any other organization and employs some of the most experienced and sophisticated human interpreters, interpretation services are already at their best in terms of quality and sophistication. Automated language processing and interpretation tools in these settings thus serve a different role from the settings discussed before: rather than performing fully automatic interpretation, they aim to support, enable and *amplify* human effort so as to achieve greater quality, speed, and scale in the face of overwhelming demand.

To date, three use cases were identified that instantiate such complementarity: (1) a generator of on-demand terminology lists and their translation, for example, if a session on “fishery” is scheduled, the system automatically serves up special terms pertaining to their domain and delivers it to the assigned interpreter along with appropriate dictionary lookups; (2) named entity and number tracking (to recall numbers and names more easily); and (3) the “Interpreter’s Cruise Control,” intended to handle repetitive (boring) segments of a session (such as, for example, voting sessions), or where human interpreters are not available. A sophisticated, multimodal interface is essential to deliver such human-machine symbiosis, seamlessly. The services access available resources (schedules, agendas, reports, dictionaries, and lexica) and deliver the desired support to EU interpreters on tablets or laptops. The services exist so far are in a prototype stage, but user studies and evaluations have shown the success of these tools, more than 60% of the interpreters were satisfied or very satisfied with the final tool [Stüker et al. 2015].

## 14.4

### Multimodal Translingual Communication

In a multilingual and multicultural environment, language barriers are not only encountered in spoken dialogues, lectures, or text documents. They occur in many other communication situations, circumstances, and media: for example, important information can be found on road signs or in short text messages (SMS), TV news, lecture transparencies, gestures, and many more. To make the vision of a multilingual, language barrier-free world come true, our efforts have to go beyond the construction of better translation systems. The goal should rather be to build user interfaces that make language barriers transparent or move their existence into the background. Successful translingual communication is achieved, when people can interact with each other without being aware of the barriers between them! In the following, we discuss multimodal system designs, where this was attempted and achieved with varying degrees of effectiveness. The processing of multiple modalities is particularly important and beneficial in two situations: (1) recovering from miscommunications that may result from occasional human misunderstandings or from machine recognition or translation errors; and (2) when responding naturally to multimodal communicative clues in varying multicultural scenarios.

#### 14.4.1 Multimodal Error Handling

Miscommunication in speech dialogue translation can result from errors during the speech recognition or the MT processes, and the causes are often not readily identifiable by the user. Worse, the translation of a misrecognized word rarely bears any resemblance to the translation of the correctly recognized word, so that the resulting output appears just confusing. Recognizing that an error has occurred, offering tools to recover from such errors, and algorithms to even learn from such correction, are subjects for considerable research on learning algorithms and effective multimodal interface design.

Several types of errors can occur in the process of cross-lingual interpretation.

- **Out-of-Vocabulary Words (OOVs).** Most commonly the problem arises when words are missing in the pronunciation dictionary of a recognizer, leading to one or more substitution errors. Named entities and specialty terms are particularly prone to this type of problem.
- **Speech confusions.** Words that are phonetically close (“forest” and “far East”) or homophones (“two” and “too”) can lead to substitution errors due to their acoustic similarity, even though they may differ semantically.

- **Translation Errors.** Words can have multiple translations. While translation and language models attempt to select the most appropriate translation, occasional inappropriate choices remain and need to be corrected.

#### **14.4.1.1 Error Detection**

Before attempting a correction, a problem has to be identified. A speaker may determine that a recognition error has occurred and intervene, if s/he pays careful attention to the transcript but this may not be possible in all situations. Translation errors may even be harder to detect for a speaker who does not know the target language. Two typical solutions are (1) Confidence measures to judge the reliability of the recognition and translation outputs and (2) back translation into a source language so that an input speaker may verify that a translation appears to be correct. A variety of confidence measures exist for ASR and MT engines; most typically compute an *a posteriori* probability of the word to be correct. Although the measures help identify errors in translation, there is unfortunately no guarantee that they will accurately flag problems or that flagged problems are actually errors. During consecutive translation, such methods are also far more likely to succeed, since both speakers have a joint interest in being understood, and have the time and interest to collaboratively resolve potential miscommunications.

#### **14.4.1.2 Error Repair**

How are miscommunications resolved, once they have been identified? Two methods have been proposed in the case of actual misrecognitions or mistranslations: (1) clarification dialogues; and (2) cross-modal repair. In the former, the system will initiate a disambiguation dialogue (triggered by a confidence measure) to get a user to resolve a potential error through a voice dialogue. In the latter, the user (or the system) may divert to another modality to clarify.

Clarification dialogue. In the former approach, once a putative error has been identified, the system attempts to initiate a clarification dialogue with the user. If the system misrecognizes “my name is Edwards,” as “my name is *at words*,” a clarification component might ask for clarification on the misrecognized word, if the error is correctly identified. In such a case, the human user is engaged to disambiguate the confusion through a clarification dialogue “is ATWORD a name?.” Errors in recognition can be caused by OOVs, homophones, or substitutions with similar sounding words, but the detection of errors is a non-trivial classification task in itself. If an error is not recognized as an error, it is missed and cannot be repaired; if a correct word is flagged as an error, it may generate an unnecessary

clarification dialogue and may be a nuisance to the user. Early versions of such clarification dialogues have already been explored in early studies [Block et al. 2000]. More rigorous evaluations using clarification dialogues were conducted under the DARPA program BOLT using simulated field data that investigated the efficacy of language-based error repair dialogues. Even though good performance in detecting and repairing errors was achieved through voice dialogues alone [Kumar et al. 2015], such dialogues take time and are generally much slower than and thus inferior to a multi- or cross-modal repair strategy. If an error is visible on the screen, and alternate input modalities are available, errors can be reliably detected by humans and corrected through typing, gesturing, handwriting or spelling, for example. Thus, unless the use-case is strictly a hands-eye-busy voice situation, multimodal repair appears to be more effective [Suhm et al. 1999, Kumar et al. 2015].

**Cross-Modal Repair.** In cross-modal repair, the error is corrected by diverting to an alternate, hopefully orthogonal modality, such as typing, handwriting, spelling, paraphrasing, etc. The advantage of this approach is that it can proceed in parallel to speaking. Generally, it is thus considerably faster to correct an error by pointing, clicking, and editing, rather than through a disambiguating dialogue [Suhm et al. 1996, 1999, Waibel et al. 1991, 1998a, 1998b, Oviatt and VanGent 1996]. The simplest form of such repair is to simply observe the error and correcting it through typing. Alternatively, however, it is possible to point to the error and spell, hand-write, paraphrase, or respeak a correction. Such correction is fast, potentially more natural, and exploits the orthogonal sources of errors in each modality to obtain a jointly optimal result [Suhm et al. 1999, Waibel et al. 1998a, Oviatt and VanGent 1996].

**Learning Words.** OOVs are particularly troublesome for speech translators, since no matter how the user may correct the input, recognition and translation will fail every time, if the missed word is not included in the processing dictionaries. A typical way to handle OOVs in MT is to simply pass the unknown word through to the other side.<sup>5</sup> If the two languages in question use the same script (say, English and Spanish), this may lead to acceptable results: an unknown name, for example, may appear in the same script as the name in the other language. However, this is not acceptable if the scripts (e.g., Chinese and English) differ. On the recognition side, OOVs are particularly problematic, since their absence from the

---

5. It is worth noting that this approach is no longer so simple in current implementations of NMT, due to the absence of phrase tables.

recognition lexicon will force another match and thus lead to substitutions errors.<sup>6</sup> In a speech translator then, such substituted words will be translated in curious, irrelevant ways that have no resemblance (neither phonetically nor semantically) from the original intended message [Kaiser 2005]. In research speech translation systems, the problem of OOVs is mostly handled by adding the missing words to the various dictionaries and language components manually. This involves a total of eight modifications: the pronunciation dictionary in language L1 has to be provided (“Paul”—[P AO L]), the language model has to be modified to include the “Paul” in a word sequence (e.g., “my name is Paul”), “Paul” has to be translated to “Pablo,” and we may need a pronunciation dictionary to properly pronounce “Pablo” in Spanish. If the system we are building is a bidirectional dialogue system, the appropriate modifications have to be made in the reserve direction as well.

The modifications involve knowledge of phonetics and statistical language models that are easily done in research labs, but they cannot be performed by non-expert users in the field. OOVs (for example, the occurrence of names) found in the field are also not predictable a priori and modifications really must be done by the user. Interactive multimodal interface solutions have thus been proposed [Waibel and Lane 2015, Kaiser 2006] that shield the required technical detail and allow a non-expert to make vocabulary additions in the field, interactively and intuitively. The interface accepts orthography of a word/name to be added, it then generates the appropriate model entries automatically in the background and modifies all system components dynamically. It provides intuitive, interactive sound checks to make sure the pronunciation is correctly represented. When the same name is then uttered again (in either language), it is recognized and translated appropriately. This functionality was extensively tested during humanitarian deployments and in a commercial deployment on a smartphone App (Jibbigo).

#### **14.4.2 Multimodal, *Flexi-modal* Communication**

In cross-lingual communication, multimodal interfaces are not only useful to recover from errors generated by speech recognition or machine translation, they also open up a broad array of cross-lingual communication channels. We recall that our goal is not just speech or text translation, but to provide humanity with a human-human communication experience in which linguistic and cultural barriers become transparent. As such, we must be concerned with the full breadth of

---

6. Here, recent character based neural approaches may offer potential solutions in the future by generating character strings directly without the use of dictionaries [Zenkel et al. 2017, Miao et al. 2015, Zweig et al. 2016].

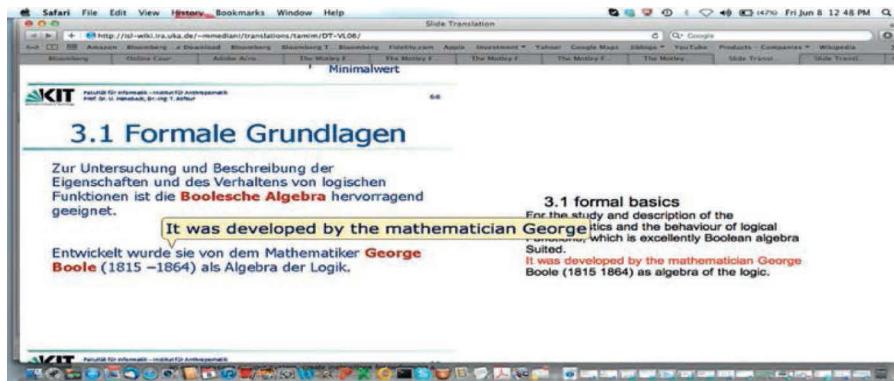


**Figure 14.15** Road sign translator (2001) and Google Translate (Wordlense) 2015.

human expression and assist in their mediation, in whatever modality, context, or situation humans may find themselves. Human language is given by speech, but also images, text, handwriting, even gestures, and facial expressions. Input and output of language may be suitable using one modality in one situation but awkward in another. A successful cross-lingual communication system design must carefully match input and output modalities with devices in each situation.

Over the last 10–20 years considerable progress has been made in achieving this goal.

- **Road sign translators.** As early as 2001 [Yang et al. 2001, Waibel 2002], first systems were developed and commercialized to read and translate road signs with the help of a mobile device and camera. Translations were inserted into the image of the scene and the system was tested first on a (then applicable) PDA platform. Translations of text found in road sign images were then displayed as subtitles under the signage. Meanwhile, similar applications have been developed and issued as iTunesTM and AndroidTM apps for smartphones. A more recent development offered by Wordlense, a startup company (now incorporated in Google Translate), combined a simple recognition engine for Western script and translation dictionaries with graphical rendering that inserts the translated word back into the original image. Both applications (shown in the left and right of Figure 14.15) demonstrate how an integrated multimodal design is equally essential to achieving our goal of language transparency as the language technology itself.



**Figure 14.16** Translation of lecture slides.

Handwriting recognizers recognize handwritten text and provide text in translation. The problem of text translation in real images has been solved partly solved by the road sign translator or OCR scanners as discussed above [Zhang et al. 2002], but using handwriting still offers additional real-time low latency opportunities. The difficulties of recognizing handwriting also require more sophisticated recognition akin to speech recognition. Early neural network-based and HMM-based systems have been proposed since [Jaeger et al. 2000, 2001, Manke et al. 1995, Starner et al. 1994], and recent solutions have matured in performance and usability,<sup>7</sup> so as to permit integration into sophisticated commercial grade multimodal communication interfaces.

- **Translation of lecture presentation material.** If foreign students have difficulties understanding lectures in a foreign language, then the lecturer's presentation slides or handouts might generate communication issues, too. For this reason, translation can also be applied to material across these different media, as well. Figure 14.16 shows a prototype translation system for PowerPoint™ slides (explored at KIT) that translates the text on a slide, when hovering with the mouse over the appropriate text. The translated text is then displayed in a speech bubble.
- **Distant speech input.** A continuing issue problem with speech translation devices is the placement of the microphone. When wearable or hand-held

7. <http://www.myscript.com>.

microphones are acceptable (e.g., lectures or mobile smartphones), this is of little concern, since the speaker's speech is well discernable and signal quality generally good. However, in meetings, noisy public places, and many other situations, signal separation, reverberation, and noise become significant factors in delivering successful translation services. One method to mitigate these factors are microphone arrays that are placed in a strategic location (e.g., on a table in a meeting room), worn (necklaces), or moved on a robot. Fujitsu and NICT are testing directional microphone technology under a national project aimed to deliver communication for the 2020 Olympics in Japan. Other mobile microphone arrays were proposed as attachments on smartphones or individual devices and (of course) for non-translation purposes conversational speech dialogue pods for the home such as Alexa and Google Home.

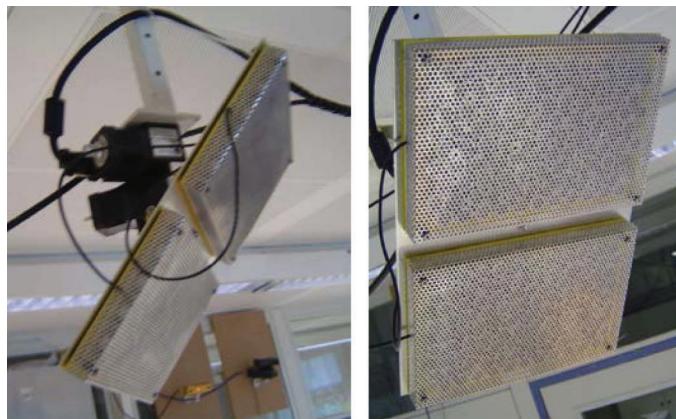
- **Silent speech input.** Speech is audible and thus perceived as noise for those for whom it was not intended. Is noise-free speech conceivable? Alternate non-vocal speech recognition systems have indeed been proposed, where articulated mouth movements are captured by electromyography, even though the language is not spoken out loud. Such “silent speech” can be recognized (although recognition is not as good as for spoken language), translated, and made audible by synthesis [Maier-Hein et al. 2005]. Subsequently, articulation of silent speech can be translated into audible speech in another language. The underlying technology is not yet mature, but the proposed prototypes (see Figure 14.17) show that input devices could be devised that can accept *silent* spoken language as an alternative modality where speaking aloud would create disturbances (or privacy concerns). Using silent speech input technology might thus be conceivable so that *anyone* can produce loud speech in *any* language, by (silent) articulatory motion in another.
- **Targeted Audio.** Synthetic speech output in a speech translation system can also be delivered selectively by directional loudspeakers. Such speakers were proposed experimentally (CHIL-project, [Waibel and Stiefelhagen 2009]), and commercially distributed (Sennheiser AudioBeam Ultrasonic Directional Loudspeaker).<sup>8</sup> By directing such speakers to different points in a room different listeners can then listen to simultaneous translation in different languages without a headset. It is as if each listener has a personal interpreter whispering the interpreted result into his or her ear. Early steer-

---

8. Sennheiser Electronic GmbH & Co. KG, product currently discontinued.



**Figure 14.17** Silent speech as input to translator.



**Figure 14.18** Translated speech delivered through targeted audio speakers: Personal, audible interpretation without headsets.

able prototypes have been developed and proposed [Waibel and Stiefelhagen 2009] for meeting rooms where the appropriate output interpretation can be positioned toward specific individuals (or guided by face recognition) (see Figure 14.18).

- **Speech translation goggles.** As alternative to personalized acoustic delivery of translation output, such output can also be delivered in visual form. In 2005, such a cross-modal speech translated was first proposed at a press



**Figure 14.19** First demonstration of heads-up display “translation goggles” at Carnegie Mellon (2005).

conference at CMU/KA where translated output from a simultaneous (lecture) translation system was delivered textually via heads-up display goggles (Figure 14.19). In this configuration, the user faces a conversation partner or a lecturer and the translation of spoken words is displayed as text as subtitles in the glasses. While this still seemed like science fiction in 2005, such configurations are now becoming reality as mobile computing platforms (smartphones, smart watches) can be connected with wearable augmented and virtual reality eye-glasses (Google “Glass”, Snapchat’s “Spectacles”, Facebook’s Oculus) that are also becoming pervasive and commonplace. Google already proposed a speech translator just like it as a feature for Google Glass (Figure 14.20), Google Glass Demo<sup>9</sup> and others are sure to follow.

- **Earplugs and Pixel-Buds.** Another form factor that has recently attracted attention are earphone style devices (perhaps inspired by the Babel Fish from the science fiction series Hitchhiker’s Guide to the Galaxy). Here, a set of earplugs provides input and output for a speaker attempting to dialogue with others. The underlying technology is similar to the systems described above, but speech is delivered through earbuds instead of to a phone’s microphone. A young start-up, “Waverly Labs,” announced to bring a product (the “Pilot”) to market, and Google recently launched a similar product called “Google Pixel Buds”. Google’s system combines Google Translate with speech I/O from and to the earbuds.

9. <https://www.youtube.com/watch?v=MqZuscmCYi4>.



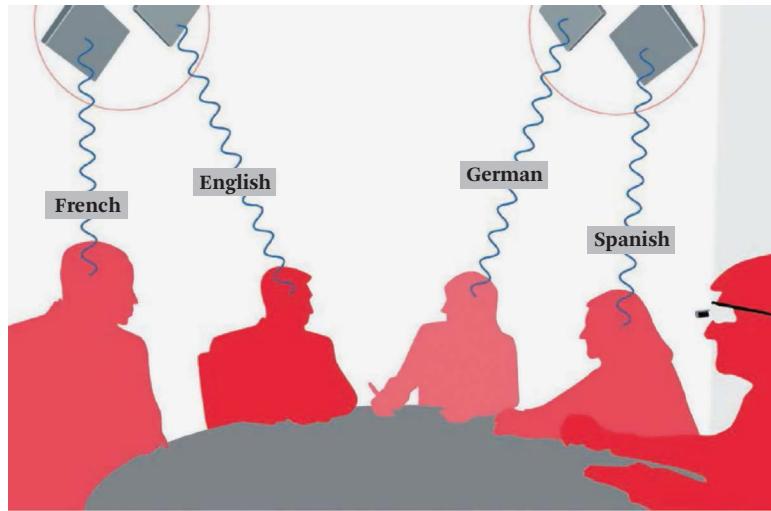
**Figure 14.20** MITE: translation via Google glass (2014).

Given all the advances with systems that can provide a translation function in a variety of situations, speaking styles and across multiple modalities, true integrated multilingual and multimodal environments become possible. With input accepted from personal, directional microphones, by electromyography speech or handwriting, and output translations delivered via directional targeted audio speakers, heads-up display goggles, personal displays on smartphones and tablet, language transparent conversations, and meetings become possible. So far, such fully integrated systems have been demonstrated only as prototypes or concept demonstrations (see Figure 14.21), but with continued progress they will likely transform the way we communicate in the global village of the future.

## 14.5

### Conclusion

Multimodal interfaces represent a critical dimension to building effective systems that support cross-lingual communication. Depending on the situation (lectures, meetings, one-on-one conversations, mobile dialogues, telephone conversations, blackboard notes, handwriting, texting, and many more), language is communicated in different modalities and at different speeds. Input must be accepted in different forms and output translation delivered in different modes and presentation styles, depending on use-case and personal preference. Perceptual input processing and translation technology has to be carefully adapted optimized to deliver the best language translation accuracy, at the appropriate speed, and latency, as required by the application. All processing steps also have to sensitive to the conversational context. With multimodal (“fleximodal”) interfaces, dialogues across language barriers can be supported effectively. Multimodal interfaces can better



**Figure 14.21** Individually adapted simultaneous translation in meetings.

compensate for errors, improve the speed of communication, adapt and scale, and respond to user communication needs and environments. Suitable interface design is as important to the success of cross-lingual communication tools in practice, as the performance of the underlying technology components.

### **Focus Questions**

**14.1.** What are the three partial tasks that need to be solved for speech-to-speech translation?

- ASR
- MT
- Speech Synthesis

**14.2.** What are the three main components of a modern speech recognition system?

- Acoustic Model
- Dictionary
- Language Model

**14.3.** What NN architecture is typically used in neural machine translation?

- Recurrent NN

**14.4. What are current research questions in speech translation?**

- Compound words
- Agreement
- Technical terms, jargon and unknown words
- Code switching
- Pronouns
- Readability
- Spontaneous Speech
- Microphones and noise
- Linguistic scalability/portability

**14.5. Which types of errors can occur during cross-lingual interpretation?**

- Out-of-Vocabulary Words
- Speech confusion
- Translation errors

## References

- D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 584, 586
- H. U. Block, St. Schachtl, and M. Gehrke. 2000. Adapting a large scale MT system for Spoken Language. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 394–410. Springer Published, Berlin/Heidelberg, Germany. 606
- H. Bourlard and N. Morgan. 1994. *Connectionist Speech Recognition, A Hybrid Approach*. Kluwer Academic Publishers.
- H. Bourlard and Ch. Wellekens. 1989. Speech pattern discrimination and multilayer perceptrons. In *Computer Speech and Language*, (3), pp. 1–19. 587
- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): S. 263–311. 590
- E. Cho, J. Niehues, T. L. Ha, M. Sperber, M. Median, and A. Waibel. 2016. Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016. In *Proceedings of the 13th International Workshop on Spoken Language Translation, IWSLT*. Seattle. 588
- E. Cho, J. Niehues, and A. Waibel. 2014a. Tight integration of speech disfluency removal into SMT. *EACL, 2014*, 43. Gothenburg, Sweden. DOI: [10.3115/v1/E14-4009](https://doi.org/10.3115/v1/E14-4009). 585

- E. Cho, J. Niehues, and A. Waibel. 2014b. Machine Translation of Multi-party Meetings: Segmentation and Disfluency Removal Strategies. *IWSLT*. Lake Tahoe, US. 601
- E. Cho, C. Fügen T. Herrmann, K. Kilgour, M. Medianı, C. Mohr, J. Niehues, K. Rottmann K. Saam, S. Stüber, and A. Waibel. 2013. A Real-World System for Simultaneous Translation of German Lectures. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. Lyon, France. 596, 598
- K. Cho, B. van Merriënboer, Ç, Gülc̄ehre, D. Bahdanau, F. Bougares, H. Schwenk. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1724–1734. Doha, Qatar. DOI: [10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179). 600
- M. Eck, I. Lane, Y. Zhang, and A. Waibel. 2010. Jibbigo: Speech-to-Speech translation on mobile devices. In *2010 IEEE Spoken Language Technology Workshop*, 165–166. DOI: [10.1109/SLT.2010.5700843](https://doi.org/10.1109/SLT.2010.5700843). 592
- The Economist. June 12, 2006. How to build a Bablefish. In *The Economist*.
- Ectaco. 1989. Ectaco eBook Readers and Translators. <http://www.ectaco.com>. 592
- C. Fügen, A. Waibel, and M. Kolss. 2007. Simultaneous Translation of Lectures and Speeches. In *Journal Machine Translation*, 21(4), 209–252. DOI: [10.1007/s10590-008-9047-0](https://doi.org/10.1007/s10590-008-9047-0). 597
- A. von Greve-Dierfeld. 2012. Uni-Übersetzungs-Automat: Don't worry about make. Spiegel Online. <http://www.spiegel.de/unispiegel/studium/dolmetscher-fuer-die-vorlesungskit-entwickelt-uebersetzungsprogramm-a-838409.html> 598
- T.-L. Ha, J. Niehues, and A. Waibel. December 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, 8–9. Seattle, WA. 586
- Handelsblatt: Übersetzung. July 1991. Handelsblatt. [http://isl.anthropomatik.kit.edu/cmu-kit/downloads/1991.07.30\\_Handelsblatt.pdf](http://isl.anthropomatik.kit.edu/cmu-kit/downloads/1991.07.30_Handelsblatt.pdf). 588
- K. Hatazaki, J. Noguohi, A. Okumura, K. Yoshida, and T. WatanabeT. 1992. INTERTALKER: an experimental automatic interpretation system using conceptual representation. In *Second International Conference on Spoken Language Processing*. 588
- S. Hourin, J. Binder, D. Yeager, P. Gamerdinger, K. Wilson, and K. Torres-Smith. 2013. Speech-to-Speech Translation Tool Limited Utility Assessment Report. Report OMB No.0704-0188. 593
- S. Jaeger, S. Manke, J. Reichert, and A. Waibel. 2001. Online handwriting recognition: the NPen++ recognizer. *International Journal on Document Analysis and Recognition*, 3(3): 169–180. DOI: [10.1007/PL00013559](https://doi.org/10.1007/PL00013559). 609
- S. Jaeger, S. Manke, and A. Waibel. September 2000. NPen++: An On-line Handwriting Recognition System. In *Proceedings of the 7th International Workshop on Frontiers in*

- Handwriting Recognition, IWFHR 2000*, Amsterdam, The Netherlands, 11–13. DOI: [10.1.1.30.158.609](https://doi.org/10.1.1.30.158.609)
- A. Jain and A. Waibel. August 1989. A Connectionist Parser Aimed at Spoken Language. In *Proceedings of the 1st International Workshop on Parsing Technologies, IWPT 1989*, Pittsburgh, PA. 28–31. DOI: [10.1109/ICASSP.1990.115782.585](https://doi.org/10.1109/ICASSP.1990.115782.585)
- M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2017. Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351. [586](#)
- E. C. Kaiser. 2006. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the 8th ACM International Conference on Multimodal Interfaces*, pp. 347–356. ACM Press. DOI: [10.1145/1180995.1181060.607](https://doi.org/10.1145/1180995.1181060.607)
- E. C. Kaiser. 2005. Multimodal new vocabulary recognition through speech and handwriting in a white-board scheduling application. In *ACM Intelligent User Interfaces Conference, IUI '05*, pp. 51–58. ACM Press, New York. DOI: [10.1145/1040830.1040851.607](https://doi.org/10.1145/1040830.1040851.607)
- N. Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1700–1709. [584](#)
- P. Koehn, Y. Zhang, C. Dugast, J. Gauthier, S. Grimsey, S. Fünfer, M. Mueller, S. Stueker, and V. Steinbiss. EU-BRIDGE D6.3 Final Evaluation Report, ([www.eu-bridge.eu](http://www.eu-bridge.eu)). [602](#)
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL, Prague, Czech Republic. [590](#)
- P. Koehn, and K. Knight. 2003. Empirical Methods for Compound Splitting. *EACL*, pp. 187–193. Budapest, Hungary. DOI: [10.3115/1067807.1067833.599](https://doi.org/10.3115/1067807.1067833.599)
- R. Kumar, S. Hewavitharana, N. Zinovieva, M. E. Roy, and E. Pattison-Gordon. 2015. Error-Tolerant Speech- to-Speech Translation. In *Proceedings of MT Summit XV, Volume 1: MT Researchers' Track*, pp. 229–239. Miami, FL. [606](#)
- A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan. 1997. JANUS III: Speech-to-Speech Translation in Multiple Languages. *International Conferences on Acoustics, Speech, and Signal Processing, ICASSP*. Munich, Germany. DOI: [10.1109/ICASSP.1997.599557.590](https://doi.org/10.1109/ICASSP.1997.599557.590)
- W. Lewis. November 2015. Skype Translator: Breaking Down Language and Hearing Barriers. *AsLing's 37th Translating and the Computer Conference*, 27–28. London. [595](#)
- L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. November 2005. Session Independent non-audible speech recognition using surface electromyography. In *Proceedings of ASRU*, Cancun, Mexico. DOI: [10.1109/ASRU.2005.1566521.610](https://doi.org/10.1109/ASRU.2005.1566521.610)

- S. Manke, M. Finke, and A. Waibel. 1995. The use of dynamic writing information in a connectionist on-line cursive handwriting recognition system. *Advances in Neural Information Processing Systems*, 1093–1100. [609](#)
- Y. Miao, M. Gowayyed, and F. Metze. 2015. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-Based Decoding. *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ. [588](#), [607](#)
- R. Miikkulainen and M. D. Dyer. 1989. A modular neural network architecture for sequential paraphrasing of script-based stories. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE. DOI: [10.1109/IJCNN.1989.118677](#). [585](#)
- T. Morimoto, Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu. 1993. ATR's speech translation system: ASURA. In *Proceedings Eurospeech '93*, pp. 1291–1294. Geneva, Italy. [588](#), [590](#)
- G. Neubig. 2016. Lexicons and Minimum Risk Training for Neural Machine Translation: NAIST-CMU at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*. [588](#)
- J. Niehues and A. Waibel. 2011. Using Wikipedia to Translate Domain-specific Terms in SMT. In *Proceedings of the Eight International Workshop on Spoken Language Translation (IWSLT)*. [600](#)
- F. J. Och, and H. Ney. 2004. The alignment template approach to statistical machine translation. In *Journal Computational Linguistics*, 30(4): pp. 417–449. DOI: [10.1162/0891201042544884](#). [590](#)
- S. Oviatt and R. VanGent. 1996. Error resolution during multimodal human-computer interaction. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Vol. 1, pp. 204–207. DOI: [10.1109/ICSLP.1996.607077](#).
- S. Oviatt and P. R. Cohen. 1992. Spoken Language in interpreted telephone dialogues. In *Computer Speech and Language*, 6(3): pp. 277–302. DOI: [10.1016/0885-2308\(92\)90021-U](#). [580](#)
- D. B. Roe. 1992. A spoken language translator for restricted-domain context-free languages. In *Speech Communication*, 11(2–3): pp. 311–319. DOI: [10.1016/0167-6393\(92\)90025-3](#). [588](#)
- S. Stüker, M. Federico, Ph., Koehn, H. Ney, M. Rödder, M. Simpson, V. Steinbiss, and A. Tescari. 2015. EU-BRIDGE Final Report. [www.eu-bridge.eu](http://www.eu-bridge.eu) [603](#)
- S. Stüker, C. Zong, J. Reichert W. Cao, M. Kolss, G. Xie, K. Peterson, P. Ding, V. Arranz, J. Yu and A. Waibel. 2006. *Speech-to-Speech Translation Services for the Olympic Games 2008, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, MLMI 2006*, Washington D.C. [592](#)
- M. Seligman, A. Waibel, and A. Joscelyne. 2017. TAUS Speech-to-Speech Translation Technology Report. *TAUS Report*.

- R. Sennrich, B. Haddow, and A. Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*. Berlin, Germany. [588](#)
- T. Starner, J. Makhoul, R. Schwartz, and G. Chou. 1994. On-line cursive handwriting recognition using speech recognition methods. *Acoustics, Speech, and Signal Processing, ICASSP*. DOI: [10.1109/ICASSP.1994.389432](#). [609](#)
- B. Suhm, B. Myers, and A. Waibel. May 1999. Model-based And Empirical Evaluation Of Multimodal Interactive Error Correction. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI 1999*, Pittsburgh, PA. DOI: [10.1145/302979.303165](#). [606](#)
- B. Suhm, B. Myers, and A. Waibel. 1996. Interactive recovery from speech recognition errors in speech user interfaces. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 861–864. Philadelphia, PA. DOI: [10.1109/ICSLP.1996.607738](#). [606](#)
- I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014* pp. 3104–3112. Quebec, Canada. [584](#), [585](#)
- T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto. 1998. A Japanese-to-English Speech Translation System: ATR-MATRIX. In *Proceedings ICSLP'98*, pp. 779–782. Sydney, Australia. [590](#)
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, l. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762. [586](#)
- Voxtec, Advancing Voice Technology™. <http://www.voxtec.com>. [592](#)
- W. Wahlster. 1993. Verbmobil—Translation of Face-To-Face Dialogues. In *Grundlagen und Anwendungen der Künstlichen Intelligenz*. Springer-Verlag, Berlin Heidelberg. DOI: [10.1.1.109.6407](#). [588](#)
- A. Waibel. 2002. Portable object identification and translation system. US Patent App. 10/090,559. [608](#)
- A. Waibel. 2015a. Translation Training with Cross-Lingual Multimedia Support, 2018, US Patent 9,892,115 B2; CIP Application #14/589,658, filed 2015.
- A. Waibel. 2015b. Translation and Integration of Presentation Materials with Cross-Lingual Multimedia Support, 2017 US Patent 9,678,953 B2, 2017; CIP Application #14/589,653, filed 2015
- A. Waibel. 2014. Translation and Integration of Presentation Materials with Cross-Lingual Multimedia Support, 2014; Patent Pub. #US 2014/0365202 A1; Appl. # 14/302,146, filed 2014
- A. Waibel and I. R. Lane. 2015. Enhanced speech-to-speech translation system and methods for adding a new word, US Patent. [607](#)

- A. Waibel and R. Stiefelhagen, editors. 2009. *Computers in the Human Interaction Loop*. Springer, London. [610](#), [611](#)
- A. Waibel and A. McNair. 1998b. Locating and correcting erroneously recognized portions of utterances by rescoring based on two n-best lists, US Patent 5,712,957b. [606](#)
- A. Waibel, B. Suhm, and A. McNair. 1998a. Method and apparatus for correcting and repairing machine-transcribed input using independent or cross-modal secondary input. US Patent 5,855,000. [606](#)
- A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann, and J. Tebelskis. May 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto. DOI: [10.1109/ICASSP.1991.150456](#). [588](#), [606](#)
- A. Waibel, A. Hanazawa, G. Hinton, K. Shikano, and K. Lang. March 1989. Phoneme Recognition Using Time-Delay Neural Networks. In *IEEE Transactions of the Acoustics, Speech and Signals Processing Society*, 347(3). DOI: [10.1109/29.21701](#). [587](#)
- A. Waibel. March 1989. Modular Construction of Time-Delay Neural Networks for Speech Recognition. In *Journal for Neural Computation*, MIT Press Journals, 1. DOI: [10.1162/neco.1989.1.1.39](#). [587](#)
- Y.-Y. Wang and A. Waibel. July 1997. Decoding Algorithm In Statistical Machine Translation. In *Proceedings of the 35th Annual Meeting of the ACL joint with the 8th Meeting of the European Chapter of the ACL 1997*, ACL/EACL 1997, Madrid, Spain. DOI: [10.3115/979617.979664](#). [590](#)
- Y.-Y. Wang and A. WaibelA. May 1991. A Connectionist Model for Dialog Processing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1991, Toronto, Canada. DOI: [10.1109/ICASSP.1991.150090](#). [585](#)
- J. Yang, J. Gao, Y. Zhang, X. Chen, and A. Waibel. 2001. An Automatic Sign Recognition and Translation System. *Workshop on Perceptual User Interfaces 2001*, PUI 2001. DOI: [10.1145/971478.971490](#). [608](#)
- T. Zenkel, R. Sanabria, F. Metze, J. Niehues, M. Sperber, S. Stüker, and A. Waibel. 2017. Comparison of Decoding Strategies for CTC Acoustic Models. In *Proceedings of Interspeech*, pp. 513–517. Stockholm, Sweden. DOI: [10.21437/Interspeech.2017-1683](#). [588](#), [607](#)
- Y. Zhang, B. Zhao, J. Yang, and A. Waibel. September 2002. Automatic SIGN Translation, 7th International Conference on Spoken Language Processing, 2nd Interspeech Event, ICSLP 2002 - Interspeech 2002, Denver, CO. [609](#)
- G. Zweig, C. Yu, K. Doppo, and A. Stolcke. March 2016. Advances in All-Neural Speech Recognition. *The 41st IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China. [588](#), [607](#)

# Commercialization of Multimodal Systems

Philip R. Cohen, Raj Tumuluri

## 15.1

### Introduction

This chapter surveys the broad and accelerating commercial activity in building products incorporating multimodal-multisensor interfaces. We discuss here multimodal-multisensor interfaces for: Aeronautics, robotics, biometrics, avatars and assistants, insurance, field force automation, and personal care. Volume 3 of this handbook also covers in separate chapters two topics for which there exist commercial products, namely multimodal-multisensor interfaces for automotive interfaces (Chapter 12) and for translation (Chapter 14). Numerous other applications are in development, but to be included in this chapter, the products must have been shipped or deployed, rather than being in a trial phase or under research. However, where directly relevant to product development, we will list emerging products and technologies. Finally, we are only considering products that offer multimodal interfaces, rather than providing multimodal data primarily just for the user's visualization, without the user's interacting with the system. Thus, we do not consider multimodal radiological imaging, combining PET, CT, MR and other types of radiographic scans, which has become a major trend in radiology. For such systems, the multimodal imaging system displays and overlays different modalities for viewing by the radiologist. Excellent surveys of the literature on multimodal imaging for various medical specialties can be found in [Uludağ and Roebroeck \[2014\]](#), [Vogler et al. \[2015\]](#), [Wu and Shu \[2018\]](#).

## 15.2

### Aeronautics

Lockheed Martin's F-35 Joint Strike Fighter is among the pre-eminent examples of information fusion in modern technology. In virtue of multi-spectral sensing,



**Figure 15.1** The F-35 helmet is an information-display device, showing targeting data, status of the aircraft systems, and visual and infrared views of the world outside the airplane. (USMC employee Public Domain)

information fusion across aircraft and ground personnel, and increased levels of automation, this high-performance aircraft can be operated by a single pilot equipped with a multimodal-multisensor interface.

The locus of multimodal-multisensor interface technology is the F-35 pilot's helmet built by a joint venture of Rockwell Collins and Elbit Systems. This \$400,000 helmet is actually an integral part of a multi-pilot information system, which fuses situational information from multiple cameras, multiple aircraft into a 360° 3D augmented reality image that is projected onto the pilot's monochromatic 30 × 40 degree field-of-view helmet-mounted display (see Figure 15.1). The pilot can see a 360° sphere around the plane, and as his head moves, the entities displayed remain fixed in their positions in 3-space. For example, the pilot can "look" through the floor of the aircraft to see the ground below. To do this, the helmet's multi-sensor fusion tracks the pilot's head optically, as well as magnetically via a magnetic field generated under the seat, and via an inertial measurement unit. The helmet also tracks the pilot's eye position, with eye-tracking aligned to the pilot's inter-pupil

distance. In order to generate the required level of precision, the helmet is carefully milled to the pilot's head shape, requiring four fittings to properly align the optical components [Mola 2017, Lemons et al. 2018].

Among the sensors that are fused are visible light and infrared images, as well as a night-vision system positioned on the helmet (see Figure 15.1) that is fused with the optical image generated from a camera on the cockpit glare shield. Because of the many sensors and novel display technologies, significant effort was expended to *declutter* the display, providing a single fused picture of the airspace. For example, because data is shared across multiple networked aircraft, the information system must fuse the icons representing real entities (called *tracks*) as observed by each aircraft, rather than duplicate them. The fusion algorithm employed is based on *Dempster-Shafer theory* [Frey et al. 2018].

A major technical advance in the F-35 is Autonomous Sensor Management, a *closed-loop fusion* system that autonomously modifies “the actions of the sensor suite to achieve a mission-level objective or behavior,” such as track accuracy [Frey et al. 2018]. Based on a concept of *sufficiency* (i.e., having sufficient information for the existing mission), the sensor manager can decide only to gather sufficient new information about existing tracks or to scan for other entities. This autonomous sensor manager enables the pilot to offload sensing decisions, enabling him/her to concentrate on flying the aircraft.

The helmet also contains an optical sensor, built by Lifebeam Systems, that monitors the pilot's state of consciousness by measuring heart rate and blood flow. If it senses that the pilot is unconscious due to high G-force or hypoxia at high altitudes, it notifies the plane's autopilot system to take over the flight until consciousness is regained. Regarding interacting with the cockpit itself, in addition to commands that can be entered by the pilot via a touchscreen, or buttons on a joystick, commands to control various airplane systems can also be issued via a noise-robust voice recognition system based on SRI International's Dynaspeak [Franco et al. 2002], integrated by Adacel Corporation. Thus, alternative use of interface modalities is supported, although given the eye-gaze system in the helmet, it may be possible to integrate selection via gaze with voice query about the selected entity.

The importance of aeronautical sensor fusion cannot be overstated. Recent crashes of the new Boeing 737 Max 8 aircraft are being attributed to the autopilot's stall-prevention system's being reliant on a single angle-of-attack sensor rather than multiple fused sensors [Tangel and Pasztor 2019]. Boeing was in the midst of modifying its software to incorporate multisensor fusion when a second crash happened on March 12, 2019 also suspected to have been caused by the

### Glossary

**Biometric Liveness measures.** In order to avoid spoofing, a biometric system can sense whether the data it is receiving is in fact coming from a live person (vs. a cadaver, or a photograph). Example measures include heart rate, heat, etc.

**Biometric spoofing.** The provision of replicas of a person's biometric data, such as a photograph or plaster cast, in order to cause a biometric algorithm to falsely recognize the presence of that person.

**Closed-loop fusion.** Based on mission/pilot objectives, an autonomous sensor manager will analyse fused track information in order to prioritize sensor tasks, including maintaining the sensing of current tracks vs. searching for new ones. Thus, sensor fusion leads to creation and prioritization of new sensor fusion tasks.

**Decluttering of a situational display.** Using multisensory fusion techniques to simplify a display by combining tracks that possibly represent single entities. Decluttering has to accommodate multisensory information from the onboard sensor suite, as well similar information being propagated from other sensing platforms (aircraft, ground stations, etc.).

**Dempster-Shafer Theory (DST)** is a mathematical theory of evidence that is often used for sensor fusion [Calderwood et al. 2017, Murphy 1998] because it enables one to combine evidential information from different independent sources without requiring prior probabilities. The theory is based on belief functions (in DST, called "mass functions") that assign a degree of belief (mass) to sets of propositions. DST starts with a "frame of discernment"  $\Omega$ , a set of mutually exclusive and exhaustive possible states of the system being studied. For example,  $\Omega$  might consist of the set {airliner, military jet, helicopter} meaning that the system is sensing one of those aircraft. A belief function  $m$  will assign evidential belief mass (perhaps zero) to the elements of the powerset of  $\Omega$ . To continue the example, the mass function may assign evidential mass to the subset {airliner, helicopter}, meaning that it is assigning evidence to its sensing either an airliner or a helicopter. Evidence from multiple independent mass functions  $m_1, m_2$  (e.g., multiple sensors) may be combined using Dempster's rule, which sums the product of those mass functions  $m_1(X)m_2(Y)$  evaluated over the non-null intersection of all sets  $X$  and  $Y$  from  $\Omega$ . Dempster's rule ignores conflicting evidence between belief functions, i.e., where  $X$  and  $Y$  do not overlap, through a normalization factor that has led some to argue that this rule can provide counterintuitive results [Zadeh 1986]. Numerous attempts have been made to overcome these problems with different rules of combination while maintaining the benefits of the approach [Haenni 2004, Sentz and Ferson 2002].

**Glossary** (*continued*)

**Kalman Filter** A Kalman filter [Kalman 1960] is an iterative estimation algorithm used to predict future states of a continuously varying system or process based on potentially noisy measurements of the current state. Because it naturally incorporates multiple sensors, it is often used as a sensor fusion algorithm. The standard Kalman filter predicts the system and minimizes the noise (hence the term “filter”) optimally if the measurement and noise are Gaussian and the sensors are linear. Variants of the Kalman Filter, such as Extended Kalman Filters [Pfeiffer and Franke 2010] and Unscented Kalman Filters [Allodi et al. 2016, Julier and Uhlmann 2004], are used when the system being modeled behaves nonlinearly. Models of the covariance of multiple sensors are used to accomplish sensor fusion.

**Myoelectricity.** The electrical signals that stimulate muscles.

**Sufficiency of Sensing.** Rather than fusing multiple sensors to determine the most precise estimate of a track’s properties (e.g., range, bearing, identity, etc.), modern aeronautical information fusion algorithms may attempt to derive an estimate sufficient only for the pilot’s understanding or mission success. Once that is accomplished, sensing resources can be redeployed to analyse other tracks.

**Track.** In aeronautical parlance, an entity that is being sensed (e.g., by radar).

**Transcode.** The transformation of one encoding of an entity into another.

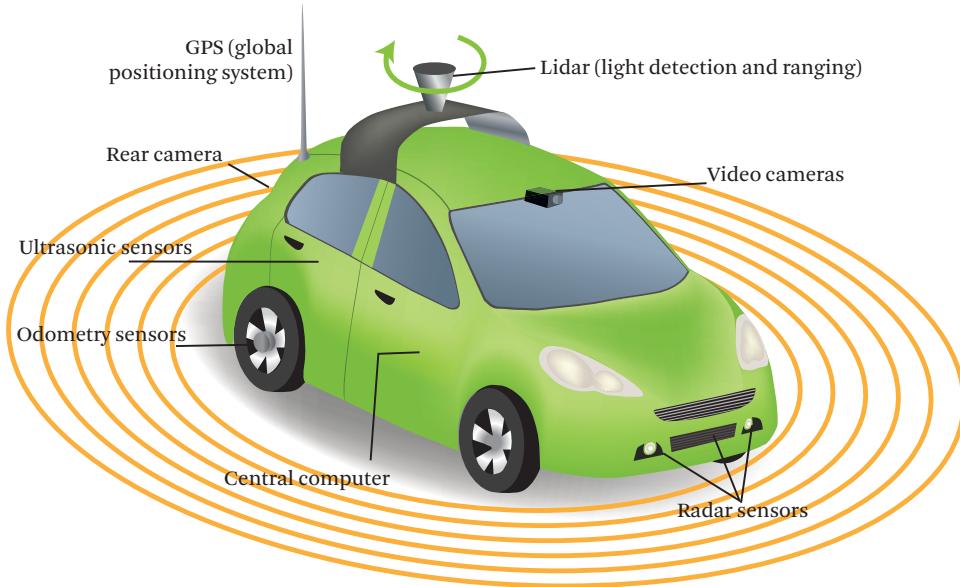
single sensor-based stall-prevention system. On March 13, 2019, the airplane was grounded worldwide until new multisensor fusion software and operating instructions have been provided.

## 15.3 **Robotics**

In this section, we examine multimodal/multisensor interaction and fusion with robots, including autonomous cars, surgical robots, prosthetic arms, and social robots.

### 15.3.1 **Autonomous Cars**

A critical factor enabling the development of autonomous cars is sensor fusion. The automobile needs to sense its surroundings, including its own position and movement, identify and track stationary and moving objects, as well as markings on the roadway. These perceptions need to be extremely accurate in all kinds of weather and road conditions. In order to do so, such vehicles are equipped with multiple sensors, including: visible light cameras, Lidar (Light Detection and



**Figure 15.2** Sensors on a self-driving car.

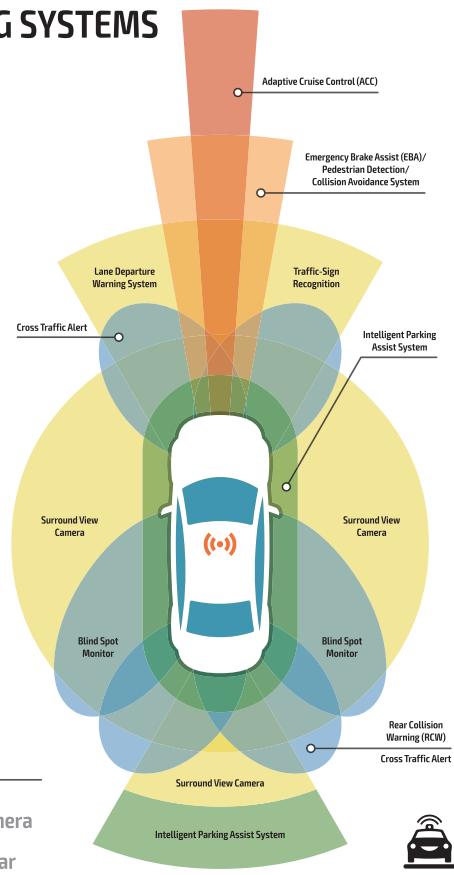
Ranging, using infrared lasers), Radar (Radio Detection and Ranging), odometric (wheel angle and displacement), ultrasound (nearby obstacle detection), cameras (object identification, distance calculation), and others (see Figure 15.2). Lidar can provide accurate position information, but can be obscured in bad weather. Radar can provide data on velocity, distance, and angular location of an object, but is not suitable for object identification. 2D and 3D visible light cameras can enable object detection and recognition, but can be impeded by bad weather. The sensor fusion task then is to combine these potentially noisy sensors in order to estimate the vehicle's location, as well as any objects in its vicinity, and to perform many other tasks (see Figure 15.3).

Schoettle [2017] argues that sensor fusion is critical to safe autonomous driving. Typical sensor fusion and tracking algorithms include Kalman filtering [Choi et al. 2013], extended Kalman filtering [Pfeiffer and Franke 2010], unscented Kalman filtering [Allodi et al. 2016, Julier and Uhlmann 2004], particle filtering [Song et al. 2015], evidential reasoning algorithms [Chavez-Garcia and Aycard 2016], and recently, deep neural nets [Liang et al. 2018], etc. The technology is advanced enough that chip manufacturers such as Intel and Nvidia are delivering fully packaged multisensor fusion chips ready for integration into autonomous vehicles. There is a huge research literature devoted to this topic, and a good overview can be found in

## AUTONOMOUS CAR SENSING SYSTEMS

-  Adaptive Cruise Control (ACC)
-  Emergency Brake Assist (EBA)
-  Pedestrian Detection
-  Collision Avoidance System
-  Lane Departure Warning (LDW)
-  Traffic-Sign Recognition
-  Surround View Camera
-  Cross Traffic Alert
-  Blind Spot Monitor
-  Rear Collision Warning (RCW)
-  Intelligent Parking Assist System

● Long-Range Radar    ● LIDAR    ● Camera  
● Medium-Range Radar    ● Ultrasonic Radar



**Figure 15.3** Uses of sensors.

[Elfring et al. 2016, Schoettle 2017]. Some vehicles, notably the Tesla Model S, do not include lidar among their sensors, relying primarily on computer vision from multiple cameras and radar, with image recognition processed by deep neural networks. Still, the (semi) autonomous driving system appears not to be able to detect stopped vehicles when driving at high speed, which has arguably led to a number of crashes [Stewart 2019]. In addition to lidar, the sensor suite could include far-field infrared used in night-vision systems to detect humans and animals [Schoettle 2017], though addition of sensors puts a corresponding load on the computer processors. Fosse [2019] points out that even if an autonomous car drove well, it would drive differently from humans, potentially causing humans to behave differently, leading to accidents or driver frustration.

Apart from uses in autonomous vehicles, modern automobiles are laden with sensors (see Figure 15.3) that provide a dramatically safer human driving experience, including blindspot detection, lane departure, collision avoidance, automatic braking, etc. Multimodal interaction with vehicle systems, allowing for alternative and simultaneous use of modalities such as speech input and output, touch input, and graphical output either on a touch screen or on a heads-up display, currently exists in the latest cars offered by major automobile manufacturers. Chapter 13 of this volume provides much more detail on this topic.

### 15.3.2 Surgical Robots

FDA-approved robotic surgery products first appeared in the year 2000 with the introduction of the da Vinci™ surgical robot by Intuitive Surgical. As of 2017, Intuitive Surgical reports that more than 4100 da Vinci robots have been delivered, which have performed 450,000 surgeries. The tele-operated robot consists of four robotic arms equipped with surgical tools on the ends that are controlled by a surgeon at a distance from the patient (see Figure 15.4). The surgeon views the operating field via a 3D stereo image, and manipulates local hand controls and foot pedals (the “master” tools), whose movements are translated electronically to the operating tools (the “slave” tools). Often, the surgeon is positioned in the same room, but in principle, the surgeon could be out of the room or on another continent, if there is sufficient communications bandwidth.

The stated benefits to the patient of tele-operated robotic surgery include: minimal invasiveness, filtering of surgeon tremor, less pain and scarring, less blood loss, less risk of infection, and faster post-operative recovery. Some of these benefits may be due to the surgeon’s being able to see the surgical field in 3D, with minimal blood pooling. Current research is building robots that can perform more precise movements than would be possible with the human hand (<https://newatlas.com/oxford-eye-surgery-robot/45382/>).

However, outcomes are still in debate [Barry et al. 2012]. A major criticism of the da Vinci robot is the lack of direct haptic feedback, which leads to increased errors [Meccariello et al. 2016, Ruurda et al. 2004, Weber and Schneider 2014, Wottawa et al. 2016], increased surgeon cognitive load resulting in difficulty to learn to operate with the device [Lendvay et al. 2013, Van der Meijden and Schijven 2009], and increased surgical time [Van der Meijden and Schijven 2009, Pacchierotti et al. 2015, Weber and Schneider 2014], which can indirectly lead to numerous potential complications [Maerz et al. 2017]. There is recent evidence that minimally invasive hysterectomies for early stage ovarian cancer, including robotic surgeries, had a



**Figure 15.4** The da Vinci robotic surgery system. (Photo © 2015 Intuitive Surgical, Inc.)

significantly lower survival rate at 4.5 years after the operation than open surgeries [Melamed et al. 2018].

Researchers are actively studying how to provide feedback to the surgeon [Gulrez 2014, Okamura 2009, Wang et al. 2017], with different output modalities being investigated. In addition to haptic force feedback, one promising approach is to *transcode* the force information at the end effector into another modality or modalities to be presented to the surgeon, such as visual and/or auditory modalities [Aviles-Rivero et al. 2018, Bethea et al. 2004]. Some new products have recently entered the market that provide force feedback. For example, HeroSurg (<https://newatlas.com/robotic-surgery-herosurg-haptic-feedback/45676/>) uses strain gauge technology incorporated into the end instrument to measure forces. Another new product provides multimodal feedback (vibratory and audio) (Verrrotouch Medical). An excellent survey of the provision of haptic feedback to the robotic surgeon can be found in Enayati et al. [2016].

### 15.3.3 Prosthetic Arms

The human hand has a very wide range of flexibility and dexterity that has been hard to mimic with prostheses. However, recently a number of companies have



**Figure 15.5** Mobius Bionics' LUKE™ arm. (Courtesy of Mobius Bionics LLC)

built prosthetic arms that offer an unprecedented range of motion, control, and feeling. New limbs are lighter weight than previously (some less than 2 kg), offer more motors driving more joints, and are able to sense and provide feedback. Modern prosthetic limbs are typically controlled by **myoelectricity**, meaning that sensors on the remaining portions of the human limb sense muscle movements. Those signals are converted to movements of the motors that control the shoulder, elbow, wrist, and individual fingers. Another way to control such arms, employed by the LUKE™ arm (see Figure 15.5) from Mobius Bionics (<http://mobiusbionics.com>) is through inertial measurement units placed in the user's shoes that convert foot motion to hand/arm motions. These new bionic hands can perform many different grips. For example, the Bebionic hand (<http://www.bebionic.com>) can perform 14 grips, which require a touch on the back of the hand with the user's other hand to invoke. Touch Bionics' hand (<http://www.touchbionics.com>) enables users to change grips with gesture control or a mobile app. Similarly, OpenBionics (<http://www.openbionics.com>) offers a 3D printed, and thus low-cost, myoelectrically controlled bionic arm, whose grips can be changed by a tensing/holding signal, causing the hand to cycle through groups of grip modes. Importantly, the LUKE and OpenBionic arms provide vibratory haptic feedback at the junction of the prosthetic and the partial limb in response to the force being exerted by the hand/fingers to which it is attached (<https://www.youtube.com/watch?v=6rloSSqiUCM>).

An emerging technology that promises to revolutionize prosthetics even more is the combination of brain-computer interfaces and prosthetics, such that a person can imagine performing a hand/arm or lower limb motion, which transmits signals to the nerves in the partial limb that are then translated to computer signals to control the prosthesis. Signals can flow the other way too, as force feedback applied by the bionic arm is provided to the user's brain via the electrodes attached

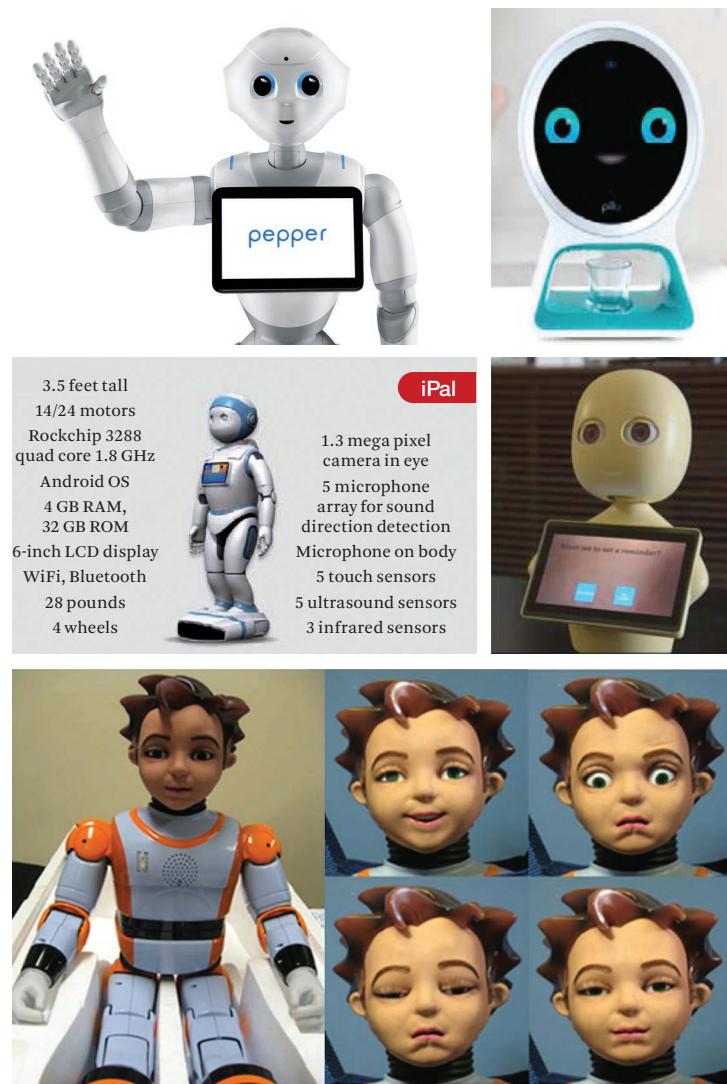
to the partial limb. This video from Johns Hopkins University's Applied Physics Laboratory shows what is now possible for prosthetic arms [https://www.youtube.com/watch?v=F\\_brnKz\\_2tI](https://www.youtube.com/watch?v=F_brnKz_2tI). Based on the rapid progress in this field, the US Food and Drug Administration has recently issued draft guidelines for clinical and non-clinical development and testing of brain-computer-driven prosthetics [US FDA 2019]. Finally a recent survey of brain-machine interfaces for prosthetics and neurorehabilitation can be found in [Lebedev and Nicolelis 2017], and Chapter 12 of this volume provides a thorough discussion of a related topic, brain-computer control of exoskeletons.

#### **15.3.4 Social Robots**

Social robots are not intended to perform specific physical tasks, such as a factory welding robot but rather are intended to engage humans in social interaction. In some cases, their task is to be a companion to elderly adults or children. Other robots (e.g., Softbank's Pepper™; <https://www.softbankrobotics.com/emea/en/pepper>) have been used to greet humans in stores, hotels, or other venues, or to attract crowds for marketing purposes.

All social robots are, of course, multisensor and multimodal, although the robots' manufacturers do not generally provide details of lower-level multisensor fusion algorithms, e.g., for navigation or object detection. Some are humanoid-shaped with heads, arms, faces, and either legs or wheels. They usually are equipped with stereo cameras, depth and touch sensors, sonar, and in some cases, infrared sensors. Usually, they provide voice and camera inputs, and for some a touchscreen. Figure 15.6 shows images of the robots discussed below, and Table 15.1 compares the various social robots relative to their multimodal-multisensor input and output capabilities.

Based on these hardware devices, the various robots support interactions with humans in support of their intended function(s). Starting with non-humanoid robots, Jibo (<http://www.jibo.com>) pioneered the concept of an emotionally attractive home assistant robot. Jibo incorporates multiple cameras, a display, speaker, and microphone array for sound source localization and voice capture. These multiple sensors enable Jibo to identify the user and to engage in spoken language question-answering dialogs, in support of informative, entertainment, and educational applications. A unique aspect of this robot is while it does not change physical positions, its segmented "body" rotates and bends, creating movements that emulate a stylized "bow," which was intended to be emotionally attractive to its home-based users. However, the body movements may not have had a significant



**Figure 15.6** Social robots. Top row left: Pepper (Courtesy of SoftBank. All rights reserved) and Pillo (Courtesy of Pillo, Inc.). Middle row, left to right: iPal (Courtesy of AvatarMind Robot Technology) and Mabu (Courtesy of Catalia Health Inc.). Bottom row: Robokind Zeno, showing some of its facial expressions. (From [Cameron et al. 2015], used with permission)

purpose, and cost valuable time to implement. Although potentially more attractive as a long-term companion, as a result of late market entry, its market niche has now largely been filled by non-robotic Amazon Alexa™ and Google Home™ devices, among others. With more limited functionality, Mabu™ from Catalia Healthcare is intended to engage patients in a long-term wellness relationship through conversation that is tailored to the patient's emotional state. Mabu reminds the user to take medications and can coordinate with both family members and healthcare personnel. Mabu has multiple microphones, cameras, and articulated eyes and head. Because it is essential that the robot creates a rapport with the elderly client in support of long-term usage, Mabu employs Affectiva's emotion recognition system (see Section 15.11), which enables it to adjust its behavior depending on the user's state. Whereas Mabu reminds users to take their medication, Pillo™ from PilloHealth is a pill dispensing robot that uses voice and face recognition to coordinate pill delivery to the right user. It can provide information about the medications, and also provide typical assistant information (e.g., weather reports) via voice input. If the user wishes, Pillo can coordinate video calls with family members.

Among the humanoid robots, SoftbankRobotics's Pepper™ is intended to be an emotionally attractive companion. With its tablet display, it can deliver corporate information as part of its conversational interaction. Pepper is now being delivered with the Google Android™ operating system, providing easy access to voice recognition in 15 languages, and tools for natural language processing. It also supports person and object recognition, as well as emotion recognition, enabling it (in theory) to tailor its output to the state of the user. Pepper is capable of omnidirectional and autonomous navigation. The PAL Robotics (<https://www.pal-robotics.com/en/>) service robots (Reem™, Reem-C™, Talos™, Tiago™) can recognize objects, people, and faces. AvatarMind's (<https://www.ipalrobot.com/>) iPal™ is a social robot designed for educating and entertaining children, elder companionship, and marketing. It incorporates multiple sensors (touch, ultrasound, and infrared), as well as multi-modal interaction via vision and speech recognition/natural language processing. The iPal robot can also function as a mobile tele-operated video camera so remote parties can engage visually and verbally with the local user (typically a child or elder). The robot is somewhat controversial as it is partly intended to educate and play with lonely children. Finally, the Furhat™ robot [Al Moubayad et al. 2012] consists of a translucent moveable head with a backprojected image of a face, which avoids having to construct articulators for facial movements. Furhat is being used for a number of applications, including public health screening, in partnership with Merck.

**Table 15.1** Multimodal-multisensor robots

		Company / Robot						
		ipal	Jibo	PAL	Softbank	Robokind	Intuitive Surgical	Verro Touch
Reem, REEM-C, Talos, Tiago								
Pepper								
Zeno R-50								
da Vinci								
Verro Touch								
<b>Active Input</b>								
Voice		X	X	X	X	X	X	
Touch on screen		X		X	X	X	X	
2D camera		X	X	X	X	X	X	
Pointing with mouse/stylus; pressing buttons						X		
Gesture					X			
Head pose				X				
Object recognition				X				
Person detection				X	X	X		
Face recognition		X		X	X	X		
Emotion recognition					X	X		
Sound-source localization		X	X		X	X		
Laser rangefinder					X			
Gaze								
<b>Passive Sensing</b>								
Microphones		X	X	X	4	8		
3D camera (stero)				X	X	X	X	X
Touch sensors		X	X		X	X	X	X
Torque sensors				X		X	X	X
Infrared sensors		X			X	X		
Depth sensors			X	X				
Ultrasound		X				X		
Lidar						X		
Vibration							X	
Sonar					X	X		
Inertial measurement units					X	X		

**Table 15.1** (*continued*)

		Company / Robot						
		ipal	Jibo	PAL	Softbank	Robokind	Intuitive Surgical	Verro Touch
		Reem, REEM-C, Talos, Tiago		Pepper, Noo		Zeno R-50	da Vinci	Verro Touch
<b>Output</b>								
Lights		X			X			
Haptic								X
Head motion			X	X	X	X		
Eyes/eyebrow motion							X	
Mouth motion							X	
2D or 3D screen graphics	X	X	X		X		X	X
Screen video	X	X				X		
Body motion		X			X	X	X	X
AR glasses						X		
Non-speech audio								X
Voice		X	X	X	X	X		
Gestures			X	X	X			

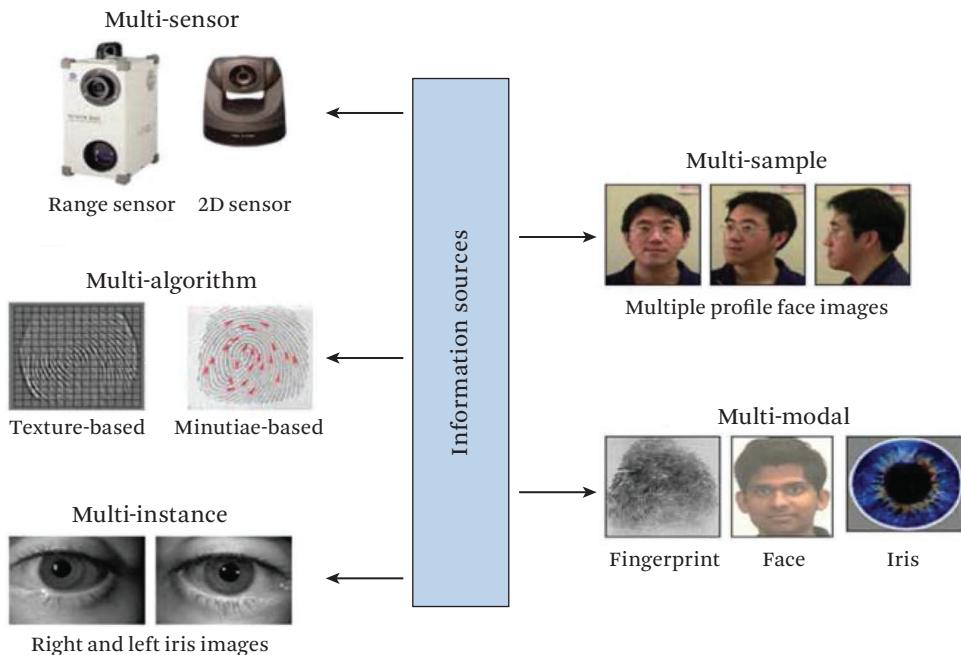
A significant potential use of socially *assistive* robots, especially emotive ones such as Robokind Zeno™, is to engage and teach children with autism spectrum disorders (ASD) [Dautenhahn et al. 2009]. The robot has an expressive face that can move its eyes, eyebrows, and mouth, can gesture with its arms, has a grasping hand, and the ability to move its legs and waist. For output, it also offers a text-to-speech system that is coordinated with face, mouth, and lip movement. Regarding use with children, it has been reported that some, but not all, children with ASD respond well to interaction with a robot [Diehl et al. 2012]. However, though the research is promising, well-controlled studies of the effectiveness of socially assistive robots for children with ASD is lacking [Dickstein-Fischer et al. 2018]. For healthy children, other studies have found that male children liked interacting with a (male-appearing) Zeno robot that showed positive and negative facial expressions better than they liked interacting with the same robot with a neutral expression. In contrast, no differences in likeability between expressive vs. neutral faces were found with female children and the male Zeno robot. An excellent survey of care robots can be found in [Bodenhagen et al. 2019].

Given the state of the art, most of these robots have limited conversational capabilities, but they are upgradeable through Software development kit (SDK)s. Although the goal is to relate to people, especially to their emotions, with the exception of Pepper's and Mabu's vision-based emotion recognition, most social robots do not provide built-in emotion recognition. Instead, they rely on researchers to develop that capability (see Volume 2, Chapter 6). In general, in order to make social robots more useful, there needs to be a complete multimodal architecture built on top of the host operating system (e.g., Robot Operating System or Android) that incorporates the various modalities with a path to upgrading.

## 15.4 Biometrics

The biometric research and development community has wholly embraced multimodality (see [Ross et al. 2006]). Typical use cases for biometric analysis include: physical access to secure areas, border control, fraud prevention, criminal, forensic, and civil applications such as voter identification, network entry (e.g., banking), cyber security, password reset, retail, etc. Depending on the level of security required, many organizations have opted for multimodal biometrics, which combine the results of multiple biometric measurements, either in sequence or simultaneously (e.g., speaking while looking at a camera). Ross and Poh [2009] describe the multi-biometric problem and distinguish the various types of data (see Figure 15.7) and approaches to combining results of multiple biometric classifiers.

Many commercial systems now incorporate multiple modalities such as: iris, fingerprint, finger vein, palm vein, face recognition, gait, and voice print, as well as behaviors such as typing on a keyboard. Figure 15.7 from Ross and Poh [2009] shows that one must distinguish between: collecting data via multiple sensors, applying multiple biometric recognition algorithms; taking multiple samples from multiple body parts (e.g., left and right iris); gathering multiple samples from the same person; and finally analyzing data from multiple modalities. However, here we will only consider the last as "multimodal biometrics." Multimodal biometric fusion technology can be used to identify a person from a huge database of people, or to verify that a given person is who s/he claims to be. Many of the same issues of time of fusion (early vs. late), and techniques for combining classifiers across different sensing modalities that are discussed in Alpaydin [2018], Kittler et al. [1998], Ross and Jain [2007], Ross et al. [2006], Volumes 1 and 2 of this handbook can be found in systems that perform multimodal biometric analyses. Techniques for combining classifiers that produce a score or probability, include applying a linear combination, taking the maximum, summing the results, etc. [Kittler et al. 1998, Sanchez



**Figure 15.7** Multiple sources of biometric data, including multimodal data. (From [Ross and Poh 2009], used with permission)

2006, Tulyakov et al. 2008, Fierrez et al. 2018]. When the data sources are independent, multibiometric solutions can have much lower false positive and false negative rates than either individual modality. Studies have shown very substantial error rate reductions in person identification by combining uncorrelated biometric data. For instance, Sanchez [2006] tested multimodal biometric classification of identity. They report that a linear combination of classifiers results in a false acceptance rate reduction from 12.67% using lip shape alone during speech to 0.29% using lip shape, two facial features, and two voice features. Likewise, false rejections were reduced from 14% using lip shape during speech to 0% using the five features above. Similar results have been found in bimodal person identification using voice and image from mobile phones [McCool et al. 2012].

A major reason for adopting multimodal biometrics for the highest security applications is to minimize vulnerability to *spoofing*. For example, fingerprint readers can be spoofed by artificial fingers [Matsumoto et al. 2002], and face recognition systems can be fooled by photographs. However, multimodal biometric systems would be much less likely to accept two such spoofs. In addition, biometric vendors

**Table 15.2** Multimodal biometric companies and the biometrics they use

Company	Iris	Fingerprint	Voice	Finger Vein	Palm	Face	Gait	Behavior
Anviz	X	X					X	
FST Biometrics				X			X	X
Daon IdentityX	X	X	X		X	X		X
HID Global	X	X			X	X		
IDTech360	X	X					X	
Incadence	X	X					X	
MorphoTrak	X	X		X	X	X		
M2Sys	X	X	X		X	X	X	
NEC Biometrics	X	X	X	X	X	X		
Nuance				X		X	X	
Rayabin Co	X	X					X	
Sensory				X			X	
TechnoBrain	X	X					X	

are now incorporating *liveness* measures, which can measure body temperature, eye-movement and blood flow. Other similar liveness measures are possible.

Companies that are offering multimodal biometric products include those in Table 15.2:

## 15.5 Assistants and Avatars

### 15.5.1 Virtual Assistants

Very few current virtual assistants are truly multimodal. The major virtual assistants supplied by Amazon, Apple, Facebook, Google, and Microsoft have so far supported voice-only or text-only interactions, with limited opportunities for follow-up interactions. Recently Amazon, Google, Microsoft, and Samsung have shipped their assistants on devices that incorporate screens, including the Amazon Echo Show<sup>TM</sup> (Figure 15.8), Samsung Bixby Home<sup>TM</sup> and the Smart Display<sup>TM</sup> for the Google Assistant, including devices from Google, Lenovo, Harman, and others. Microsoft's Cortana runs on phones and computers using the Windows Operating System, and Apple Computer's iPad<sup>TM</sup> tablet has been running Siri for many years.



**Figure 15.8** Amazon Echo Show. (From Daylen [CC BY 4.0])

These major vendors have realized that multimodal interaction will require that the interaction paradigm be rethought to be more tightly integrated across modalities, and have begun research efforts in their laboratories. Google is now providing developers with multimodal design guidelines for its DialogFlow developer toolkit, but has yet to provide a multimodal toolkit. However, developers can develop code that can sense the type of device the user has and provide system output accordingly. Furthermore, users can request that output be sent to their phone or Android tablet device. Amazon has added Alexa Presentation Language (APL), which enables third-party developers to build “multimodal” skills that coordinate Alexa’s natural-language understanding systems with on-screen graphics on Alexa devices, such as the Echo Show™ (see Figure 15.8). Recently, Apple has added follow-up query support via touch, such that if one selects an item from a list, such as a restaurant, it becomes the only item on the screen, and one can then make follow-up requests about it. Other virtual assistants can do this too, but it is a very simple use of multimodality.

Samsung’s Bixby™ Version 1 is designed such that whatever can be executed via touch on the graphical user interface can be accomplished via voice, and vice versa. Some Bixby applications are inherently multimodal. For example, Bixby Vision, part of the camera app on Samsung’s Galaxy and Note phones (versions 8 and

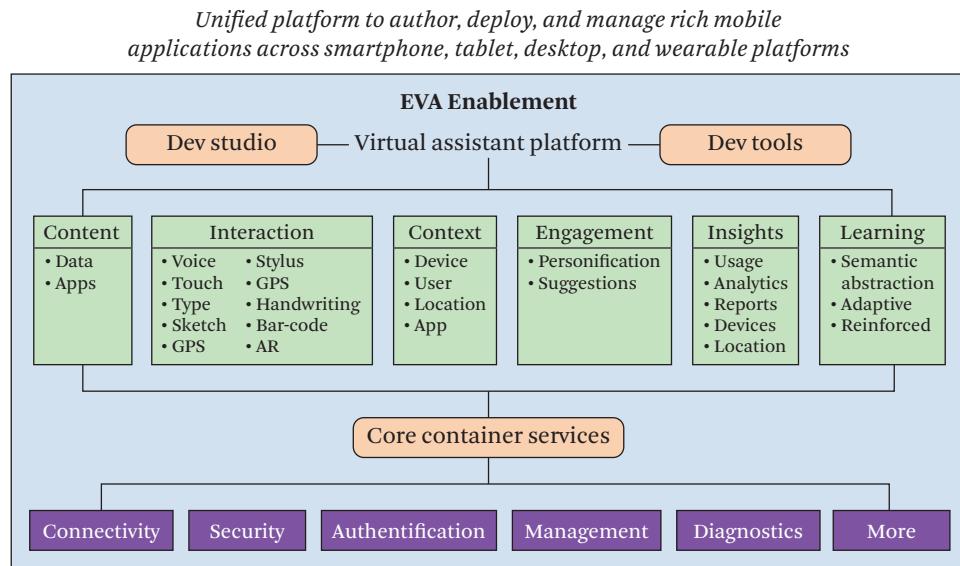
above), can identify objects in real time, search for information about those objects through various on-line services, and enable the user to request to purchase them. The capability can be invoked with a series of menu selections, or via a single voice command. In addition, like Google's Assistant, Bixby Vision can translate text, read QR codes, and recognize physical landmarks.

Almost all of the devices will allow speech coordinated with touch input on graphical output. However, for a conversational experience that at least allows follow-up questions, the systems need to know what information they have provided in the various modalities. For example, if a set of sweaters is provided graphically as an answer to a question, the user should be able to say "buy the red one," even if the word "red" was never spoken by either party, or does not even appear on the screen. Likewise, the user should be able to say "buy the larger one," even if the items displayed are "small" and "medium" sized.

The academic multimodal research community has worked through such issues, aptly captured by the slogan "No presentation without representation" [Wahlster et al. 1991, Wahlster 2006]. Presentation planning software has been designed to derive multimodal output from a representation of the content to be presented, which is then divided across modalities (multimodal fission). Chapter 4 of this volume details some of these efforts.

The most capable multimodal virtual assistant is from Openstream, which has developed the Enterprise Virtual Assistant (EVA™) that is built with the EVA Café™ (earlier called CueMe Studio) development kit for multimodal systems. EVA Café is based on the World Wide Web Consortium (W3C) international standard reference architecture (see Chapter 9), which includes State Chart XML (SCXML), a W3C standard XML-based markup for state machine [Barnett et al. 2015] representation of the multimodal application's Interaction Manager (IM), and EMMA, an XML-based message content standard [Johnston 2009]. The EVA Café software framework (Figure 15.9) can capture user input from a number of modalities, including speech, sketch, handwriting, keyboard, stylus, touch, GPS, bar-code scanners, etc. The input modalities are sent to an Interaction Manager server (IM, per W3C standard, see Chapter 9) for (late) hybrid multilevel fusion and understanding, which then invokes the back-end software. EVA Café has been used for filling forms, annotating drawings (including synchronized voice and drawing), question-answering, personal information management, and other functionality.

EVA Café implements a Delivery Context component that lets the IM adjust the system output based on the output capabilities of delivery contexts, including the type of screen, presence of audio speakers, and their state (e.g., the speakers are muted). The IM can coordinate audio information presentation across multiple



**Figure 15.9** EVA multimodal virtual assistant layered on multimodal and other components that are instantiated in the W3C Multimodal Architecture.

connected devices, including cell phone, tablet, and personal computer, determining at runtime the type of commands (via XML markup) it needs to send to a given device (delivery context) based on the system state. For example, it can allow the visual modality to show more content when a larger screen is detected. Likewise, when rapid motion is detected, such as when the user is driving, the IM can automatically suppress displaying information, and read (some of) it aloud instead.

Using EVA Café, Openstream and its clients, including Walmart, Merck, Roche, and Bank of NY/Mellon, Thomson Reuters, and Chubb, have developed a series of multimodal applications, such as mobile field-force automation, financial, pharmaceutical, and other applications. Figure 15.9 provides an example of some of EVA's multimodal capabilities.

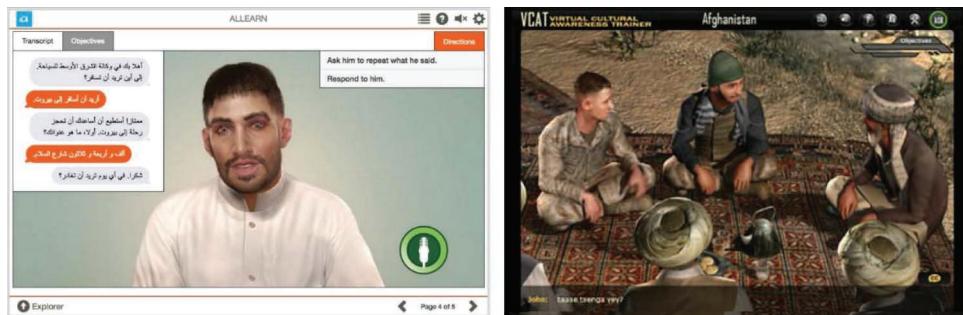
An emerging trend is to offer multimodal development tools. In addition to Openstream's EVA Café as of January 2019, other organizations and device manufacturers have announced the availability of multimodal development tools, notably Microsoft's Platform for Situational Intelligence (PSI) [Bohus et al. 2017], and Oppo's breeno (<https://newsbeamer.com/zimbabwe/oppo-launches-smart-assistant-breeno-and-developer-finance/>), both of which are still in beta.



**Figure 15.10** The Soul Machines Ava and Will avatars (top left and right). Ava is used by Autodesk as its virtual assistant, while Will is used by Toshiba. The avatars are derived from 3D images and motion capture of a real human but with facial muscle movement driven by trained biologically inspired neural networks (bottom left) prototyped in the BabyX avatar (bottom right). (Courtesy of Soulmachines.com)

### 15.5.2 Avatars

Soul Machines (<http://www.soulmachines.com>) creates lifelike 3D multimodal “digital humans” (hereafter “avatar”) that perceive speech, natural language, and emotion, and are capable of rendering fine-grained facial movements to signal emotional output. The avatars perceive the user’s voice and facial input, recognize speech and faces, and process natural language with technologies supplied by a client company. The avatars are powered by a collection of biologically inspired neural networks (Figure 15.10, bottom), that (on output) drive lips, eyes, facial muscles, and head motion. Detailed models of the motion of muscles (individual or groups), fascia, fat, and connective tissue in response to simulated neural networks drive realistic skin deformation. Emotions are combinations of brain-body



**Figure 15.11** Alelo's multimodal avatars for language and culture training. (© Alelo Inc. Reprinted with permission)

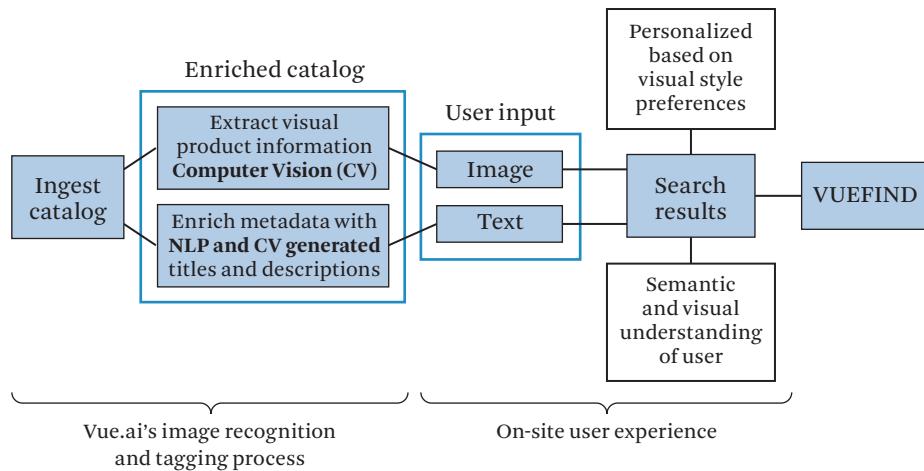
states that alter yet other behavioral circuits (e.g., lowering the threshold for the pattern of facial muscles that constitute crying, laughing, etc.). Because emotions are decomposed into the interaction of multiple muscle groups, essentially a superset of facial action codings [Ekman and Friesen 1978], an avatar can express multiple emotions simultaneously (akin to forming chords from piano keys). The customer scripts the generated emotions, based on the desired interactions of the supplied conversational agent. The avatars are licensed to various companies, such as Autodesk, ANZ Bank, and Daimler, to be the face of their virtual assistants.

Alelo (<http://www.alelo.com>) builds multimodal avatars for language and culture training. Students engage these avatars via voice input, often to learn a natural language (e.g., Arabic, see Figure 15.11, left or English), or to participate in realistic cultural encounters, such as to visit a village elder (Figure 15.11, right). The avatars generate scripted voice outputs and coordinated body movements built with the Alelo authoring platform.

## 15.6

### Product Search

Many vendors are beginning to offer combine multimodal input/output and machine-learning in order to provide better product search and interaction capabilities. For example, Vue.ai (<https://Vue.ai>) enables search of clothing with text and image inputs, and allows a variety of personalization features (see Figure 15.12). Its ingestion engine consumes various product catalogs, performing feature extraction on images and descriptions of products, and creating a multimodally searchable knowledge base. When a user inputs an image of a dress and/or describes it as evening gown, the system will retrieve items matching that criteria filtered by user's



**Figure 15.12** Multimodal catalogue search. (Source: Vue.ai)

preferred colors and sizes. Multimodal search is a topic of considerable research, see for example Laenen et al. [2018].

## 15.7 Virtual and Augmented Reality

Virtual and augmented reality (AR) applications generally require the user to wear an opaque or see-through headset on which is displayed either the virtual world, or information overlaid onto the actual world, respectively. The user's head motion is tracked using optical, magnetic, gyroscopic, or other technology, such that s/he can view the entire scene by turning his/her head. In addition to projecting the visual scene, the headsets usually offer spatial audio output, as well as noise-canceling microphone input. Thus, the headset device is inherently a multimodal-multisensor device. Some headsets such as Microsoft's Hololens 2™ and HTC Vive Pro Eye™ also incorporate eye-tracking. Headsets are offered both to consumers, usually for gaming, and more often to businesses. For example, Hololens 2 is only offered to commercial customers. Among the AR devices currently being offered are as follows:

1. Microsoft's Hololens 2 (display, spatial audio)—cameras, IMU (gyroscope, magnetometer, accelerometer), collaboration (remote assist), time-of-flight camera enabling hand gesture recognition in the scene, voice commands, and the avoice-enabled Cortana virtual assistant with voice output. Entities in the scene can be selected based on what the user is looking at.



**Figure 15.13** Augmented reality glasses and use in a warehouse. (Photographs courtesy of Vusiz)

2. The Magic Leap One™ headset incorporates multimodal input and output, including audio input/output, camera input, six degree of freedom (DOF) head-tracking, a trackpad, haptics in the headset, spatial audio, six DOF device for pointing, including “mappable” buttons.
3. The Vuzix Blade augmented reality headset (see Figure 15.13) offers voice control, camera, head-tracking (gyroscope, magnetometer, accelerometer), touchpad upon which the user can perform gestures, two noise-canceling microphones, buttons on the headset, and speakers.

Virtual Reality (i.e., opaque headsets) are primarily offered for gaming applications, and often require instrumented environments that support the head tracking technologies. Two main headsets are:

1. Oculus Rift™: optical head tracking, spatial audio, microphones, voice input and output, two 6 DOF 3D pointing devices; and
2. HTC Vive Pro Eye: laser-based head tracking, spatial audio, microphone, handheld sensors with buttons for pointing and selection, and eye-tracking.

Often, the headsets are coupled with 3D point devices which, unfortunately, are frequently used as 3D “mice,” rather than employing more natural methods for interacting with a 3D scene. A different approach can be found in [Kaiser et al. \[2003\]](#). Although the devices may capture multiple modalities of input, to date, few applications have made serious use of true integrated multimodal interaction. Table 15.3 compares these multimodal-multisensor devices.

**Table 15.3** Comparison of augmented reality and virtual reality headsets

Augmented Reality					Virtual Reality	
Company / Product						
Microsoft Hololens 2	Magic Leap One	Vuzix Blade	Oculus Rift	HTC Vive Pro		
<b>Input Modalities and Sensors</b>						
Cameras	2		X		X	
Inertial	X		X			
<b>Measurement</b>						
Microphones	X	X	2	X	X	
Voice	X	X	X			
Head-tracking	X	X	X	Optical	Laser	
Haptics			X			
Eye-tracking	X	X			X	
Trackpad		X	X			
Gestures	in scene		on trackpad			
Time-of-flight camera for 3D	X					
Altitude						
Humidity						
Ambient light						
<b>Output Modalities and Devices</b>						
Spatial audio	X	X	X	X	X	
Mixed reality	X	X	X			
Voice	X	X	X	X	X	
Pointing device		X	X	X	X	
Speakers	X	X	X	X	X	
Buttons			X	X	X	

## 15.8

### Field Force Automation

Over the last decade there have been many digital applications that enable field workers to use multimodal interaction to complete data gathering, image capturing, form-filling, or even chat/instant-messaging via their mobile devices and measuring/scanning instruments.

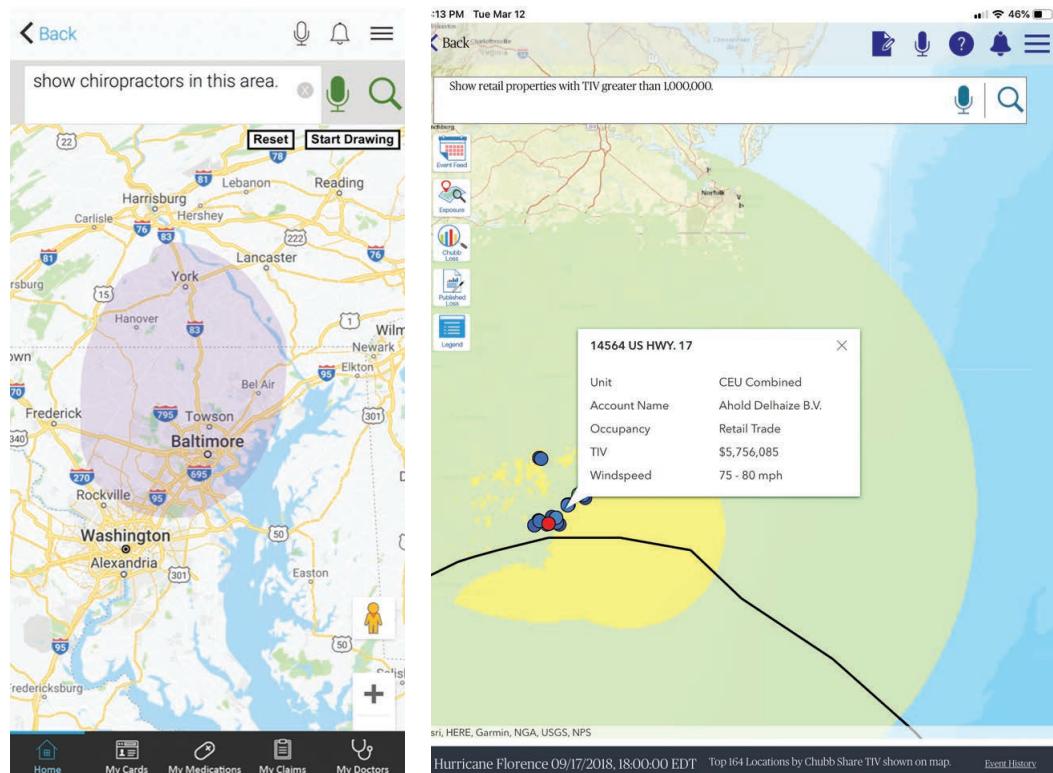
These devices have evolved over time, and typically now feature a stylus, touch-screen, microphone and speakers, barcode/QRcode scanners, cameras, and other sensors. They enable hands-free input through speech where appropriate and help faster and accurate product scans through barcode/QR code. For example, warehouse personnel can be directed verbally by a computer to pick certain items off a shelf. The personnel scan the desired items (e.g., with a ring-mounted scanner), and speak a confirmation to the computer of the number of those items picked.

Field workers typically ask questions using their voice regarding what they see in the field of view of their augmented reality glasses or head-mounted display. A spoken and/or graphical display of the response will be given to the user, who can further interact with the display through a change of gaze or a touch/tap on the side of the glasses. Wearable multimodal technologies such as augmented reality glasses have had such a major impact on field force solutions that we find them everywhere from warehouses to supply-chain logistics. AR glasses are commercially available from several vendors like Vuzix (see Section 15.7) and are in use by many fieldforce/customer relationship management solutions from vendors like Salesforce, SAP, and Openstream.

## 15.9 Insurance

The insurance industry is going through the process of digital transformation, where many new insurance solution providers are offering self-service options, simplifying the process of finding and buying insurance policies as well as submitting claims and paying bills. One company offering multimodal solutions for insurance applications is Openstream.

Openstream's (<http://www.openstream.com>) virtual assistant-based insurance solution, allows users to track catastrophic events on a map with touch and speech inputs near the properties of their interest (see Figure 15.14). Users can capture images and highlight the damaged areas on the property (a home or auto) with speech & ink annotations while submitting their claims. The system allows mixed-initiative multimodal interaction while filling the forms, combining speech and other sensory inputs. Users can find relevant locations on a map, in the event of a catastrophe by using plan based dialog methods. They can also search through policy documents for relevant information by talking/typing, and they can receive a section of the explanation of benefits translated to their chosen language by selecting the region of text (paragraph or from/to) and specifying the language to translate and speak aloud (see also Chapter 13). The solution has the ability to ingest images/documents with multimodal annotations and determine the extent



**Figure 15.14** Left: Multimodal insurance query with speech + drawing: “Show chiropractors in this area?” Right: Show retail properties with TIV(total insured value) greater than 1,000,000. (Courtesy of Openstream)

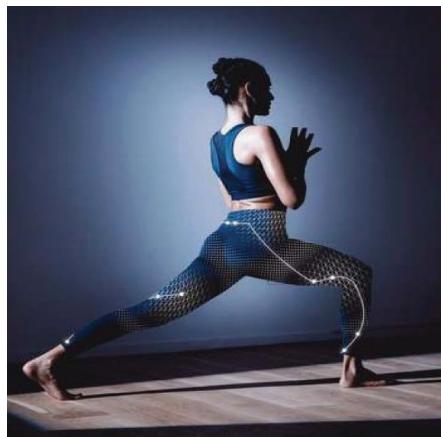
of damage, and recommend a claim amount that can further be approved by a human agent with appropriate authority.

## 15.10 Personal Care Products

The hair and skin care industries have adopted multisensor technologies. Kérastase, in partnership with Withings (a Nokia company) and L’Oréal, have created a smart hairbrush that incorporates a microphone that analyzes brushing patterns in relation to the type of hair (frizzy, dry, split ends, etc.), load sensors that analyze brushing force, accelerometer and gyroscope that analyze brushing pattern, and haptic feedback to signal the user if the brushing is too vigorous. The user receives a summary of the brushing session on a mobile app (see Figure 15.15).



**Figure 15.15** Kérastase Hair Coach<sup>TM</sup>. (Courtesy of L'Oréal)



**Figure 15.16** Nadi X<sup>TM</sup> multisensor yoga leggings. (Courtesy of Wearable X)

Neutrogena is offering Skin360<sup>TM</sup>, a combination of a multisensory “skin scanner” and a smartphone application. The scanner, which fits over an iPhone camera lens, visually analyzes the skin for fine lines, pores, wrinkles, etc., and also contains a moisture meter that reads subcutaneous hydration levels. The app tells the user where to place the scanner, accumulates the results, produces a facial analysis and product recommendations. Other multimodal-multisensor skin care applications are offered by Shiseido and Modiface.ai.

Relevant to Chapters 7 and 8 in this volume, Wearable X (<https://www.wearablex.com>) sells multisensor yoga leggings called Nadi XTM, which communicate wirelessly with a mobile phone. The user attaches a battery-operated Bluetooth-enabled transducer package to a connector mounted behind the knee. Then, the user chooses the desired pose on the phone. Sensors in the garment sense the position of the leg and trunk (see Figure 15.16), and transmit that data via the transducer to the Nadi X app on the user's mobile phone. The app compares the leg and trunk positions with the normative ones for the chosen pose, and sends haptic vibrations back to the user's body to signal which body parts need to be adjusted.

Kaia Health has developed the Perfect Squat Challenge app for iPhones (<https://kaia-perfect-squats.app.link/challenge>). Using the iPhone's camera, the app analyses the user's squat positioning by tracking 16 points on the user's body. Then, it makes verbal and pictorial recommendations on how to alter the pose to achieve the best result.

## 15.11 Emotion Recognition

Volume 2 Chapter 6 discusses the state of the art in research on multimodal emotion recognition. In this section, we discuss how some of that research is beginning to be translated into commercial products. Affectiva Corp (<https://www.affectiva.com>) sells multimodal emotion recognition capabilities to a variety of other companies for a variety of purposes via a software development kit. For example, emotion recognition is being used in order to assess how an advertisement, product, movie, or TV show is being received by an audience, or how drivers and passengers are feeling in an automobile, such as when a person is tired or angry. Affectiva's emotion recognition is primarily accomplished with analyses of facial expressions derived from cameras, and includes seven states: anger, contempt, disgust, fear, joy, sadness, and surprise. Currently, under development is real-time multimodal emotion recognition that also incorporates prosodic information.

Companies such as Behavioral Signals (<https://www.behavioralsignals.com>), Cogito (<https://www.cogito.com>), CompanionMX (<https://www.companionmx.com>), and Audeering (<https://www.audeering.com>) are emphasizing emotion recognition and sentiment analysis derived from verbal or textual data. These companies are targeting numerous industries, including: call centers (analyzing callers' verbally expressed emotions), automotive (analyzing both people's emotions and car noises for diagnostic reasons), virtual reality, healthcare, and forensics. Companies using such products include BMW, Daimler, Humana, Mercedes, MetLife, and others. Their products are also being incorporated into multimodal emotion recognition solutions (e.g., from Emteq <https://www.emteq.net>). CompanionMX is

offering a mobile multimodal-multisensor device-based solution to identify users who suffer from mood disorders are at risk for suicide. The product, Companion™, includes a patient-facing app, cloud-based analytics and a clinician-facing dashboard. The solution is based on recent research [Place et al. 2017] that showed that classification of users' mood could be derived from a combination of vocal features (prosody, voice quality, and tone) from an audio diary. The solution also models users' diminished interest in activities, fatigue, and avoidance of people/places, based on smartphone meta-data obtained from measures of distance traveled, and outgoing call/message data. Finally, a Chinese company Emotibot (<http://www.emotibot.com/EN.html?n=42>) sells multimodal emotion recognition technology based on facial and vocal analyses, for sales, call centers, and financial applications.

## 15.12

### Summary

This chapter has provided a broad survey of commercial products that incorporate multimodal-multisensor interfaces. Of course, these are today's products, many of which may fail in the marketplace, and new ones may take their places. However, it should be clear that the pace of multimodal-multisensor interface technologies is accelerating, which will continue to spawn entirely new types of applications. A major question for researchers and technology developers is how to build applications with multimodal-multisensor technologies in ways that respect human privacy and ethics, topics discussed in Chapter 16.

### Focus questions

- 15.1. Why is multisensor fusion important for aerospace applications?
- 15.2. How is multisensor fusion accomplished for aerospace applications?
- 15.3. How are various modalities being incorporated for surgical robots?
- 15.4. What advances in multimodal control have taken place with prosthetic arms?
- 15.5. How do social robots make use of multiple modalities for human-robot interaction?
- 15.6. Which industries have need of multimodal biometrics? How have you seen multimodal biometric solutions deployed so far?
- 15.7. How can virtual assistants and avatars (e.g., Siri, Alexa, etc.) make use of multiple modalities? Which is the most advanced for input, and for output?
- 15.8. How can field workers with mobile devices and/or head-mounted displays take advantage of multimodal interfaces?

## References

- S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. *Cognitive Behavioural Systems. Lecture Notes in Computer Science*. Springer. DOI: [10.1007/978-3-642-34584-5\\_9](https://doi.org/10.1007/978-3-642-34584-5_9). 633
- M. Allodi, A. Broggi, D. Giaquinto, M. Patander, and A. Piroletti. 2016. Machine learning in tracking associations with stereo vision and lidar observations for an autonomous vehicle. In *IEEE Intelligent Vehicles Symposium (IV), 2016*, pp. 648–653. DOI: [10.1109/IVS.2016.7535456](https://doi.org/10.1109/IVS.2016.7535456). 625, 626, 777
- E. Alpaydin. 2018. Classifying multimodal data. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. DOI: [10.1145/3107990.3107994](https://doi.org/10.1145/3107990.3107994). 636
- A. Aviles-Rivero, S. M. Alsaleh, J. Philbeck, S. P. Raventos, N. Younes, J. K. Hahn, and A. Casals. August 2018. Sensory substitution for force feedback recovery: a perception experimental study. *ACM Transactions on Applied Perception (TAP)*, 15(3). DOI: [10.1145/3176642](https://doi.org/10.1145/3176642). 629
- J. Barnett, R. Akolkar, R. J. Auburn, M. Bodell, D. C. Burnett, J. Carter, S. McGlashan, T. Lager, M. Helbing, R. Hosn, T. V. Raman, K. Reifernrath, and N. Rosenthal. 2015. State Chart XML (SCXML): State machine notation for control abstraction. <http://www.w3.org/TR/scxml/> 640
- M. J. Barry, P. M. Gallagher, J. S. Skinner, and F. J. Fowler Jr. February 2012. Adverse Effects of Robotic-Assisted Laparoscopic Versus Open Retropubic Radical Prostatectomy Among a Nationwide Random Sample of Medicare-Age Men. *Journal of Clinical Oncology*, 30(5): 513–518. DOI: [10.1200/JCO.2011.36.8621](https://doi.org/10.1200/JCO.2011.36.8621). 628
- B. T. Bethea, A. M. Okamura, M. Kitagawa, T. P. Fitton, S. M. Cattaneo, V. L. Gott, W. A. Baumgartner, and D. D. Yuh. June 2004. Application of Haptic Feedback to Robotic Surgery. *Journal Laparoendosc Advanced Surgical Techniques*, 14(3): 191–195. DOI: [10.1089/1092642041255441](https://doi.org/10.1089/1092642041255441). 629
- L. Bodenhagen, S. D. Suvei, W. K. Juel, E. Brander, and N. Krüger. 2019. Robot technology for future welfare: meeting upcoming societal challenges—an outlook with offset in the development in Scandinavia, *Health Technology*. DOI: [10.1007/s12553-019-00302-x](https://doi.org/10.1007/s12553-019-00302-x). 635
- D. Bohus, S. Andrist, and M. Jalobeanu. November 2017. Rapid development of multimodal interactive systems: a demonstration of platform for situated intelligence. *ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction*. DOI: [10.1145/3136755.3143021](https://doi.org/10.1145/3136755.3143021). 641
- J. Cacace, A. Finzi, and V. Lippiello. 2016. Multimodal interaction with multiple co-located drones in search and rescue mission. In *Proceedings of IEEE International Conference on Robot and Human Interactive Communication*. RoMan.

- S. Calderwood, K. McAreavey, W. Liu, J. Hong. 2017. Context-dependent combination of sensor information in Dempster-Shafer theory for BDI. *Knowledge Information System*. 51, 259–285. DOI: [10.1007/s10115-016-0978-0](https://doi.org/10.1007/s10115-016-0978-0). 624, 767
- D. Cameron, A. Millings, S. Fernando, E. Collins, R. Moore, A. Sharkey, V. Evers, and T. Prescott. April 2015. The effects of robot facial emotional expressions and gender on child-robot interaction in a field study. *4th International Symposium on New Frontiers in Human-Robot Interaction*, 632
- R. O. Chavez-Garcia and O. Aycard. 2016. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 17(2): 525–534. DOI: [10.1109/TITS.2015.2479925](https://doi.org/10.1109/TITS.2015.2479925). 626
- J. Choi, S. Ulbrich, B. Lichte, and M. Maurer. 2013. Multi-target tracking using a 3d-lidar sensor for autonomous vehicles. *16th International IEEE Conference on Intelligent Transportation Systems-(ITSC)*, pp. 881–88. DOI: [10.1109/ITSC.2013.6728343](https://doi.org/10.1109/ITSC.2013.6728343). 626
- K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. Assif Mirza, and M. Blow. 2009. KASPAR—a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3–4): 369–397. 635
- L. A. Dickstein-Fischer, D. E. Crone-Todd, I. M. Chapman, A. T. Fathima, and G. S. Fischer. 2018. Socially assistive robots: current status and future prospects for autism interventions. *Innovation and Entrepreneurship in Health*, 5: 15–25. 635
- J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell. 2012. The clinical use of robots for individuals with autism spectrum disorders: a critical review, *Research in Autism Spectrum Disorder*, 6(1): 249–262. DOI: [10.1016/j.rasd.2011.05.006](https://doi.org/10.1016/j.rasd.2011.05.006). 635
- P. Ekman and W. V. Friesen. 1978. Facial action coding system. *A Technique for Measuring Facial Movement*, Consulting Psychologists Press. Palo Alto. 643
- J. Elfring, R. Appeldoorn, S. van den Dries and M. Kwakkernaak. 2016. Effective world modeling: Multisensor data fusion methodology for automated driving. *Sensors*, 16: 1668. DOI: [10.3390/s16101668](https://doi.org/10.3390/s16101668). 627
- N. Enayati, E. De Momi, and G. Ferrigno. February 2016. Haptics in robot-assisted surgery: challenges and benefits. *IEEE Reviews In Biomedical Engineering*. DOI: [10.1109/RBME.2016.2538080](https://doi.org/10.1109/RBME.2016.2538080). 629
- J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho. November 2018. Multiple classifiers in biometrics. Part 1: Fundamentals and review. *Information Fusion* 44. DOI: [10.1016/j.inffus.2017.12.003](https://doi.org/10.1016/j.inffus.2017.12.003). 637
- Food and Drug Administration. 2019. Implanted brain-computer interface (BCI) devices for patients with paralysis or amputation—Non-clinical testing and clinical considerations. US Food and Drug Administration.
- P. Fosse. November 2019. Deep dive into Tesla's autopilot & self-driving architecture vs. Lidar-based systems. *Clean Technica*. <https://cleantechnica.com/2018/11/04/deep-dive-into-teslas-autopilot-self-driving-architecture-vs-lidar-based-systems/> 627

- H. Franco, J. Zheng, J. Butzberger, F. Cesari M. Frandsen, J. Arnold, V. R. R. Gadde, A. Stolcke, and V. Abrashy. March 2002. DynaSpeak: SRI's scalable speech recognizer for embedded and mobile systems. *HLT '02 Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc. San Francisco CA. DOI: [10.1.1.156.7028](https://doi.org/10.1.1.156.7028). 623
- T. L. Frey, C. Aguilar, K. Engebretson, D. Faulk, and L. G. Lenning. 2018. F-35 information fusion. *2018 Aviation Technology, Integration, and Operations Conference*. Atlanta, GA. DOI: [10.2514/6.2018-3520](https://doi.org/10.2514/6.2018-3520). 623
- T. Gulrez. 2014. Role of Haptic Interfaces in Robot-Assisted Minimally Invasive Surgery. *International Journal of Swarm Intelligence and Evolutionary Computing*, 3(2). 629
- R. Haenni. July 2004. Shedding new light on Zadeh's criticism of Dempster's rule of combination. *7th International Conference on Information Fusion*, Stockholm, Sweden. DOI:[10.1109/ICIF.2005.1591951](https://doi.org/10.1109/ICIF.2005.1591951). 624, 767
- M. Johnston. 2009. Building multimodal applications with EMMA. *International Conference on Multimodal Interfaces*. Cambridge, MA. 640
- G. Jones, N. Berthouze, R. Bielski, and S. Julier. 2010. Towards a situated, multimodal interface for multiple UAV control. *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1739–1744. DOI: [10.1109/ROBOT.2010.5509960](https://doi.org/10.1109/ROBOT.2010.5509960).
- S. J. Julier and J. K. Uhlmann. 2004. Unscented filtering and nonlinear estimation. In *Proceedings of the IEEE*, 92(3): 401–422. DOI: [10.1109/JPROC.2003.823141](https://doi.org/10.1109/JPROC.2003.823141). 625, 626, 777
- E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. *Proceedings of the 5th International Conference on Multimodal Interfaces*, pp. 12–19. ACM Press. DOI: [10.1145/958432.958438](https://doi.org/10.1145/958432.958438). 645
- R. E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82: 35. DOI: [10.1115/1.3662552](https://doi.org/10.1115/1.3662552). 625, 777
- S. Kane, K. McGurgan, M. Voshell, C. Monnier, S. German, and A. Ost. 2017. A multimodal interface for natural operator teaming with autonomous robots (MINOTAUR). In J. Chen editor, *Advances in Human Factors in Robots and Unmanned Systems*, pp. 99–108. Springer.
- R. Kasper and S. Schmidt. September 2008. Sensor-data-fusion for an autonomous vehicle using a Kalman-filter. *6th International Symposium on Intelligent Systems and Informatics*, pp. 26–27. Subotica, Serbia. DOI: [10.1109/SISY.2008.4664905](https://doi.org/10.1109/SISY.2008.4664905).
- J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3): 226–239. DOI: [10.1109/34.667881](https://doi.org/10.1109/34.667881). 636
- J. Kocic, N. Jovićic, and V. Drndarevic. November 2018. Sensors and sensor fusion in autonomous vehicles. *6th Telecommunications Forum TELFOR*.

- K. Laenen, S. Zoghbi, and M.-F. Moens. 2018. Web search of fashion items with multimodal querying. In *Proceedings Of the 18th International Conference on Web Search and Data Mining*, pp. 342–350. ACM Press. DOI: [10.1145/3159652.3159716](https://doi.org/10.1145/3159652.3159716). 644
- M. A. Lebedev and M. A. L. Nicolelis. 2017. Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiological Reviews*, 97(2): 767-837. American Physiological Society. DOI: [10.1152/physrev.00027.2016](https://doi.org/10.1152/physrev.00027.2016). 631
- G. Lemons, K. Carrington, T. Frey, and J. Ledyard. June 2018. F-35 mission systems design, development and verification. *AIAA AVIATION Forum, 2018 Aviation Technology, Integration, and Operations Conference*, pp. 25–29. Atlanta, GA. DOI: [10.2514/6.2018-3519](https://doi.org/10.2514/6.2018-3519). 623
- T. S. Lendvay, B. Hannaford, and R. M Satava. 2013. Future of robotic surgery. *The Cancer Journal* 19(2): 109–119. 628
- M. Liang, B. Yang, S. Wang, and R. Urtasun. 2018. Deep continuous fusion for multi-densor 3D object detection. *Proceedings of European Conference on Computer Vision (ECCV)*. 626
- D. A. Maerz, L. N. Beck, A. J. Sim, and D. M. Gainsburg. 2017. Complications of robotic-assisted laparoscopic surgery distant from the surgical site. *British Journal of Anaesthesia*, 118(4): 492–503. DOI: [10.1093/bja/aex003](https://doi.org/10.1093/bja/aex003). 628
- T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. 2002. Impact of artificial “Gummy” fingers on fingerprint systems. *Proceedings of SPIE Vol. #4677, Optical Security and Counterfeit Deterrence Techniques*. DOI: [10.1117/12.462719](https://doi.org/10.1117/12.462719). 637
- I. Maza, F. Caballero, R. Molina, N. Pena, and A. Ollero. 2010. Multimodal interface technologies for UAV ground control stations: A comparative analysis. *Journal of Intelligent and Robotic Systems*, 57(1–4): 371–391. DOI: [10.1007/s10846-009-9351-9](https://doi.org/10.1007/s10846-009-9351-9).
- C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernocký, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes. 2012. Bi-modal person recognition on a mobile phone: using mobile phone data. *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 635–640. DOI: [10.1109/ICMEW.2012.116](https://doi.org/10.1109/ICMEW.2012.116). 637
- G. Meccariello, F. Faedi, S. AlGhamdi, F. Montevercchi, E. Firinu, C. Zanotti, D. Cavaliere, R. Gunelli, M. Taurchini, A. Amadori, and C. Vicini. 2016. An experimental study about haptic feedback in robotic surgery: may visual feedback substitute tactile feedback? *Journal of Robotic Surgery*, 10(1): 57–61. DOI: [10.1007/s11701-015-0541-0](https://doi.org/10.1007/s11701-015-0541-0). 628
- A. Melamed, D. J. Margul, L. Chen, N. L. Keating, M. G. Del Carmen, J. Yang, B. L. Seagle, A. Alexander, E. L. Barber, L. W. Rice, J. D. Wright, M. Kocherginsky, S. Shahabi, and J. A. Rauh-Hain. 2018. Survival after Minimally Invasive Radical Hysterectomy for Early-Stage Cervical Cancer. *New England Journal of Medicine*. 15,379(20): 1905-1914. Epub. DOI: [10.1056/NEJMoa1804923](https://doi.org/10.1056/NEJMoa1804923). 629
- R. Mola. 2017. Super Helmet: F-35 pilots get X-ray vision and other magic powers. *Air and Space Magazine*, September, 2017. <https://www.airspacemag.com/military-aviation/super-helmet-180964342/>. 623

- C. Monnier. May 2017. Developing a multimodal ugv robot control interface. *Aerospace and Defense Technology*.
- R. Murphy. 1998. Dempster-Shafer theory for sensor fusion in autonomous mobile robots. *IEEE Transactions on Robotics and Automation*, 14(2): 197–206. DOI: [10.1109/70.681240](https://doi.org/10.1109/70.681240). 624, 767
- A. M. Okamura. 2009. Haptic feedback in robot-assisted minimally invasive surgery. *Current Opinions in Urology*, 19(1): 102–107. DOI: [10.1097/MOU.0b013e32831a478c](https://doi.org/10.1097/MOU.0b013e32831a478c). 629
- C. Pacchierotti, L. Meli, F. Chinello, M. Malvezzi, and D. Prattichizzo. 2015. Cutaneous haptic feedback to ensure the stability of robotic teleoperation systems. *The International Journal of Robotics Research*, 34(14): 1773–1787. DOI: [10.1177/0278364915603135](https://doi.org/10.1177/0278364915603135). 628
- D. Pfeiffer and U. Franke. 2010. Efficient representation of traffic scenes by means of dynamic stixels. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 217–224. DOI: [10.1109/IVS.2010.5548114](https://doi.org/10.1109/IVS.2010.5548114). 625, 626, 777
- S. Place, D. Blanch-Hartigan, C. Rubin, C. Gorrostieta, C. Mead, J. Kane, B. P. Marx, J. Feast, T. Deckersbach, A. Pentland, A. Nierenberg, and A. Azarbeyjani. 2017. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of Medical Internet Research*, 19(3): e75. Published online 2017 Mar 16. DOI: [10.2196/jmir.6678](https://doi.org/10.2196/jmir.6678). 651
- A. Ross and A. K. Jain. 2007. Fusion techniques in multibiometric systems. In R. I. Hammoud, B. R. Abidi, M. A. Abidi, editors, *Face Biometrics for Personal Identification. Signals and Communication Technology*. Springer, Berlin. DOI: [10.1007/978-3-540-49346-4\\_12.pdf](https://doi.org/10.1007/978-3-540-49346-4_12.pdf). 636
- A. Ross, K. Nandakumar, and A. Jain. 2006. *Handbook of Multibiometrics*. Springer, Berlin Heidelberg. 636
- A. Ross and N. Poh. 2009. Multibiometric systems: overview, case studies, and open issues. In M. Tistarelli, S. Z. Li, R. Chellappa, editors, *Handbook of Remote Biometrics for Surveillance and Security*, pp. 273–292. Springer. DOI: [10.1007/978-1-84882-385-3\\_11](https://doi.org/10.1007/978-1-84882-385-3_11). 636, 637
- J. P Ruurda, IAMJ Broeders, B. Pulles, F. M. Kappelhof, and C. Van der Werken. 2004. Manual robot assisted endoscopic suturing: time-action analysis in an experimental model. *Surgical Endoscopy and Other Interventional Techniques*, 18(8): 1249–1252. DOI: [10.1007/s00464-003-9191-9](https://doi.org/10.1007/s00464-003-9191-9). 628
- U. R. Sanchez and J. Kittler. 2006. Fusion of talking face biometric modalities for personal identity verification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 5, V-V. 636, 637
- B. Schoettle. 2017. Sensor Fusion: A comparison of sensing capabilities of human drivers and highly automated vehicles. *Report SWOT-2017-12*. University of Michigan, Sustainable Worldwide Transportation. 626, 627
- K. Sentz and S. Ferson. April 2002. Combination of Evidence in Dempster-Shafer Theory. *Sandia Report Sand2002-0835*. 624, 767

- S. Song, Z. Xiang, and J. Liu. 2015. Object tracking with 3d lidar via multi-task sparse learning. *IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 2603–2608. DOI: [10.1109/ICMA.2015.7237897](https://doi.org/10.1109/ICMA.2015.7237897). 626
- J. Stewart. 2019. Why Tesla's Autopilot Can't See a Stopped Firetruck, Wired, August 27, 2018. <https://www.wired.com/story/tesla-autopilot-why-crash-radar/> 627
- A. Tangel and A. Pasztor. March 2019. Boeing to Make Key Change in 737 MAX Cockpit Software. Wall Street Journal. 623
- S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann. 2008. Review of classifier combination methods. In S. Marinai and H. Fujisawa, editors, *Machine Learning in Document Analysis and Recognition*, pp. 361–386. DOI: [10.1007/978-3-540-76280-5\\_14](https://doi.org/10.1007/978-3-540-76280-5_14). 637
- K. Uludağ and A. Roebroeck. November 2014. General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage* Vol. 102, Part 1, pp. 3–10. DOI: [10.1016/j.neuroimage.2014.05.018](https://doi.org/10.1016/j.neuroimage.2014.05.018). 621
- US FDA. February 2019. *Implanted Brain-Computer Interface (BCI) devices for patients with paralysis or amputation—non-clinical testing and clinical considerations draft guidance for industry and Food and Drug Administration Staff*. United States Food and Drug Administration. <https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM631786.pdf> 631
- O. A. J. Van der Meijden, and M. P. Schijven. 2009. The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: a current review. *Surgical Endoscopy*, pp. 1180–1190. DOI: [10.1007/s00464-008-0298-x](https://doi.org/10.1007/s00464-008-0298-x). 628
- N. Vogler, S. Heuke, T. W. Bocklitz, M. Schmitt, and J. Popp. July 2015. Multimodal Imaging Spectroscopy of Tissue. *Annual Review of Analytical Chemistry*, 8: 359–387. DOI: [10.1146/annurev-anchem-071114-040352](https://doi.org/10.1146/annurev-anchem-071114-040352). 621
- W. Wahlster, E. Andre, B. Bandyopadhyay, W. Graf, and T. Rist. 1991. WIP: The coordinated generation of multimodal presentations from a common representation. In O. Stock., J. Slack, and A. Ortony, editors, *Computational Theories of Communication and their Application*, Springer, Berlin. DOI: [10.1007/978-3-642-58146-5\\_7](https://doi.org/10.1007/978-3-642-58146-5_7). 640
- W. Wahlster, editor. 2006. SmartKom: Foundations of Multimodal Dialogue Systems. Springer Berlin. 640
- T. Wang, B. Pan, Y. Fu, S. Wang, and Y. Ai. 2017. Design of a new haptic device and experiments in minimally invasive surgical robot. *Innovation in Biomedical Science and Engineering*, pp. 240–250. Journal Computer Assisted Surgery Vol. 22, 2017. Taylor and Francis Online Journal. DOI: [10.1080/24699322.2017.1389402](https://doi.org/10.1080/24699322.2017.1389402). 629
- B. Weber and S. Schneider. 2014. The effects of force feedback on surgical task performance: a meta-analytical integration. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, 150–157. Springer. DOI: [10.1007/978-3-662-44196-1\\_19](https://doi.org/10.1007/978-3-662-44196-1_19). 628

- C. R. Wottawa, B. Genovese, B. N. Nowroozi, S. D. Hart, J. W. Bisley, W. S. Grundfest, and E. P. Dutson. 2016. Evaluating tactile feedback in robotic surgery for potential clinical application using an animal model. *Surgical Endoscopy*, 30(8): 3198–3209. DOI: [10.1007/s00464-015-4602-2](https://doi.org/10.1007/s00464-015-4602-2). 628
- M. Wu and J. Shu. 2018. Multimodal molecular imaging: current status and future directions, *Contrast Media & Molecular Imaging*. DOI: [10.1155/2018/1382183](https://doi.org/10.1155/2018/1382183). 621
- L. A. Zadeh. 1986. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI Magazine*, 85–90. 624, 767

# 16 Privacy Concerns of Multimodal Sensor Systems

Gerald Friedland, Michael Carl Tschantz

## 16.1 Introduction

This chapter explains that ignoring the privacy risks introduced by multimodal systems could have severe consequences for society in the long term. The enormous growth in the availability of technology using multimodal sensors, the increase in sharing such data, and advances in multimedia content analysis threatens privacy. We discuss notions of privacy, prior work on violating and protecting it, and what the future may bring. We focus on how data collection and analysis can enable unexpected and invasive inferences about people. We end with directions for research that may ensure the continued survival of privacy as we know it. Newly introduced key terms are defined in the [Glossary](#).

Increasingly ubiquitous multimodal sensors produce vast quantities of data, for example, the location and acceleration of cell phones, the license plates of passing cars, the pulse of a watch's wearer, the lighting level of a bedroom, the temperature of the bedroom, and the sleeping patterns in that bedroom. The popularity of online web services and distributed file storage, the cloud, along with people's perceived need to share content, leads to much of this data being held by companies whose business model is profiling people. Meanwhile, advances in multimedia content analysis (face recognition, speaker verification, location estimation, etc.) provides the ability to analyze and integrate information like never before.

While discussion of these changes have focused on the “cool” applications they have enabled, they have also enabled novel opportunities for the unethical use of

multimodal sensors and multimedia data. In particular, sensors and data warehousing have created privacy concerns, often ignored by researchers, engineers, and business people. Even the users whose privacy is affected are often unaware of the privacy consequences of using an application or granting an app permission to share data. As this chapter explains, ignoring privacy risks could have severe consequences for society in the long term. The scientific community therefore has an obligation to understand these risks, mitigate the effects, and educate the public on the issues.

In small-scale or isolated multimedia, data analytics have always been a powerful but reasonably contained privacy threat. However, when analyses link together data on an Internet scale, the threat can be severe, pervasive, and difficult to comprehend. The severity of the threat stems from the ability of advance analytics to draw highly personal conclusions about people. The threat can become pervasive if sensors are widely deployed, making monitoring common and difficult to avoid. The incomprehensible nature of the threat follows from the difficulty of anticipating the inferences that sensors might enable.

Even without pervasive monitoring, difficult-to-predict inferences about personal topics have already been made using more traditional sources of information. For example, Target reportedly used data about a person's purchases of scent-free hand lotion and multivitamins to infer personal information about early-stage pregnancy [Duhigg 2012] (cf. [Piatetsky 2014]). Here, the everyday and unavoidable act of shopping for seemingly innocuous products is used to infer something about a person's health. Similarly, Visa has used seemingly quotidian purchases to predict a person's credit worthiness, with carbon-monoxide detectors, premium birdseed, and felt pads for the bottoms of chair legs predicting good credit and buying chrome skull car accessories predicting bad credit [Ciarelli 2010]. Inferences are bound to become even more inscrutable as data from sensors are added to the mix.

Volume 2 of this series, Signal Processing, Architectures, and Detection of Emotion and Cognition, contains two examples of multimodal analyses that both address a pressing problem while introducing the potential for privacy violations. [Cohn et al. \[2018\]](#) discusses using facial expressions, speech patterns, and body movements as features for the automated detection of depression. [Burzo et al. \[2018\]](#) presents methods for using similar features for detecting deception. The diagnosis of depression and the detection of those using deception to commit or get away with crimes are both important tasks, and we should welcome improved approaches to them. However, the features used by them can be collected by anyone with a small budget for the needed sensors, regardless of their intents. For example, by re-purposing the video cameras that often surround sites of employment,

potential employers could covertly screen out applicants with mental illnesses or who can't honestly say they lack any commitments beyond work.

Consider the following example: Imagine a future where multimedia query engines *just work*. Everybody can search multimodal sensor data by topic, location, person, camera identity, and time—even when the uploader did not explicitly include such information as tags or otherwise. For example, when searching for a name, the search engine could bring up images of the person from a webcam supposedly just showing the weather conditions but capturing her in the background. The image could be labeled with the time and place along with a pointer to the chronologically next known image. While each of their own reveal only one point in time, together multiple images enable the reconstruction of her travel, perhaps to and from an area with a clinic. Such information can be further combined with thermal images to detect a fever and with audio to detect a cough. As more and more data is combined, more and more can be learned. Understanding the privacy risks requires understanding all the available data, not just looking at what one source can provide in isolation.

Furthermore, unlike traditional, typically text, sources of data about people, which is explicitly about them, such background information may be unintended by either the person shown in the photo, the photographer, or the uploader. As multimedia analysis improves, the ability to extract information from the background of photos, video, and audio, or similar incidental information from other media, will increase the amount of unintended data returned by the search engine about a person. As the amount of searchable data increases, people would find it increasingly difficult to understand what inferences could be made about them from it.

An unscrupulous attacker could, for example, query for videos recently recorded at resorts and then find videos taken with the same camera in nearby wealthy residential neighborhoods. This would produce an ideal “hit list” of targets who are likely away from home, which a thief could then refine. As reported in previous publications (see Section 16.3), such *cybercasing* already occurs, but with a multimedia query engine, simple methods of anonymizing posts and suppressing metadata will no longer be enough. Rather, the multimedia community must work to educate the public about the risks of analytics at the Internet scale, invent methods to identify when information (such as the “identity” of the camera) is being unintentionally leaked, and develop mitigation techniques to reduce the potential harm.

Providing such privacy protections will not be easy. They must work not just for a single online post but rather for a person’s entire information footprint, which can be hard to know. Furthermore, they must work for not just the information intended to be shared, but for all the unintentional background information as well. These

are moving targets as multimodal sensors proliferation, data become more abundant and retrievable, and multimedia analyses improve. Education efforts must be appropriate for the level of background most people have, for example, by providing concrete illustrations of threats instead of abstract analyses.

After defining the topic and presenting some definitions and early work (Sections 16.2 and 16.3) in the area, this chapter outlines existing and future multimedia content analysis and linking techniques that could support unethical use and describe possible attack vectors (Section 16.4). This chapter then continues with describing the landscape of currently ongoing data sharing of multimodal sensor data in the contemporary economy. Finally, we outline mitigation and educational techniques (Section 16.6) and conclude that this is a new topic to be explored (Section [Focus Questions](#)).

## 16.2 Calls for and Types of Privacy

Multimodal data collection, data warehousing, and automated data processing are touching a wide range of fields including predictive policing [[Perry et al. 2013](#)], credit rating [[Alloway 2015](#)], healthcare [[Murdoch and Detsky 2013](#)], education [[Baker et al. 2015](#)], and defense [[PRNewswire 2015](#)]. Concerns over its impacts have motivated numerous calls for privacy in data processing including from the U.S. Federal Trade Commission [[Secretary's Advisory Committee on Automated Personal Data Systems 1973](#)], the White House [[The White House 2012, Podesta et al. 2014, President's Council of Advisors on Science and Technology 2014, Executive Office of the President 2016](#)], the European Commission [[2016](#)], and the [Council of the OECD \[1980\]](#).

Two chapters, located in two different volumes of this three-volume series of books, observe that ensuring privacy is a pre-condition for the new technologies they discuss being adopted. Chapter 11 of Volume 2 [[Oviatt et al. 2018](#)] relates the need to protect the privacy of students while using multimodal sensing in classrooms. Chapter 13 in this book discusses that robots will not be trusted if people have no control over the information that the robots collect to interact smoothly with them and practical issues with providing such control. In addition Chapter 10 mentions the importance of considering privacy while assembling data sets. Furthermore, Chapter 13 mentions privacy as a concern about the deployment of multimodal sensors for health monitoring.

Innovations in media, in particular, have long motivated legal disputes over their effects on privacy. Warren and Brandeis wrote the first important U.S. law review article on privacy in response to the ability of new cameras to take pic-

## Glossary

**Aggregation** refers to two concepts: (i) the creation of aggregate data and (ii) the combining of information from multiple sources to enable new inferences, similar to those used in *linkage attacks*. Solove [2006] uses the second sense when referring to privacy violations that result from *information processing* that combines diffuse pieces of information together to make new inappropriate inferences possible. Given that aggregate data tends to be less privacy sensitive than microdata, the two uses of the term are almost at odds.

**Anonymization** modifies data to make it unlinkable to the people measured by the data. Often attempts at anonymization fail and *de-anonymization* is possible.

**Auxiliary information** is side information in addition to supposedly *anonymized* data used in an attempt to *de-anonymize* it. The difficulty of predicting the available auxiliary information makes preventing de-anonymization difficult.

**Contextual integrity** is a philosophical theory of privacy concentrating on how privacy norms about information sharing depend upon the social context in which the information is shared [Nissenbaum 2009].

**Cybercasing** uses technology to determine which buildings are uninhabited and, therefore, prime targets for crime.

**De-anonymization** takes data thought to be anonymized and reverses its attempted *anonymization* for the purpose of *identification*, that is, to link the data to the people measured by the data. Typically, de-anonymization uses a *linkage attack*, which compares the available information in a record for a person to various records provided as *auxiliary information*, looking for a match.

**De-identification** is a method that attempts the *anonymization* of microdata by removing identifiers from records. De-identification often fails to anonymize data because data fields that do not appear to be identifiers can often, in combination with one another and with *auxiliary information* identify people in surprising ways.

**Differential privacy** is an approach to *anonymization* and *statistical disclosure limitation*, which ensures that the released data will not depend much on any one person's data [Dwork et al. 2006, Dwork 2006]. Differential privacy does particularly well on aggregate data.

**Exclusion** is the privacy violation that exists when data subjects do not know enough about how their information will be used [Solove 2006].

**Identification**, a form of *information processing*, links information about a person to that person's identity by way of an *identifier*, and can be a privacy violation [Solove 2006].

**Identifiers** are types of data that pick out the person or entity about whom the record is concerned. Sometimes the term is restricted to data fields of a record that obviously does so, as in *De-identification*. However, subtle patterns in data can serve as unintended identifiers, making *de-anonymization* possible.

**Glossary (continued)**

**Increased accessibility** can be a privacy violation when it inappropriately makes information easier to obtain by those who should not know it [Solove 2006].

**Information collection** can be a privacy violation when it inappropriately makes observations through surveillance or seeks information through interrogation [Solove 2006].

**Information dissemination** shares information with people who might have otherwise not have had access to it, which can violate privacy, for example, through *increased accessibility* [Solove 2006].

**Information processing** converts and combines data in ways that can violate privacy, including by *aggregation* and by *identification* [Solove 2006].

**Invasions** is the privacy violation that happens when people experience inappropriate interference with the private sphere of their lives [Solove 2006].

**k-Anonymity** is a syntactic approach to *anonymization* that ensures that no one record is unique in released microdata [Samarati and Sweeney 1998a, 1998b, Sweeney 2002a]. This approach and its extensions has lost ground to *Differential privacy*, at least for aggregate data.

**Linkage attacks** are methods of *de-anonymization* that operates on *anonymized* microdata by comparing the available information in a record for a person to records in *auxiliary information* looking for a match.

**Quasi-identifiers** are attributes that can be used, often in combination, to pick out a single person from a data set. They can enable *linkage attacks*.

**Re-identification** is another term for *de-anonymization*, but emphasizes that the data was merely *de-identified* and never really *anonymized* in the first place.

**Secondary uses** of data are uses for purposes other than the one for which the data was collected, which can violate privacy [Solove 2006].

**Statistical disclosure prevention** attempts to prevent all disclosures about individuals from releasing data. In general, this goal is impossible; a motivation for *Differential privacy*. “Statistical disclosure limitation/control” sometimes is used to mean limited attempts to prevent statistical disclosures [Dalenius 1977].

tures quickly enough to capture images of unwilling subjects [Warren and Brandeis 1890]. Courts have since taken a keen interest in examining the legal privacy impacts of new technology from wire tapping [Supreme Court of the United States 1928, Supreme Court of the United States 1967], surveillance by aerial observation [Supreme Court of the United States 1986, Supreme Court of the United States

1989], tracking devices [Supreme Court of the United States 1983, Supreme Court of the United States 1984], hidden video cameras [New Hampshire Supreme Court 1964], and thermal imaging [Supreme Court of the United States 2001]. Philosophers have argued that privacy is important for dignity [Bloustein 1964], enabling intimate relations [Fried 1970, Gerety 1977, Gerstein 1978, Cohen 2002], interpersonal relations in general [Rachels 1975], conforming to contextually appropriate norms [Nissenbaum 2009], or for controlling access to the person [Gavison 1980]. While the multitude of justifications might make privacy appear to be a confused or even a non-existent concept, Mulligan et al. [2016] argue that privacy is rather an essentially contested concept, a concept whose meaning is and should be up for debate, not because it is confused, but because the debate is productive in clarifying what our values are.

The contest over the meaning of privacy has produced numerous definitions of it. Merriam-Webster [2013] defines privacy as “a) the quality or state of being apart from company or observation and b) freedom from unauthorized intrusion.” Given a wide range of philosophical justifications, legal scholars have generally viewed privacy as a collection of related rights rather than a single concept. This view started at least with Prosser [1960], who summarizes that “privacy is the right to be left alone” but also viewed it as a multifaceted concept. More recently, U.S. President Barack Obama’s “Framework for Protecting Privacy” defines privacy in terms of multiple related rights [The White House 2012].

Perhaps the most extensive exploration of the many facets of privacy is Solove’s [2006] taxonomy of possible privacy violations. He collects these related violations into four groups: invasions, information collection, information processing, and information dissemination. We will summarize these groups, focusing on violations related to multimedia.

*invasions* represent interference in what is traditionally considered the private sphere of life, including intrusions on private property and interfering with decisions. *information collection* includes making observations through surveillance and seeking information through interrogation.

Four types of *information processing* seem particularly relevant to large-scale multimedia analysis. *aggregation* combines diffuse pieces of information together to make new inferences possible. Multimedia analysis can take small pieces of information spread across numerous recordings and combine them in ways that provide new insights. *identification* links information about a person by way of an identifier. Multimedia analysis, including face and speaker recognition, can link someone from one recording to the next. Large-scale multimedia analysis can enable *secondary uses* of data, that is, uses of data for purposes for which it was

not originally collected. The possibility of an open-ended set of secondary uses highlights the importance of avoiding *exclusion*, the condition that exists when data subjects do not know how their information will be used.

Last, the *information dissemination* can harm privacy by sharing information in inappropriate ways, such as ways that violate trust or paints an inaccurate picture of someone. Of the dissemination risks Solove considers, *increased accessibility* is the one most relevant to multimedia analysis. Advances in image and audio search make it easier to access information about a person, possibly encouraging unintended uses of that information.

The computer security community also lacks a consensus on the definition of privacy. Privacy research is viewed as different from the field of secure communication, which is “securing the properties of the communication itself” [Wikipedia 2013] through methods of cryptography, steganography, identity hiding, and other well-known computer science topics. Privacy research is not about securing a communication line between several parties, but rather tends to cover controlling what people learn about one another even if they choose to communicate with one another.

For example, multimedia features such as geotagging, which adds meta-data about where a multimedia recording was made to the recording, raises privacy concerns because they can result in people conveying more information than intended while voluntarily sharing content. More generally, an important goal for multimedia privacy is practically limiting the implications of communication to ensure shared multimedia conveys only the information the author intended. We acknowledge that this goal will most likely never be achieved perfectly. However, improvements in methods can make communication more private. In the rest of this chapter we focus on this privacy goal, particularly to cases that pose an actual criminal threat and/or directly influence life-changing decisions.

A variety of formal definitions of this sort of privacy exist in computer science. The most straightforward, *statistical disclosure prevention*, is based upon the idea that an adversary with the released information learns nothing about any one person that the adversary could not have known without the information [Dalenius 1977]. However, if the released information is the exact value of a useful statistic (where “useful” is measured in terms of a utility function), it is impossible for this requirement to hold for all adversaries [Dwork 2006, Dwork and Naor 2008].

*Differential privacy* takes a different approach. Rather than attempting to limit the inferences made possible by releasing some data, it places restrictions on how much any one person’s data may affect the to-be-released statistic [Dwork

[et al. 2006](#), [Dwork 2006](#)]. While this approach promises less, it is a promise that algorithms can keep, making it a leading approach to privacy in computer science. (See Focus Question [16.10](#) for more about differential privacy.)

[Halpern and O'Neill \[2008\]](#) formalize privacy relevant concepts such as secrecy and anonymity using logics of knowledge. In response to Gavison's [[1980](#)] desire for "protection from being brought to the attention of others", [[Chawla et al. 2005](#)] formalize a notion of an individual's record being conspicuously different from the other records in a set; they characterize this notion in terms of high-dimensional spaces over the real numbers.

## 16.3

### Prior Work on Privacy Threats and Responses

The scientific community has extensively investigated privacy threats and mitigations for structured data—data organized with attribute fields. After reviewing this work, we consider the less developed literature on privacy for multimedia data, much of which is *unstructured*.

#### 16.3.1 Some Notable Privacy Threats on Structured Data

Solove's concepts of aggregation, linkage, and increasing accessibility relate to many of the privacy threats explored in prior research on structured data. In the context of databases and [anonymization](#), researchers have studied [linkage attacks](#). In its simplest form, an attacker compares supposedly anonymized records to records with [identifiers](#) in hopes of finding matches that enable [de-anonymizing](#) the "anonymized" records. For example, suppose an attacker receives a [de-identified](#) medical database that contains detailed information about each person's measurements (height, weight, etc.). If that attacker also gains access to a tailor's records with names, the attacker might be able to link the names back to the medical records using the measurements as a key.

[Sweeney \[1997\]](#)] showed that anonymously published medical records can be de-anonymized when correlated with external data, triggering a large body of follow-up work on designing anonymous statistical databases as well as understanding their limitations [[Dwork et al. 2006](#), [Dwork 2006](#), [Dwork 2008](#), [Dwork and Naor 2008](#), [Sweeney 2002b](#), [Dinur and Nissim 2003](#), [Aggarwal 2005](#)].

[Narayanan and Shmatikov \[2008\]](#) present an algorithm for de-anonymizing sparse data sets. They apply their algorithm to anonymized Netflix movie ratings: given knowledge of a subset of movies a person has rated (e.g., learned from a lunch conversation or public ratings), then their system is able to identify *all* the movies in the database that the user has rated. In later work [[Narayanan and Shmatikov](#)

2009], they use the same idea to de-anonymize a social network graph by leveraging a graph from a second network with real identities as auxiliary information.

Researchers from Parc investigated inference using web search engines to analyze whether anonymized (or obfuscated) private documents that are going to be released publicly can be de-anonymized [Staddon et al. 2007, Chow et al. 2008]. They do not consider multimedia content nor inference between information that is already publicly available.

Griffith and Jakobsson [2005] correlate public birth, death, and marriage records from the state of Texas to derive the mother's maiden name of more than 4 million Texans. Baldazzi et al. [2010] automatically query 8 social networks with a list of 10 million e-mail addresses to retrieve the associated user profiles. They then correlate that profile information across the networks and are able to identify mismatches between them (i.e., they find users who chose different names, age, etc. in different networks). More generally, Bishop et al. [2010] discuss the need to go beyond "closed worlds" when sanitizing a data set and consider external knowledge explicitly.

With geo-location information being a popular key to image and video retrieval, another area of related research is locational privacy. The Electronic Frontier Foundation published an overview of locational privacy [Blumberg and Eckersley 2009]. Locational privacy in vehicular systems, such as toll collection, is addressed in Popa et al. [2009] and Hoh et al. [2008]. Zhong et al. [2007] present protocols for secure privacy-preserving location sharing. The HTML 5 standard will include APIs to query a client's location. The *Cree.py* [Walker-Morgan 2011] application uses geo-location data from social networks and media hosting services to track a person's movements.

### 16.3.2 Some Notable Privacy Approaches on Structured Data

Despite the impossibility of statistical disclosure prevention in general, a variety of attempts have been made to come up with anonymization approaches for providing such a guarantee or for avoiding the attacks mentioned in Section 16.3.1. In the statistics community, such work is typically called "statistical disclosure limitation" and attempts to preserve privacy despite releasing statistics. (For an overview, see Federal Committee on Statistical Methodology [2005].) Methods include those for releasing tables of aggregate data (summary statistics about groups of people) or microdata (individual responses). Despite appearing to work well in practice, these works tend to lack strong formal properties. We cover a few approaches studied in computer science.

### 16.3.2.1 ***k-Anonymity***

Within computer science, a series of approaches, starting with ***k-Anonymity***, place syntactic requirements on the anonymized database [[Samarati and Sweeney 1998a](#), [Samarati and Sweeney 1998b](#), [Sweeney 2002a](#), [Machanavajjhala et al. 2007](#), [Li et al. 2017](#), [Xiao and Tao 2007](#)]. However, this line of work has not yet produced a robust, compositional syntactic approach to privacy. (See Focus Question 16.10 for more about why.) Nevertheless, we'll discuss it in detail since it remains one the most well-known privacy definitions.

*k*-anonymity is defined over databases, which are modeled as matrices. The rows correspond to data about different people and the columns correspond to different attributes about those people. The following table is an example of a small database, which we will call  $D_0$ :

Name	Age	Sex	Weight (kg)	Cancer
Sue Jones	34	Female	50	Lung
Kate Gross	37	Female	55	Lung
Bob Smith	31	Male	65	Throat
Guy Fox	39	Male	60	Lung
Tom Om	50	Male	70	Throat
Bill Lee	55	Male	75	Larynx

Some of the attributes (column headers) intuitively identify the person the row is about, for example, the name “attribute.” Other attributes are intuitively private, such as the type of cancer a patient has. Some attributes fall somewhere in between. For example, age and sex might, in some cases, be sufficient to identify a person in a small database. These attributes are called ***quasi-identifiers***, but henceforth, we'll abuse nomenclature and refer to both identifiers and quasi-identifiers as quasi-identifiers.

The first step in *k*-anonymity is to determine the quasi-identifiers found in the table. Let  $Q$  represent the set of quasi-identifiers. For a database  $D$ , let  $D[Q]$  be the database that results from projecting out just those attributes in  $Q$  and dropping the rest. For example, letting  $Q_0 = \{\text{Name, Age, Sex, Weight}\}$ ,  $D_0[Q_0]$  produces the following table.

Name	Age	Sex	Weight (kg)
Sue Jones	34	Female	50
Kate Gross	37	Female	55
Bob Smith	31	Male	65
Guy Fox	39	Male	60
Tom Om	50	Male	70
Bill Lee	55	Male	75

A database  $D$  satisfies  $k$ -anonymity for the quasi-identifiers  $Q$  iff every row that appears in  $D[Q]$  shows up at least  $k$  times (i.e., is repeated  $k - 1$  times) [Samarati and Sweeney 1998a, Samarati and Sweeney 1998b]. For example, the database  $D_0$  only has 1-anonymity for  $Q_0$ . A combination of suppressing attributes values (represented as  $-$ ) and generalizing them to ranges of values can produce the following database  $D'_0$  with 2-anonymity for  $Q_0$ .

Name	Age	Sex	Weight (kg)	Cancer
—	30s	Female	50s	Lung
—	30s	Female	50s	Lung
—	30s	Male	60s	Throat
—	30s	Male	60s	Lung
—	50s	Male	70s	Throat
—	50s	Male	70s	Larynx

In this table, for example, “30s” refers to the range of ages from 30–39.

While higher values of  $k$  would provide even more privacy, in principle even 2-anonymity provides a degree of privacy protection. The intuition is that an adversary trying to learn the sensitive value of “Cancer” about a person  $p$  in the database would want to first determine which row holds the data about  $p$ . The adversary might already know the values of the quasi-identifiers for  $p$ . However, since the table has 2-anonymity, the adversary will not be able to determine which of two rows corresponds to  $p$  using just the quasi-identifiers.

Do you buy the above reasoning? How much privacy does  $k$ -anonymity provide? Does it capture privacy? Can it leak information in certain cases? Before going on, think about these issues yourself.

As stated,  $k$ -anonymity has two issues when it comes to preserving privacy. The first is that if all  $k$  people with the same values for the quasi-identifiers have the same value for a sensitive attribute, then the released data will reveal that they

each have that value for that sensitive attribute [Machanavajjhala et al. 2007]. The second is that releasing two databases, each with  $k$ -anonymity, could provide enough information for the receiver of those databases to construct one without  $k$ -anonymity [Ganta et al. 2008]. (We explore these issues in more detail in Focus Question 16.10.) For these reasons, some researchers have looked for definitions that provide more robust guarantees than  $k$ -anonymity does.

#### **16.3.2.2 Differential Privacy**

*Differential privacy* provides such a robust guarantee, albeit one for a more limited notion of privacy than statistical disclosure prevention. Like  $k$ -anonymity, differential privacy is defined over databases [Dwork et al. 2006, Dwork 2006]. However, rather than being a property of a database, differential privacy is a property of an algorithm that produces statistics about databases. While such statistics could resemble a database itself, we will focus on simpler statistics, such as the average value of some column in the database. We will model such algorithms as a function  $a$  that given a database  $D$  returns a random variable  $a(D)$  over possible output values. A single output value  $o$  is drawn according to this random variable  $a(D)$  and shared as the privacy-preserving statistic about the database. We let  $\mathcal{O}$  be the set of all possible output values. For example,  $a(D)$  might count up the number of people with lung cancer in  $D$  and release that count plus some random noise added to protect privacy.

A randomized algorithm  $a$  has  $\epsilon$ -differential privacy iff all databases  $D_1$  and  $D_2$  that differ in a single row, for all measurable subsets  $S$  of  $\mathcal{O}$ ,

$$\Pr[a(D_1) \in S] \leq e^\epsilon * \Pr[a(D_2) \in S].$$

When  $\mathcal{O}$  is discrete, as it is for any computer producing a finite range of outputs, this  $\epsilon$ -differential privacy can be simplified to requiring that all databases  $D_1$  and  $D_2$  that differ in a single row, for all outputs  $o$  in  $\mathcal{O}$ ,

$$\Pr[a(D_1) = o] \leq e^\epsilon * \Pr[a(D_2) = o].$$

Intuitively, one can picture an adversary seeing an output  $o$  drawn from  $a(D_1)$  or  $a(D_2)$  and trying to figure whether it came from  $a(D_1)$  or  $a(D_2)$ . If the adversary learns this, the adversary will learn whether the database is  $D_1$  or  $D_2$ , which tells the adversary whether the person's whose data is in only one of them is in the database. Differential privacy means that the adversary learns little about whether the database was  $D_1$  or  $D_2$  since the probability that the output was  $o$  is almost equal (up to a factor of  $e^\epsilon$ ) for the two databases.

Algorithms exist that can release versions of aggregate statistics with a small amount of random noise added to provide both  $\epsilon$ -differential privacy and a reasonably high level of accuracy. (For an overview, see [Dwork and Roth \[2014\]](#). For a framework for implementing or using such algorithms, see [McSherry \[2009\]](#).) The person running such algorithms gets to choose the value of  $\epsilon$ , where a lower value provides more privacy, adds more noise, and results in less accurate statistics than would a higher value. If more accuracy is desired, the value of  $\epsilon$  can be increased, which decreases privacy, or a larger sample can be used, which leaves the level of privacy the same but could increase the costs of a study.

Differential privacy behaves more predictably under composition with itself than  $k$ -anonymity does with itself. In particular, if two algorithms each have  $\epsilon$ -differential privacy, then their combination has  $2\epsilon$ -differential privacy. Other sorts of compositions are also well behaved [[McSherry 2009](#)].

However, this does not mean that differential privacy satisfies everyone's desires for a privacy definition. In particular, in some cases, arguably, someone's privacy may be violated by releasing an approximately correct aggregate statistic about a group of people. This violation could be caused by some adversary who knows something about a particular person that depends upon the statistic, such as Bob having the same illness as the most common illness in a particular group.

Alternately, it could be caused by the whole group facing consequences due to their average condition. For example, suppose that a survey of smokers finds a higher rate of cancer than in the population at large. Health insurers may raise the premiums of all smokers as a result.

This premium increase may not seem to be much of a privacy violation since it is just science running its course. However, we can look at a related example that may sound more like a privacy invasion. To borrow an example from Kifer and Machanavajjhala [2011, Ex. 2.1], suppose a survey is looking at how common a contagious disease is within a family living under one roof. Most of time, the members of the family either all have the disease or none of them do. Thus, seeing that the count is approximately equal to the number of family members strongly suggest that each member of the family has the disease whereas seeing a count close to zero suggest that none of them do. This can allow an adversary to determine with high confidence that an individual member has the disease.

Note that in each of the three examples above, whether or not the particular individual whose privacy is in question participated in the study has little to do with the adversary learning the sensitive information. For example, consider a particular member of the family found in the example we considered last, such as the oldest family member. Whether I learn that this family member has the disease does not depend upon him joining the study. As long as enough other family mem-

bers join to produce a noisy count far from zero, I will learn with high probability that the oldest member has the disease. Given this subtlety, there has been considerable debate about the degree to which differential privacy captures privacy, with some believing it only works for certain settings [Kifer and Machanavajjhala 2011, Kifer and Machanavajjhala 2012, Li et al. 2013, Kifer and Machanavajjhala 2014, He et al. 2014, Chen et al. 2014, Zhu et al. 2015, Liu et al. 2016] and others disagreeing [Bassily et al. 2013, Kasiviswanathan and Smith 2014, McSherry 2016a, McSherry 2016b].

#### **16.3.2.3 Logics**

A variety of works attempt to make privacy properties precise using formal languages and logics [Powers and Schunter 2003, Cranor 2002, Cranor et al. 2006, Barth et al. 2006, Halpern and O'Neill 2008]. For example, Barth et al. [2006] provide a privacy enforcement framework based on *contextual integrity*, a philosophical theory of privacy concentrating on how privacy norms about information sharing depend upon the social context in which the information is shared [Nissenbaum 2009].

#### **16.3.3 Some Notable Works on Privacy of Multimodal Data**

The above section covered current work on structured data. History has shown that work on multimedia data follows in the footsteps of structured data with a delay (for example, work on compression, messaging capabilities, or even World Wide Web content itself). As a result, we see an initial growth in multimedia articles that present work on privacy.

Most closely related to this chapter are works surveying potential privacy issues. Adams and Sasse [2001] survey privacy risks that arise from differences between users' expectations and what multimedia enables. When users find their expectations violated, they can lose trust in the technology, decreasing their willingness to use it. They provide a model for assessing this risk, which looks at how sensitive the captured information is, who receives the information, how it is used, the awareness level of the users tracked by the technology, and the context of its use.

Thuraisingham [2007] focuses on security issues raised by multimedia database management systems, but also considers privacy. She discusses the need to mine multimedia databases for national security reasons but that this must be subject to the requirement of preserving people's privacy. She highlights providing statistical disclosure prevention, that is, controlling the inferences made possible from multimedia data, as a central challenge.

Ribaric et al. [2016] survey de-identification methods for multimedia content. They focus on unstructured biometric identifiers, such as face images, but also

consider structured non-biometric identifiers, such as license plate numbers. For each of a selection of such media, they survey *identification*, *De-identification*, and *re-identification* methods for de-anonymization.

We highlight a few such methods below to show the difficulties of providing privacy when sharing multimedia data. Lukas et al. [2006] propose a method for the problem of digital camera identification from images based on the sensor's pattern noise. For each camera under investigation, they first determine its reference pattern noise, which serves as a unique identification fingerprint. This is achieved by averaging the noise obtained from multiple images using a de-noising filter. To identify the camera from a given image, they consider the reference pattern noise as a spread-spectrum watermark, whose presence in the image is established by using a correlation detector. Experiments on approximately 320 images taken with consumer digital cameras are used to estimate false alarm rates and false rejection rates. This work shows the ability of artifacts to serve as unintended *identifiers*.

Many researchers have worked on automatic video blurring (e.g., Dufaux and Ebrahimi [2006], Koshimizu et al. [2006], Fan et al. [2005]). However, Neustaedter et al. [2006] showed that many of the proposed techniques are not effective. In response to this problem, Dufaux and Ebrahimi [2010] presented an initial framework to validate video privacy. This series of work shows the difficulty of removing identifiers, even those as obvious as a person's face.

Finally, some work has looked at both structured meta-data attached to unstructured multimedia and the unstructured data itself to show the real-world consequences of such data. Friedland and Sommer [2010] analyzed the privacy implications of geotagging. Specifically, the article examines the risk that such geotags pose for what was termed *cybercasing*, using online data and services to mount real-world attacks. Turning to unstructured data, follow-up work showed that geo-tags are not needed as they can be replaced by techniques that analyze multimedia for clues about the location represented by it [Friedland and Choi 2011].

## 16.4

### Privacy Risks and Possible Attacks

In this section, we describe some existing and future multimedia analytic techniques that pose a privacy risk including how these risks could be exploited. This is by no means an exhaustive list. Where we are aware of pre-existing work on such threats, we provide citations. Other risks are more hypothetical. It is very hard to distinguish between threats that are hypothetical and actual ones. Friedland and Sommer speculated about cybercasing in 2010, only to find out that in 2012 that

it had already happened in 2009. We have to assume that all hypothetical threats could become or may even already be real threats.

### 16.4.1 Risks

**Location Estimation.** Multimedia location estimation formed the genesis of our interest in privacy in multimedia, and was reported in previous work [[Friedland and Sommer 2010](#)], [[Friedland and Choi 2011](#) (discussed in Section 16.3.3)]. Using multimodal methods, state-of-the-art algorithms can estimate the location of about 40% of Flickr videos with an accuracy better than 100 m, and over 50% with an accuracy better than 1 km [[Choi et al. 2010](#)]. This extends the amount of exactly trackable multimedia by a significant factor without requiring actual GPS sensors.

**Time Estimation.** The date and time that a multimedia document was recorded can be estimated using cues such as sun location or measuring shadow lengths. More powerfully, if you can determine that Video A was recorded at the same time and place as Video B, and you know or can infer Video A's time, you now know Video B's time. Just excluding time/date metadata from *your* vacation video does not protect you if somebody else includes it in theirs.

**Person Detection.** In the image realm, this is usually known as face detection; in audio, speaker recognition. While the uploader can take active methods to anonymize the foreground participants if privacy is an issue (e.g., replacing their face with a black box, replacing their audio with a bleep sound), the privacy of background participants is problematic because the uploader may not care about incidental privacy breaches of the background participants.

**Object Detection.** Detecting an iPhone in a person's hand might make them a more desirable robbery target. Marketers could target people based on the furniture quality in the background of a video. Note that mitigation techniques are particularly problematic with object detection, since one cannot simply remove *all* objects from a multimedia document without severely impacting the document's content.

**Environmental Acoustic Noise.** Uploaders often recognize the need to obscure faces. However, when recording video data they often forget that the audio track includes a unique signature that might break their anonymity. This has been shown in several studies, including our previous work [[Friedland and Choi 2011](#)]. Also, the combination of such linking methods with other methods such as location estimation leads to even more powerful privacy invading possibilities.

**Sensor Detection.** It is already possible to narrow down or even uniquely identify what camera was used to record a video or what microphone was used to record audio based on the artifacts of the sensor [Lukas et al. 2006]. For example, pixel noise is unique to a particular camera; the exact frequency response of a microphone might be used to narrow down the possible microphones. This provides a whole new avenue of linking, completely bypassing other means of anonymization.

**3D Recordings.** Time-of-flight cameras, light field cameras, stereo cameras, and microphone arrays are all becoming more pervasive. It is clear that similar devices will continue to be developed. Each comes with its own sets of issues, and have the potential to capture even more unwanted data. This is especially true since the idea for 3D cameras is to not focus on anything but to capture the entire scene and allow for focusing on specific objects in a post processing step. Since this trend will only accelerate, it is necessary for the multimedia community to address these issues.

**Exotic Sensors.** Everything from air pressure sensors to heart rate monitors are becoming more common, and it is likely data from these sensors will be incorporated into multimedia documents much as GPS is now. Since users often have no real notion of what is being collected or how accurate it is, they have little or no intuition on the privacy implications. A prominent historic example is GPS: it was only recently that the profound privacy implications of geotagging became commonly known.

#### 16.4.2 Attacks

We outline a small number of specific attacks that can now or could shortly be used to invade privacy in detrimental ways using Internet-scale multimedia analytics and linking.

Today, one can readily access much of the structured information available online via programmatic interfaces: major services like Google, Facebook, Twitter, Flickr, YouTube, and LinkedIn all offer extensive APIs that make automatic retrieval trivial. These APIs often offer more comprehensive access than the corresponding web interface, and their availability is the primary driver behind the wide range of third party “apps” that constitute a key part of today’s social networking space.

As multimedia retrieval technology matures, it might eventually become part of such APIs, making the capabilities available to everybody able to write a few lines of Python code. For example, Google already provides simple forms of image and video search, and rumor has it that face recognition is ready for mass deployment as part of their Google Goggles service. Facebook has already integrated face recognition

into their platform, and though it is not yet exposed via the Facebook API, third-party companies such as face.com are already providing programmable access to face recognition of Facebook content.

Having large-scale multimedia retrieval at one's fingertips provides an opportunity for amazing next-generation online services. However, we believe that it will also open up a new dimension of privacy threats that our community has not yet understood.

The availability of Internet-scale multimedia retrieval capabilities allows a wide range of attacks that threaten users' privacy. Whereas today's search queries remain limited to mostly textual information, attackers will eventually query for audio and video content. Criminals could leverage that to reliably locate promising targets. For example, they may first identify individuals owning high-value goods within a target area and then pinpoint times when their victims' homes are unattended.

Another threat is background checks becoming much more invasive than today. Many companies have strong incentives to examine their customers' private life for specifics impacting business decisions. An insurer, for example, might refuse payment to a customer receiving disability if the insurer finds Facebook photos of the customer skiing. Likewise, an employer seeking new hires might check a candidate's Twitter followers for potentially embarrassing information that could be used against the company in the future and refuse to hire such candidates.

A whole new realm of marketing techniques are enabled by multimedia retrieval and linking. A company could extract all videos of people wearing branded merchandise, cluster them by location and time, and target that location for direct marketing. The privacy implications of such broad and automatic analysis have been insufficiently studied.

The new capabilities make stalking easier by providing the means to not only quickly locate the victims, but also profile their typical behavior patterns, friends, relatives, and acquaintances.

## **16.5 Case Studies of Privacy Violations Using Multimedia Analyses**

### **16.5.1 Finding Corruptibles**

In this section, we will exemplify the power of multimedia retrieval in combination with structured-data retrieval with a mockup scenario adopted from [Friedland et al. \[2011\]](#).

Consider the following business: Fred works for Schooner Holdings and wishes to gain (possibly illicit) inside information on future profits at the chip maker Letin.

Fred hires Eve, who runs an “expert network.” Eve puts Fred in touch with Bob, a Letin employee. In the process of consulting for Fred, Bob is encouraged to reveal information about Letin’s upcoming products.<sup>1</sup>

Currently, the greatest limit on this process is Eve finding experts like Bob who (perhaps unknowingly) possess potential insider information and are willing to act as consultants. Eve would greatly improve her business if she could find “corruptibles”: individuals in the business of interest who might be favorable to legitimate or illegitimate offers.

Thus, Eve starts searching social networks for individuals who are compatible with her desired level of (il)legality. She instructs her crawler to begin with LinkedIn and web searches, crawling the names and contact information for personnel at companies of interest. Then, her crawler shifts to Facebook, Twitter, other social networks, and blogs, beginning with all candidates found in the first pass. This crawler does not just look at the candidates but also at friends of candidates.

She also searches any media, including images and videos, for links to other people that the social network might not provide directly. Face recognition for example can provide probable connections to other profiles. She also examines media for any compromising material, such as illegal acts, drug paraphernalia, or party photos. Eve knows that her automated content analysis does not need to be perfect: she leverages crowd-sourcing services such as Mechanical Turk<sup>2</sup> to validate potential candidate matches using human labor at a very low cost.

Eve’s crawler also queries further public and semi-public records. There are commercial services that map an email address to a mailing address. Her crawler uses these to discover where candidates live and how much their property is worth (e.g., by using Zillow.com’s access to property tax data and sales history).

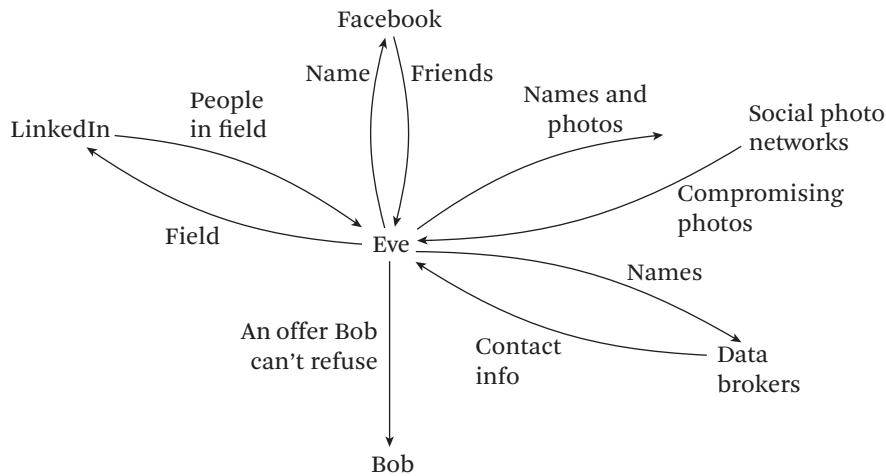
With all this data, Eve’s crawler can now create “inference chains” which estimate the probability that any given candidate in her set has a potential weakness, enabling Eve to search for possible points of corruptibility (see Figure 16.1). An individual who is dating someone with a reputation as a gold digger, or who purchased their house at the height of the real estate bubble, might have financial problems. Such candidates could be corrupted by offering consulting positions, allowing Eve to expand her expert network.

Eve might also contract with those operating outside the law. Then blackmail becomes an attractive option, especially if considering guilt by association. Some-

---

1. In many countries, this practice is possibly illegal but exists in a gray area and is seemingly routine practice. The Galleon insider trading trial [Cohan 2009] was based largely on the use of expert network consultants.

2. <https://www.mturk.com/mturk/welcome>



**Figure 16.1** Example Inference Chain. Eve uses information to get more and information and puts it all together to pressure Bob into doing her bidding.

one with a security clearance may be vulnerable if his associates are drug abusers, or if he is having an affair that can be inferred through social patterns.

Nothing in the preceding scenario is unrealistic: every step Eve takes can be constructed using today's technology. It is simply a matter of putting all the pieces together to collect and analyze the reams of data which exist on today's social networks and other databases.

Unfortunately, there is also hardly any protection in place against somebody like Eve. Furthermore, while structured data still plays a dominant role in this scenario, it is easy to see how multimedia data will blur the boundaries even more. For example, if we assume that face recognition technology reaches close to perfection, user names will no longer provide a boundary as long as a face photo is part of the website. Moreover, speaker recognition, location estimation, and other techniques described in Sections 16.3.3 and 16.4.1 will add even more possibilities. Finally, note that the methods need not be perfect—Eve needs only a small number of likely hits to follow up on to allow nefarious actions to proceed.

### 16.5.2 Cybercasing from Data Brokers

In a previous study [Friedland and Sommer 2010], Friedland, an author of this chapter, and Sommer defined cybercasing as crimes enabled by the sharing and eased discovery of online information. A simple example is shown in Figure 16.2. A bike advertised for sale on a website can be located by the GPS coordinates embedded



**Figure 16.2** Cybercasing potential based on website advertising (theft). The left photo shows a photo of a bike advertised for sale. The photo was taken with an iPhone, which stored the geo-location at which the photo was taken. The right photo shows the pin from Google Street View placed at almost the exact point where the photo was taken based on the information involuntarily stored in the image by default.

in the image, potentially allowing a thief to cherry pick desirable models online. Friedland and Sommer also pointed out that burglary is enabled by cybercasing. They searched for vacation videos with geotags and then discovered home videos from the same account with geotags. An example is shown in Figure 16.3. Friedland later set out to determine whether he could reproduce the success of such cybercasing without utilizing geo information. With Choi, he succeeded in doing so by inferring geo-locations using multimedia analyses [Friedland and Choi 2011]. Now, in an ongoing study with Weaver, Friedland is attempting to replicate the results without even needing complex multimedia analyses [Choi et al. 2018]. Rather, the goal is to buy location information from data brokers, companies whose business model is selling data about people.

In this new study, they began by searching for videos based on the vacation topic (list) and extracting the Google username. To gain data about named individuals, they submitted, to a data broker, append requests, which ask the data broker to take a partial record about a person (just the username in this case) and append to it any additional data the broker has on the person that the partial record describes. After excluding obviously bogus usernames, they submitted 2,824 names in such requests. We do not name the data broker used since the contract with it appears to prohibit disclosing its identity. The overall cost was a \$500 setup fee plus an additional \$0.10 for each match successfully appended.



**Figure 16.3** Cybercasing potential based on video uploading (burglary). This video was uploaded while the family was on vacation, clearly indicating the location and the fact that vacation just started (note the “First Day” in the photo’s caption). Videos from the same user have geo-location information indicating the family’s residence.

The result was surprisingly negative: out of the 2 824 addresses submitted, only 9 were successfully appended. We believe this is due to three factors: a lack of connection between Google accounts and purchasing behavior, that the data broker used focused on fulfilling list requests (requests for a list of people fitting some criteria) instead of append requests, and the relative quality of this particular data broker. We consider each below.

First, if a user doesn’t utilize their Google account for making purchases, there will be no link between the mailing address and email address available to a data broker to sell. Google itself may have information about the user’s address, but Google has no incentive to sell this information to others as it represents a competitive advantage.

Second, list-focused data brokers do not prioritize complete coverage as highly as append-focused data brokers, such as credit reporting agencies. In credit reporting, a small number of brokers strive for complete coverage. If a credit agency only had information on 50% of the population, it would not be competitive in the marketplace since the data consumers select the queried names. A list-centric data broker’s incentives are different: they do not need complete coverage, rather they need quality in the data they have since the broker gets to select its best data matching the requested criteria to share with the data consumer. This broker in

particular focuses as a reseller of lists with a wide variety of topics, including religious affiliation (Catholic, Jewish, Islamic, etc), economic profile (credit score), ethnicity, political donation habits, holders of handgun concealed carry permits, and even such esoteric lists as “boat owners in LaGuna Niguel, California.”

Third, they selected this broker mostly due to setup cost. Most data brokers are only interested in large orders. Even this broker required a \$500 setup cost for the append query, and this may represent a case of “you get what you pay for.”

They also did some spot checking on results, and found that append data may be of marginal quality. For example, although it correctly identified one author’s father’s address and one cousin, the address for another cousin and the author himself were completely wrong: not even in the correct state.

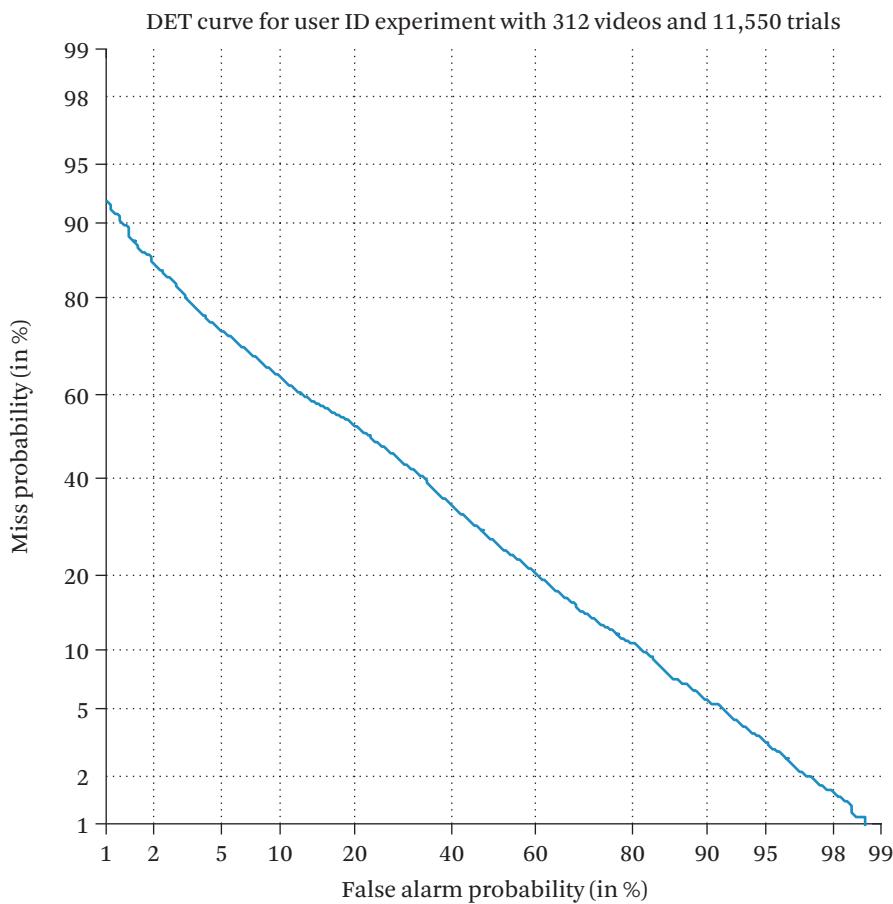
Nevertheless, devices in the Internet of Things are frequently cloud based and are therefore recording data and sending it to the provider of the service. It is no secret that companies like Salesforce integrate cloud-based services like IFTTT ([ifttt.com](http://ifttt.com)), a system that coordinates across devices, with customer relationship management systems (CRMs). With these abilities, such companies can already act as data broker for high-volume customers. As brokers gain more data through such means, they become a more useful source of location data for cybercasing.

### 16.5.3 Matching Anonymous Social Media Accounts

The article by [Lei et al. \[2011\]](#) presents the task of linking personas based on modeling the audio tracks of random Flickr videos. The experiment they describe aims to answer the question “Do these two videos belong to the same Flickr user?” The experiment is modeled after speaker recognition experiments, where two audio recordings are analyzed for a possible match of the speaker. While using only the audio tracks for persona linking ignores other cues, such as the video tracks, the results are impressive and frightening. The article shows that linking the uploaders of videos based on the audio track of the videos is feasible.

Using a subset of the MediaEval Placing Task’s Flickr video set, which is labeled with the uploader’s name, they conducted an experiment with a similar setup as a typical NIST speaker recognition evaluation run. Based on the assumption that the audio might be matched in various ways (speaker, channel, environmental noise, etc.), they trained a speaker recognition system on the audio tracks of the Flickr videos. Note that since the selection of videos is essentially random, the audio track can contain any sound.

They obtain an equal error rate of 36.7% on 312 videos with 11,550 trials. When the false alarm rate is reduced to 1% in favor of misses, the accuracy is at 7.9%. In



**Figure 16.4** False alarm and miss rates for deanonymizing videos based on audio tracks, as described in Section 16.5.3.

other words, a matched video is about eight times as likely to be a match than a false alarm. The result is shown in Figure 16.4. False negatives are highly tolerable since matching a user across an anonymized account only requires one true positive match. As the number of videos in the Internet grows steadily, both in general and for a targeted anonymized account in particular, the number of chances for that match increases.

Furthermore, this study shows that even highly tuned systems, like current speaker recognition systems, are generic enough to be “abused” for a different

tasks, such as violating privacy. In future work, the tuning of the speaker recognition system to this specific task would likely improve the accuracy. More importantly, we and the authors expect that combining audio with other cues, such as text or video features, will improve results dramatically. Personas have been linked before using textual information as well as visual information.

A second study [Goga et al. 2013] links accounts across social networking sites as having the same owner. Examining on Yelp, Flickr, and Twitter, they use only innocuous features that come inherently with posted content: the geo-location attached to a user's posts, the time stamp of posts, and the user's writing style as captured by language models. They show that among these three features the location of posts is the most powerful feature for identifying accounts that belong to the same user in different sites. When they combine all three features, the accuracy of identifying Twitter accounts that belong to a set of Flickr users is comparable to that of existing attacks that exploit usernames. Their attack can identify 37% more accounts than using usernames when they correlate Yelp and Twitter.

These two studies have significant privacy implications as they present two sets of evidence that just using different user names to keep sets of posts unlinked will not work in the future: the posts themselves can provide enough information to link accounts. Such attacks exploit users' tendency to assume that, if they maintain different personas with different names, the accounts cannot be linked together. Unexpected results could include law enforcement matching surveillance videos of a suspect against a public video database in an attempt to identify the perpetrator.

#### 16.5.4 Discussion

In summary, it is the aggregate set of a user's complete online footprint that needs protection, not just content on individual sites. Furthermore, it is hard to defend against attacks where the information that enables them comes intrinsically with the very activity one wants to publish. While all the work above examines a specific set of websites and correlation techniques as case studies, it demonstrates the broader potential, and risk, of cross-site correlation. The approaches remain conceptually simple, yet we expect that soon more sophisticated variants will emerge for exploiting the increasing volume of seemingly innocuous user information that websites now offer via convenient APIs. In particular, and as explained before, we anticipate that automated content analysis technology—such as face recognizers and natural language processing—in combination with inference will enable correlations more powerful than what is described here.

# 16.6

## Future Directions For Research

Countering the attacks described above is not straightforward since filtering out sensitive information from audio and video content is fundamentally harder than with structured text data. We therefore propose a new topic in multimedia devoted to considering both privacy research as well as education.

### 16.6.1 Mitigation Research

A major challenge for conserving privacy in consumer produced videos is the development of methods to identify the foreground information that the user considers important from the background information. It is this background data that has the highest risk of incidentally leaking private information.

We believe that machine learning will play a key role in detecting such unnoticed information leaks. For example, one can label who is an “extra” in a movie by the number of times they appear and the number of lines they speak. The extras form the semantic background to the movie—they are noticeable, but not directly relevant. A machine learning algorithm could use “star” vs. “extra” as ground truth, and learn models to distinguish the two. Applied to consumer-produced videos, the system could then identify foreground vs. background participants using the trained model.

Once the information that is breaking privacy is identified, it must also be removed or distorted sufficiently to reduce the threat. This is difficult with most existing multimedia analysis algorithms, since they are statistical in nature. If we understood the specific cues the statistical methods learn, we could obscure those cues, hopefully without distorting the rest of the content. For example, if the background semantic “bird call of a Nene” is detected, you are leaking location information (Hawaii). Just damping that sound may be enough to obscure the location. This sort of cue detection is in the nascent stages for some methods (e.g., concept detection as in TrecVID MED [[National Institute of Standards and Technology 2016](#)]), and nearly non-existent for others. It is incumbent on the multimedia community to develop an understanding of the cues so that mitigation techniques can be developed.

Previously, issues had been raised about automatic face detection algorithms compromising privacy [[Simonite 2014](#)], and as discussed earlier also geo-location estimation algorithms compromising privacy. Geo-location estimation is potentially more pernicious, since users are likely unaware of the conditions under which a typical geo-location algorithm can automatically predict the locations of their images. However, the situation is not hopeless; rather, it requires systematic research

to provide valuable insights that can be provided by research in the area of visual information retrieval. For example, there is a chance that filters and other simple convolutional operation can geo-cloak an image. These filters may be part of the standard practice already. Also, research dedicated, for example, to techniques that approximate a background collection, could take us a step closer to fitting a collection in memory, and thereby a step closer to running a geo-privacy aware camera from a mobile phone. We envision such an application would also recommend to the user filters that add Instagram style enhancements, but that are specially chosen to protect geo-privacy. The key challenge of the geo-privacy aware camera is that it should fit seamlessly into currently widespread photographic practices, that is, photo-taking behaviors and the use of image enhancements.

For other methods, more direct mitigation may be possible. Google has already released a tool for blurring all the faces in YouTube videos [Conway 2012, Halliday 2012]. A more advanced upload tool could blur just the semantically background faces in a video (however, this might not be enough; see also the discussion in Section 16.3). A query tool could refuse to perform speech recognition and indexing on background voices. This would be very similar to today's common practice of copy machines refusing to copy bank notes. A key component of such a system would be to ensure, possibly with the interaction of the uploader, that foreground content is not compromised.

### 16.6.2 Education on Privacy

Independent of any technological protection, we believe a key ingredient to comprehensive mitigation must be education. University computer science curricula usually include an abundance of material on how to improve retrieval based on the underlying multimedia content analysis but only rarely talk about the negative impacts of these technologies. Privacy content is mostly limited to traditional topics in secure communications such as steganography, encryption, and other well-known techniques and/or even removed from consideration, when ethical concerns are considered not to be part of engineering. Therefore, even when acknowledged as a problem, many new technologists lack the knowledge of how to react to society's concerns and even how to mitigate easy-to-address risks. An argument often heard from students is: "We'll deal with privacy and social issues later—right now we need to focus on development." The truth, however, is that, if privacy and security had been a concern in the early stages of developing the Internet, many of today's issues, such as spam and phishing email, would most likely be much less of a problem. Undergraduate and graduate engineering education curricula should

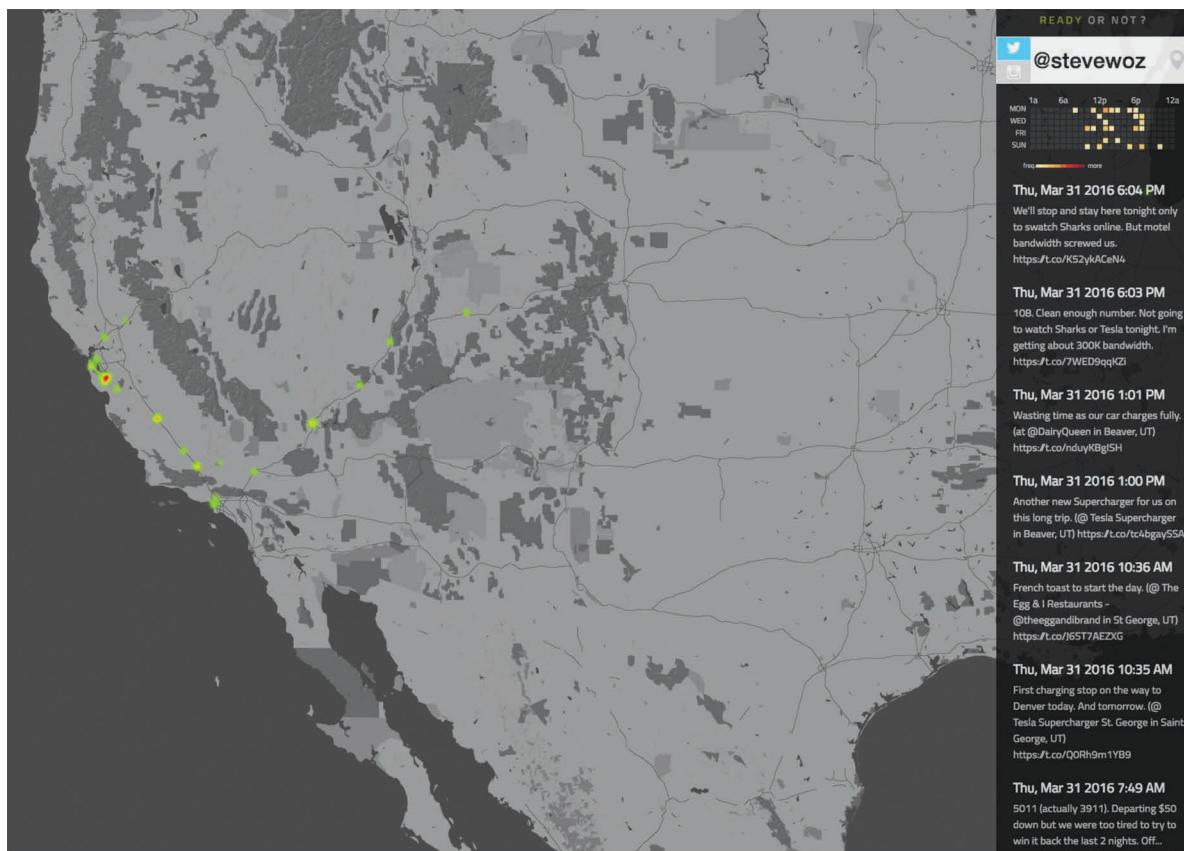
therefore include a strong component on privacy that makes future technologists aware of the societal implications of their research and development.

The second line of education should concern users, especially young people. Among the groups most affected by privacy concerns are high-school students [[InsideFacebook 2010](#)]. They are the most frequent users of social-networking sites and apps, but often do not have a full understanding of the potential consequences their current online activities might have later in their lives. For example, a Facebook posting that a high schooler's friends think is "cool" might be seen by a much larger audience than she or he expected, perhaps including future employers who won't agree with the high schooler's judgement. In addition, not understanding—or not thinking about—the consequences of posting often leads to oversharing information about other people, including friends and relatives. Consequently, users can take steps to protect themselves once they realize the power that modern content analysis tools yield in the hands of adversaries. They might even choose not to post certain content in the first place.

The Teaching Privacy Project (TPP) has the aim to develop a set of learning tools to help educators demonstrate what happens to personal information online, and the effects of sharing it [[International Computer Science Institute 2016](#)]. For example, Figure 16.5 shows a teaching tool that we created as part of the Teaching Privacy project for social media privacy education for teenagers. The input for the web-based tool is an arbitrary Twitter or Instagram handle. The feed is then analyzed for timestamps and geo-tagging data. If found, the feed is displayed along with a heat map of the recent geotags. Also, the tool shows a calendar with frequent times. This way, a user can see the habits and typical patterns exposed by the feed. This example shows the output for "Steven Wozniak," the co-founder of Apple Computers. Figure 16.6 shows a famous image illustrating the amount of metadata stored in a tweet which is significantly more than even exploited in the education tool.

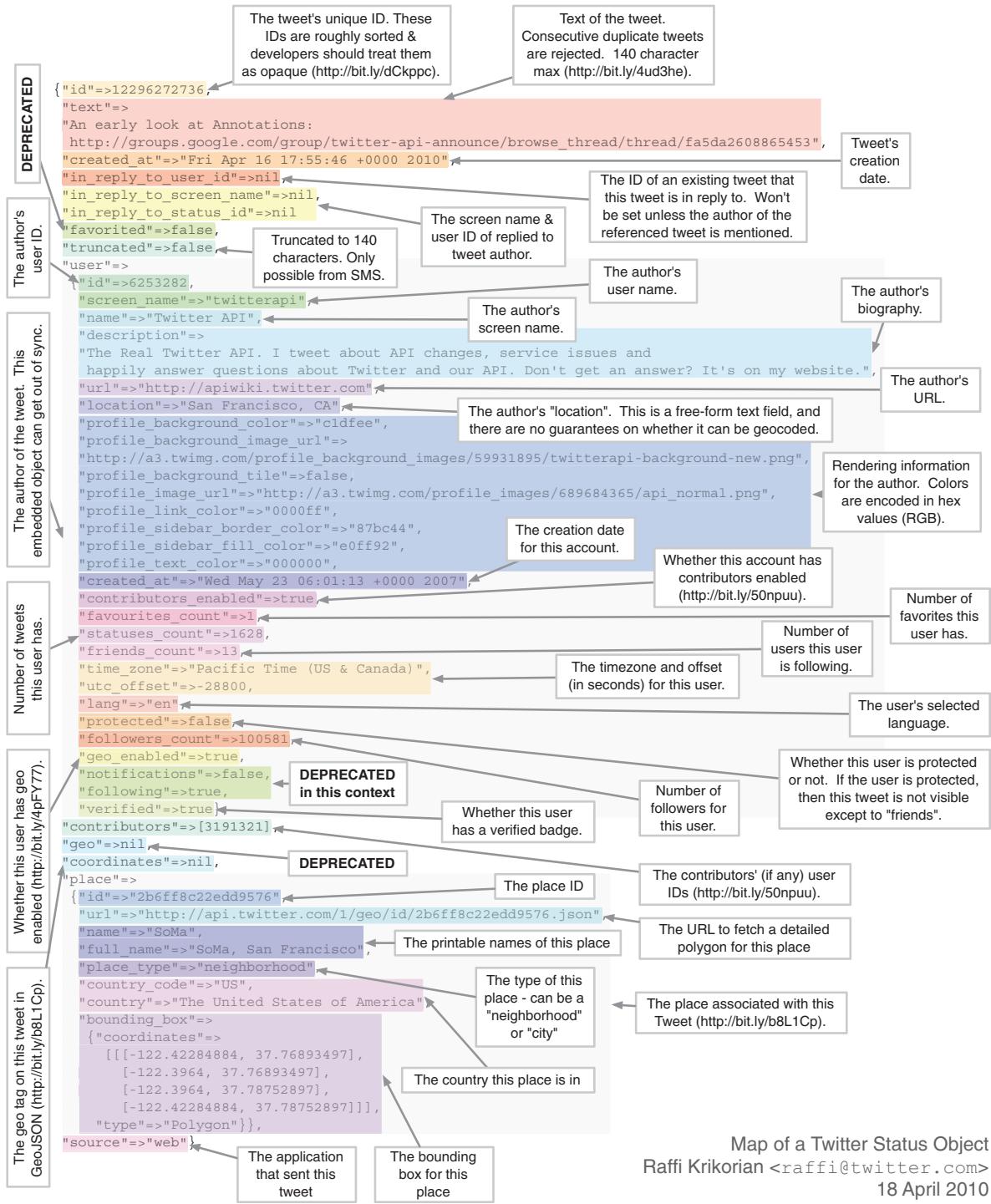
To provide context for the demonstrations that we developed, the project has identified Ten Principles for Online Privacy that describe at a high level how online privacy works, technically and socially. These principles form the basis of the TPP curriculum; each principle features an explanation of what it means and why it is important, as well as guidance on how certain privacy threats can be mitigated or outright avoided. We share these principles in Table 16.1.

The Teaching Privacy Project has used these principles as the basis for developing a teachers' kit with classroom-ready learning modules and a teachers' guide. This effort, the Teachers' Resources for Online Privacy Education (TROPE), aimed to provide high school and college instructors with resources to teach young people



**Figure 16.5** Education is part of the solution. Ready or Not: An educational browser tool showing that online image often includes meta-data that allows inference beyond the content of the image. Given a Twitter or Instagram handle, the tool shows a map with the recently reported GPS coordinates from the Twitter or Instagram feed. The GPS locations are clickable and lead to a Google Street View image. On the right, we also see a week schedule where the frequent posting times are visualized as a heatmap.

about why and how to protect their privacy online. Each of the TROPE teaching modules is centered around one of the ten principles and includes flexible lesson elements that can be used “out of the box” or adapted to supplement teachers’ existing lesson materials. These elements include explanations, discussion questions, and interactive demonstrations. Each module is structured around the 5E constructivist learning model for lesson planning: engagement, exploration, explanation, elaboration, and evaluation. The project is also developing a teachers’



**Figure 16.6** The anatomy of a Twitter post: The information shown in a Twitter post is less than the metadata stored and propagated with each post. (Courtesy of Raffi Krikorian)

**Table 16.1** Ten Principles for Online Privacy

You are Leaving Footprints

*Description:* Your information footprint is not just what you intentionally post online. It consists of all of the information that you post or that others post about you, the hidden data attached to those posts by the services you use, the record of your online activities, and also the inferences that can be drawn from putting that collective information together.

*Guidance:* Periodically check your privacy settings and update them to limit unintentional sharing.

There is no Anonymity

*Description:* Your information footprint on the Internet is like your body in the physical world: it defines your identity. Like seeing some part of your body, seeing some part of your information footprint (such as location of the device you're posting from or the pattern of your language) may allow someone to uniquely identify you, even when there is no name or other explicit identifier attached.

*Guidance:* Do not do anything online that you would not do in public.

Information is Valuable

*Description:* Every piece of information, public or not, has value to somebody: to other people, to companies and organizations, or to governments. They will use your information however benefits them, which may be contrary to your interests—and possibly even embarrassing or dangerous to you.

*Guidance:* If you're not sure how your information will be used, don't share it.

Someone Could Listen

*Description:* Unencrypted communication over the Internet works a lot like sending a postcard: it can be read by anybody along its route. Communication is routed through intermediary computers and systems, which are connected to many more computers and systems. Encryption, or encoding information so it appears scrambled to anyone who doesn't know the key, is a way to wrap a postcard in an envelope. While it can never be 100% secure, stronger encryption makes it harder for people to get to the contents.

*Guidance:* Use strong passwords and only communicate sensitive information over secure channels.

Sharing Releases Control

*Description:* Any time you interact online, that information is recorded in the network. And, as with in-person communication, once you've shared something, you can't control what happens to it—or how people will interpret it. Other people can repost or forward content to any audience without your permission, websites can sell information to other businesses, and data can be legally subpoenaed. Websites and search engines automatically pick up and duplicate content, making it impossible to "unshare"—the Internet never forgets!

*Guidance:* Think before sharing online; ask yourself if you would be comfortable becoming famous for it.

**Table 16.1** (*continued*)

**Search is Improving**

*Description:* Every day, more data is being put online. Search engines are getting better, allowing “deeper” searching of more types of data. Techniques for extracting and connecting information from different sources are getting more powerful. Furthermore, information that is not retrievable today may be retrievable tomorrow due to changes in terms of service, public policy, or privacy settings.

*Guidance:* Actively monitor your information footprint.

**Online is Real**

*Description:* Your online activities are as much a part of your life as your offline activities; they are interconnected and can affect your life and relationships in the same way.

*Guidance:* Share online as if everyone could see it, and would interpret it in the worst possible way.

**Identity is not Guaranteed**

*Description:* Creating an identity on the Internet or impersonating somebody else is often just a matter of a few clicks. Currently, there is no foolproof way to match a real person with their online identity. This means that you can never be sure with whom you are communicating, and that someone could steal your online identity and impersonate you!

*Guidance:* Before you share any information online, consider what you would be risking if the other party was not who you thought they were.

**You Cannot Escape**

*Description:* Even if you’re not actively using the Internet, someone else may be sharing information about you, intentionally or unintentionally. So, avoiding the Internet does not guarantee privacy.

*Guidance:* Share what you have learned with your friends and family—it will improve your own privacy.

**Privacy Requires Work**

*Description:* Most Internet technology is not designed to protect the privacy of those who use it; in fact, most technology providers make money by leveraging your private information. “Privacy policies” are generally written to protect providers from lawsuits, not to protect users’ privacy. Laws and regulations cover only certain aspects of privacy and vary from place to place. So, like it or not, your privacy is your own responsibility, and requires your constant attention.

*Guidance:* Encourage policymakers to develop comprehensive privacy regulations, educate yourself and others, check your options regularly.

guide that provides background information and context on privacy fundamentals, classroom discussion guides and support for implementing the lesson modules, and suggested lesson plans. All of the TROPE materials are being made available via a Teachers' Portal on [teachingprivacy.org](http://teachingprivacy.org), which will also include a discussion forum to solicit teacher feedback and answer questions.

However, we consider this only the first step in a long journey to integrate lessons like this into professional and university settings. Building effective educational components that transfer knowledge on privacy protection and the consequences of multimedia retrieval to younger adults who are not yet capable of understanding deep research results constitutes a new domain for research. Here, educational research needs to team up with HCI and other multimedia-related fields to attack this part of the new topic. The question is how to enable educators to master an up-to-date, scientifically-informed understanding of privacy, without having to rely on (often exaggerated) newspaper articles.

### **Focus Questions**

**16.1.** Have you ever posted something and then wished you had not? What did you do about it?

**16.2.** Take a look at a three different photographs: What do they depict? Now, what other information do you see?

**16.3.** Take a random video you took with your cell phone (unedited). Watch it. Now watch it again; close your eyes. What information can you identify in the background sounds?

**16.4.** Take a look at the privacy settings in your cell phone. When was the last time you looked at them? Did they change? Do you need to allow everything the apps ask for?

**16.5.** Name three sensors that can probably be abused to data mine your life.

**16.6.** Consider one of these sensors. What are the privacy consequences of its deployment? Given the multifaceted nature of privacy, there is no complete checklist of factors to consider when analyzing the privacy consequences of new technology. Some of these questions might provide relevant starting points for your analysis of the sensor you selected.

- (a) What information does it collect?
- (b) What is the intended use of this information?
- (c) What other uses exist for this information?

- (d) How can this information be combined with other sorts of information to draw new inferences about people?
- (e) How can the intended and unintended uses impact privacy?
- (f) Can the data be processed in a way that mitigates these impacts?
- (g) Can people who do not approve of the sensor's privacy consequences avoid the sensor?
- (h) Do you think the benefits of the intended use outweigh the privacy impacts, either with or without the mitigations?

**16.7.** Perform a similar privacy analysis for a type of multimedia analysis.

**16.8.** For either the sensor or analysis whose privacy impacts you examined above, design and implement a mitigation approach.

**16.9.** For either the sensor or analysis whose privacy impacts you examined above, design and implement an educational lesson or tool explaining its privacy risks and how to mitigate them.

**16.10.** Coming up with definitions of privacy is hard. In this and the next question, we'll look more closely at two definitions discussed above. Both definitions were originally stated with structured databases in mind. In Focus Question [16.12](#), we'll consider adapting these definitions for unstructured multimedia data.

This question will focus on  $k$ -anonymity, defined in Section [16.3.2.1](#). After a warm-up exercise, we will look at a series of questions about how much privacy  $k$ -anonymity provides.

- (a) Prove that for each database  $D$  and set of quasi-identifiers  $Q$ , there exists a value of  $k$  such that for all  $k'$  such that  $0 \leq k' \leq k$ , the database has  $k'$ -anonymity for  $Q$ , and for all  $k'' > k$ , the database does not have  $k''$ -anonymity for  $Q$ . This allows us to talk about the highest value of  $k$  such  $D$  has  $k$ -anonymity for  $Q$ .
- (b) Let's consider what an adversary can learn from a released database that does not have much variation in the value of the sensitive attribute.

Consider the example database  $D'_0$  that has 2-anonymity for  $Q_0$  from Section [16.3.2.1](#). Suppose a clinic wants to post  $D'_0$  publicly so that people can use it to study the incidence of cancer.

- (i) Suppose an adversary knows that Guy Fox is in the database and is 39, is male, and weighs 60 kg. Can the adversary infer from just that information and  $D'_0$  what type of cancer Guy Fox has?

- (ii) Suppose an adversary knows that Kate Gross is in the database and is female. Can the adversary infer from just that information and  $D'_0$  what type of cancer Kate Gross has?

- (iii) Does  $k$ -anonymity protect privacy as intuited above? Is there any value of  $k$  that is large enough to ensure that this issue will not arise for any database of any size?

(For more on this issue, see [Machanavajjhala et al. \[2007\]](#).)

- (c) Let's consider how the choice of quasi-identifiers affects privacy. Suppose the hospital releasing the data thinks that weight is not a quasi-identifier and considers it to instead be sensitive information. Further suppose it wants to release that data along side of the cancer data since it suspects there might be a relationship between patients' weights and types of cancer they get. Thus, the hospital instead releases the database  $D''_0$  shown in the following table.

Name	Age	Sex	Weight (kg)	Cancer
—	30s	Female	50	Lung
—	30s	Female	55	Lung
—	30s	Male	65	Throat
—	30s	Male	60	Lung
—	50s	Male	70	Throat
—	50s	Male	75	Larynx

- (i) Suppose an adversary knows that Guy Fox is in the database and is 39 and is male. Can the adversary infer from just that information and  $D''_0$  what type of cancer Guy Fox has?

- (ii) Suppose an adversary knows that Guy Fox is in the database and is 39, is male, and weighs 60 kg. Can the adversary infer from just that information and  $D''_0$  what type of cancer Guy Fox has?

- (iii) Is there any value of  $k$  that is large enough to ensure that this issue will not arise for any pair of databases that differ in whether a single attribute is considered a quasi-identifier?

- (d) Now, let's look at combining multiple released databases. Suppose a different hospital has the following database  $D_1$ .

Name	Age	Sex	Weight (kg)	Cancer
Carl Davis	33	Male	67	Larynx
Guy Fox	39	Male	60	Lung

Further suppose that the hospital releases the database  $D'_1$  using a similar method achieving 2-anonymity.

Name	Age	Sex	Weight (kg)	Cancer
—	30s	Male	60s	Larynx
—	30s	Male	60s	Lung

- (i) Suppose an adversary knows that Guy Fox is in both the databases and is 39, is male, and weighs 60 kg. Can the adversary infer from just that information,  $D'_0$ , and  $D'_1$  what type of cancer Guy Fox has?
- (ii) Is there any value of  $k$  that is large enough to ensure that this issue will not arise for any pair databases considered together?
- (iii) One heuristic for determining whether an action is morally good is ask yourself what the world would be like if everyone did it. Similarly, one way to judge whether a privacy definition is useful is to ask yourself what would happen if everyone were to use it. Ideally, in such a case, people still have a high degree of privacy, at least as measured by that privacy definition itself. What does the above result say about  $k$ -anonymity as a privacy definition?

(For more information on this issue, see [Ganta et al. \[2008\]](#), from which the presented example is adapted.)

**16.11.** This question will focus on differential privacy, defined in Section [16.3.2.2](#).

- (a) Prove that when the set  $\mathcal{O}$  of possible outputs of the algorithm  $a$  is finite,  $a$  has  $\epsilon$ -differential privacy if and only if for all outputs  $o$ :

$$\Pr[a(D_1)=o] = e^\epsilon * \Pr[a(D_2)=o]$$

- (b) Suppose the desired statistic is simply whether or not anyone in the database has lung cancer where the database has the same format as  $D_0$  in Focus Question [16.10](#). Design a simple algorithm that provides  $\epsilon$ -differential privacy for answering this query and prove that it provides  $\epsilon$ -differential privacy.
- (c) Prove that if an algorithm has  $\epsilon$ -differential privacy, then for all  $\epsilon'$  such that  $\epsilon' > \epsilon$ , it has  $\epsilon'$ -differential privacy.
- (d) Give an example of an algorithm that does not have  $\epsilon$ -differential privacy for any value of  $\epsilon$ . This means we cannot directly speak of the best (lowest) value of differential privacy offered by an algorithm as we could speak of the best (highest) value of  $k$ -anonymity had by a database. However, we can recover this ability by allowing  $\epsilon$  to take the value of  $\infty$  and letting  $\infty$ -differential privacy hold for any algorithm.

- (e) Let the parallel composition of two algorithms  $a_1$  and  $a_2$  be  $a_3$  where  $a_3(D) = \langle a_1(D), a_2(D) \rangle$ . If  $a_1$  has  $\epsilon_1$ -differential privacy and  $a_2$  has  $\epsilon_2$ -differential privacy, what can be said about best (lowest) degree of privacy offered by  $a_3$ ?

**16.12.** We now consider some issues involved in applying these definitions, which were designed for structured data, to unstructured multimedia.

- (a)  $k$ -anonymity requires selecting a subset of attributes about people to be quasi-identifiers. Multimedia data is often raw outputs from sensors, which may be about many people at once and stored in attributes without obvious semantic meanings, such as pixels in frames of videos. What challenges does this pose for determining a reasonable set of quasi-identifiers?
- (b) Differential privacy depends upon identifying the data associated with a single person. What challenges does this pose for multimedia data?
- (c) One might wish to run multimedia data through an algorithm that provides differential privacy and provides output that is also multimedia data of the same type. For example, one could imagine running video through an algorithm that randomly either adds or removes each possible person from that video. What difficulties does this task pose?

## Acknowledgments

Some of the material in Sections 16.2 and 16.3 is based upon MCT's prior work [Tschantz and Wing 2009]. Section 16.5.1 draws from GF's prior work [Friedland et al. 2011]. Section 16.5.2 comes from ongoing work with Nicholas Weaver [Friedland and Weaver 2018]. Section 16.6.2 is based upon the Teaching Privacy Project [International Computer Science Institute 2016]. Figure 16.6 is from Raffi Krikorian.

The National Science Foundation (Grant CNS 1514509) supported writing this chapter. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. government.

## References

- A. Adams and A. M. Sasse. 2001. *Privacy in Multimedia Communications: Protecting Users, Not Just Data*, pp. 49–64. Springer London. ISBN 978-1-4471-0353-0. DOI: [10.1007/978-1-4471-0353-0\\_4. 673](https://doi.org/10.1007/978-1-4471-0353-0_4)
- C. Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In *Proceedings of the International Conference on Very Large Data Bases*. Trondheim, Norway. 667

- T. Alloway. 2015. Big data: Credit where credit's due. *Financial Times*. **662**
- R. Baker, E. Y. Wang, and L. Paquette. 2015. Big data in education. <http://www.edx.org/course/big-data-education-teacherscollege-x-bde1x>. **662**
- M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. 2010 Abusing social networks for automated user profiling. In *RAID'2010, 13th International Symposium on Recent Advances in Intrusion Detection*. **668**
- A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. 2006. Privacy and contextual integrity: framework and applications. *IEEE Symposium on Security and Privacy*. DOI: [10.1109/SP.2006.32](https://doi.org/10.1109/SP.2006.32). **673**
- R. Bassily, A. Groce, J. Katz, and A. Smith. 2013. Coupled-worlds privacy: exploiting adversarial uncertainty in statistical data privacy. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 439–448. IEEE Computer Society, 2013. DOI: [10.1109/FOCS.2013.54](https://doi.org/10.1109/FOCS.2013.54). **673**
- M. Bishop, J. Cummins, S. Peisert, A. Singh, B. Bhumiratana, D. Agarwal, D. Frincke, and M. Hogarth. 2010. Relationships and data sanitization: a study in scarlet. In *Proceedings Workshop on New Security Paradigms*. **668**
- E. Bloustein. 1964. Privacy as an aspect of human dignity: An answer to dean prosser. *New York University Law Review*, 39: 962. **665**
- A. J. Blumberg and P. Eckersley. 2009. On locational privacy, and how to avoid losing it forever. Electronic Frontier Foundation whitepaper. <https://www.eff.org/wp/locational-privacy>. **668**
- M. Burzo, M. Abouelenien, V. Perez-Rosas, and R. Mihalcea. 2018. Multimodal deception detection. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces*, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition. Morgan & Claypool Publishers, San Rafael, CA. **660**
- S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. 2005. Toward privacy in public databases. *2nd Theory of Cryptography Conference (TCC 2005)*, pp. 363–385. **667**
- R. Chen, B. C. Fung, P. S. Yu, and B. C. Desai. 2014. Correlated network data publication via differential privacy. *The VLDB Journal*, 23(4):653–676. DOI: [10.1007/s00778-013-0344-8](https://doi.org/10.1007/s00778-013-0344-8). **673**
- J. Choi, I. E. Akkus, S. Egelman, G. Friedland, R. Sommer, M. C. Tschantz, and N. Weaver. 2018. Cybercasing 2.0: You Get What You Pay For. ArXiv 1811.06584. **680**
- J. Choi, A. Janin, and G. Friedland. 2010. The 2010 ICSI video location estimation system. In *Proceedings of Mediaeval 2010*. Pisa, Italy. **675**
- R. Chow, P. Golle, and J. Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, USA. **668**
- N. Ciarelli. April 2010. How Visa predicts divorce. *The Daily Beast*. **660**

- P. Cohan. 2009. Why executives risk their job to tip a hedge fund. <http://meet-the-street.blogspot.com/2009/10/expert-networks-what-every-iro-needs-to.html>. 678
- J. Cohen. 2002. *Regulating Intimacy: A New Legal Paradigm*. Princeton University Press. Princeton, NJ. 665
- J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer. 2018. Multimodal assessment of depression and related disorders based on behavioural signals. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 660
- A. Conway. 2012. Face blurring: when footage requires anonymity. <https://youtube.googleblog.com/2012/07/face-blurring-when-footage-requires.html>. 686
- L. F. Cranor. 2002. *Web Privacy with P2P*. O'Reilly. 673
- L. F. Cranor, P. Guduru, and M. Arjula. 2006. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction*, 13(2):135–178, 2006. DOI: <http://doi.acm.org/10.1145/1165734.1165735>. 673
- T. Dalenius. 1977. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444. 664, 666, 786
- I. Dinur and K. Nissim. 2003. Revealing information while preserving privacy. *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*. 667
- F. Dufaux and T. Ebrahimi. 2006. Scrambling for video surveillance with privacy. *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pp. 160–160. DOI: <10.1109/CVPRW.2006.184>. 674
- F. Dufaux and T. Ebrahimi. 2010. A framework for the validation of privacy protection solutions in video surveillance. *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pp. 66–71. DOI: <10.1109/ICME.2010.5583552>. 674
- C. Duhigg. 2012. Psst, you in aisle 5: How companies learn your secrets. *The New York Times Magazine*, p. MM30. 660
- C. Dwork. 2006. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, vol. 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer. DOI: [10.1007/11787006\\_1](10.1007/11787006_1). 663, 666, 667, 671, 768
- C. Dwork. 2008. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, vol. 4978 of *Lecture Notes in Computer Science*, pp. 1–19. Springer, Berlin/Heidelberg. 667
- C. Dwork and M. Naor. 2008. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107. 666, 667

- C. Dwork, F. McSherry, K. Nissim, and A. Smith. 2006. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, pp. 265–284. Springer. [663](#), [667](#), [671](#), [768](#)
- C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4): 211–407. [672](#)
- European Commission. 2016. General data protection regulation (GDPR). Regulation (EU) 2016/679, L119. [662](#)
- Executive Office of the President. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016\\_0504\\_data\\_discrimination.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf). Accessed October 17, 2016. [662](#)
- J. Fan, H. Luo, M.S. Hacid, and E. Bertino. 2005. A novel approach for privacy-preserving video sharing. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pp. 609–616, New York. ACM. DOI: [10.1145/1099554.1099711](https://doi.org/10.1145/1099554.1099711). [674](#)
- Federal Committee on Statistical Methodology. Statistical disclosure limitation methodology. Statistical Policy Working Paper 22, 2005. [668](#)
- C. Fried. 1970. *An Anatomy of Values*. Harvard University Press. Cambridge, MA. [665](#)
- G. Friedland, G. Maier, R. Sommer, and N. Weaver. September 2011. Sherlock holmes evil twin: On the impact of global inference for online privacy. In *Proceedings of the New Security Paradigms Workshop (NSPW)*. [677](#), [696](#)
- G. Friedland and J. Choi. 2011. Semantic computing and privacy: a case study using inferred geo-location. *International Journal of Semantic Computing*, 5(01): 79–93. [674](#), [675](#), [680](#)
- G. Friedland and R. Sommer. 2010. Cybercasing the joint: on the privacy implications of geo-tagging. In *Proceedings USENIX Workshop on Hot Topics in Security*. [674](#), [675](#), [679](#)
- G. Friedland and N. Weaver. 2018. How's the trip? In J. Choi, I. E. Akkus, S. Egelman, R. Sommer, and M. C. Tschantz, “Cybercasing 2.0: You Get What You Pay For.” [696](#)
- S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. 2008. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 265–273. ACM, 2008. DOI: [10.1145/1401890.1401926](https://doi.org/10.1145/1401890.1401926). [671](#), [695](#)
- Ruth Gavison. 1980. Privacy and the limits of law. *Yale Law Journal*, 89(3): 421–471. [665](#), [667](#)
- T. Gerety. 1977. Redefining privacy. *Harvard Civil Rights-Civil Liberties Law Review*, 12: 233–296. [665](#)
- R. Gerstein. 1978. Intimacy and privacy. *Ethics*, 89: 76–81. [665](#)
- O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 447–458. ACM. [684](#)

- V. Griffith and M. Jakobsson. 2005. Messin' with texas deriving mother's maiden names using public records. In *Proceedings of the International Conference on Applied Cryptography and Network Security (ACNS)*. 668
- J. Halliday. 2012. Google introduces face-blurring to protect protesters on YouTube. *The Guardian*. <http://www.theguardian.com/technology/2012/jul/19/face-blurring-technology-youtube-protestors>. 686
- J. Y. Halpern and K. R. O'Neill. 2008. Secrecy in multiagent systems. *ACM Transactions Information and System Security*, 12(1): 5:1–5:47. DOI: [10.1145/1410234.1410239](https://doi.org/10.1145/1410234.1410239). 667, 673
- X. He, A. Machanavajjhala, and B. Ding. 2014. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2014)*. ACM. 673
- B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson. 2008. Virtual trip lines for distributed privacy-preserving traffic monitoring. *MobiSys '08: Proceeding of the 6th International Conference on Mobile Systems, Applications, and Services*. 668
- InsideFacebook. 2010. December data on facebook's us growth by age and gender: Beyond 100 million. Adweek Blog Network's SocialTimes. <http://teachingprivacy.org>. 687
- International Computer Science Institute. 2016. Tools for teaching privacy to k12 and undergraduate students. <http://teachingprivacy.icsi.berkeley.edu>. 687, 696
- S. P. Kasiviswanathan and A. Smith. 2014. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1): 1–16. 673
- D. Kifer and A. Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204. ACM. 672, 673
- D. Kifer and A. Machanavajjhala. 2012. A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 77–88. ACM. DOI: [10.1145/2213556.2213571](https://doi.org/10.1145/2213556.2213571). 673
- D. Kifer and A. Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1): 3: 1–3:36. DOI: [10.1145/2514689](https://doi.org/10.1145/2514689). 673
- T. Koshimizu, T. Toriyama, and N. Babaguchi. 2006. Factors on the sense of privacy in video surveillance. In *Proceedings of the 3rd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, CARPE '06, pp. 35–44, New York. ACM. DOI: [10.1145/1178657.1178665](https://doi.org/10.1145/1178657.1178665). 674
- H. Lei, J. Choi, A. Janin, and G. Friedland. 2011. User verification: Matching the uploaders of videos across accounts. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 2404–2407. IEEE. 682
- N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang. 2013. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer*

- and Communications Security*, CCS '13, pp. 889–900. ACM. [10.1145/2508859.2516686](https://doi.org/10.1145/2508859.2516686). 673
- N. Li, T. Li, and S. Venkatasubramanian. 2017. *t*-closeness: Privacy beyond  $k$ -anonymity and  $l$ -diversity. *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 106–115. [10.1109/ICDE.2007.367856](https://doi.org/10.1109/ICDE.2007.367856). 669
- C. Liu, S. Chakraborty, and P. Mittal. 2016. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *Network and Distributed System Security Symposium (NDSS)*. The Internet Society. DOI: [10.14722/ndss.2016.23279](https://doi.org/10.14722/ndss.2016.23279). 673
- J. Lukas, J. Fridrich, and M. Goljan. 2006. Digital camera identification from sensor pattern noise. *Information Forensics and Security, IEEE Transactions on*, 1(2): 205–214. DOI: [10.1109/TIFS.2006.873602](https://doi.org/10.1109/TIFS.2006.873602). 674, 676
- A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. 2007.  $\ell$ -Diversity: Privacy beyond  $k$ -anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1): 3. DOI: [10.1145/1217299.1217302](https://doi.org/10.1145/1217299.1217302). 669, 671, 694
- F. McSherry. 2016a. Lunchtime for data privacy. <https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md>. 673
- F. McSherry. 2016b. Differential privacy and correlated data. <https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-29.md>. 673
- F. D. McSherry. 2009. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30. ACM. 672
- Merriam-Webster. 2013. Privacy. <http://www.merriam-webster.com/dictionary/privacy>. 665
- D. K. Mulligan, C. Koopman, and N. Doty. 2016. Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A*, 374(20160118). 665
- T. B. Murdoch and A. S. Detsky. 2013. The inevitable application of big data to health care. *JAMA*, 309(13): 1351. 662
- A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy*. Oakland, CA. 667
- A. Narayanan and V. Shmatikov. 2009. De-anonymizing social networks. In *Proceedings of the IEEE Symposium on Security and Privacy*. Oakland, CA. 667, 668
- National Institute of Standards and Technology. 2016. TREC video retrieval evaluation home page. <http://www-nlpir.nist.gov/projects/trecvid>. 685
- C. Neustaedter, S. Greenberg, and M. Boyle. March 2006. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions of Computer-Human Interaction*, 13(1): 1–36. DOI: [10.1145/1143518.1143519](https://doi.org/10.1145/1143518.1143519). 674
- New Hampshire Supreme Court. 1964. Hamberger v. Eastman. *Atlantic Reporter*, 206: 239. Stanford, CA. 665
- H. Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books. Stanford University Press. 663, 665, 673, 765

- Council of the OECD. 1980. OECD guidelines on the protection of privacy and transborder flows of personal data. <http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldatal.htm>. 662
- S. Oviatt, J. F. Grafsgaard, L. Chen, and X. Ochoa. 2018. Multimodal learning analytics: Assessing learners' mental state during the process of learning. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, editors, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool Publishers, San Rafael, CA. 662
- W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood. 2013. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation. 662
- G. Piatetsky. 2014. Did target really predict a teen's pregnancy? The inside story. *KDnuggets*. <http://www.kdnuggets.com/2014/05/target-predict-teen-pregnancy-inside-story.html> 660
- J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren, and J. Zients. 2014. Big data: Seizing opportunities, preserving values. [https://obamawhitehouse.archives.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf) Accessed January 26, 2014. 662
- R. A. Popa, H. Balakrishnan, and A. Blumberg. 2009. VPriv: Protecting Privacy in Location-Based Vehicular Services. In *Proceedings of the USENIX Security Symposium*. Montreal, Canada. 668
- C. Powers and M. Schunter. 2003. Enterprise privacy authorization language (EPAL 1.2). W3C Member Submission. Cambridge, MA. 673
- President's Council of Advisors on Science and Technology. 2014. Big data and privacy: A technological perspective. Report to the president, Executive Office of the President, Washington, DC. 662
- PRNewswire. 2015. Big data in government, defense and homeland security 2015–2020. <http://www.prnewswire.com/news-releases/big-data-in-government-defense-and-homeland-security-2015—2020-300133749.html>. 662
- W. L. Prosser. 1960. Privacy. *California Law Review*, 48: 383. 665
- J. Rachels. 1975. Why privacy is important. *Philosophy and Public Affairs*, 4: 323–333. 665
- S. Ribaric, A. Ariyaeenia, and N. Pavescic. 2016. De-identification for privacy protection in multimedia content: A survey. *Image Communications*, 47(C):131–151. DOI: [10.1016/j.image.2016.05.020](https://doi.org/10.1016/j.image.2016.05.020). 673
- P. Samarati and L. Sweeney. 1998a. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report. <http://www.csl.sri.com/papers/srir-98-04/>. 664, 669, 670, 777
- P. Samarati and L. Sweeney. 1998b. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART*

- Symposium on Principles of Database Systems*, pp. 188–. ACM. DOI: [10.1145/2754872.275508](https://doi.org/10.1145/2754872.275508). **664, 669, 670, 777**
- Secretary's Advisory Committee on Automated Personal Data Systems. 1973. Records, computers, and the rights of citizens. Technical report, U.S. Department of Health, Education, and Welfare. **662**
- T. Simonite. 2014. Facebook creates software that matches faces almost as well as you do. *MIT Technology Review*. <https://www.technologyreview.com/s/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do/> **685**
- D. J. Solove. 2006. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3): 477–560, 2006. **663, 664, 665, 761, 771, 775, 776, 784**
- J. Staddon, P. Golle, and B. Zimny. 2007. Web-based inference detection. In *Proceedings of 16th USENIX Security Symposium*. **668**
- Supreme Court of the United States. 1928. Olmstead v. United States. *United States Reports*, 277:438. **664**
- Supreme Court of the United States. Katz v. United States. 1967. *United States Reports*, 389:347. **664**
- Supreme Court of the United States. 1983. United States v. Knotts. *United States Reports*, 460:276. **665**
- Supreme Court of the United States. 1984. United States v. Karo. *United States Reports*, 468:705. **665**
- Supreme Court of the United States. 1986. Dow Chemical Co. v. United States. *United States Reports*, 476:227. **664**
- Supreme Court of the United States. 1989. Florida v. Riley. *United States Reports*, 488:455. **664, 665**
- Supreme Court of the United States. 2001. Kyllo v. United States. *United States Reports*, 533:27. **665**
- L. Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics*, 25(2–3). **667**
- L. Sweeney. 2002a. *k*-Anonymity: A model for protecting privacy. *International Journal of Uncertainain Fuzziness and Knowledge-based Systems*, 10(5): 557–570. DOI: [10.1142/S0218488502001648](https://doi.org/10.1142/S0218488502001648). **664, 669, 777**
- L. Sweeney. 2002b. *k*-Anonymity: A model for protecting privacy. *Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5). **667**
- The White House. 2012. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. *Journal of Privacy and Confidentiality*, 4(2):9 5–142. **662, 665**
- B. Thuraisingham. 2007. Security and privacy for multimedia database management systems. *Multimedia Tools Appliances*, 33(1): 13–29. DOI: [10.1007/s11042-006-0096-1](https://doi.org/10.1007/s11042-006-0096-1). **673**

- M. C. Tschantz and J. M. Wing. 2009. Formal methods for privacy. In *Proceedings of the 2nd World Congress on Formal Methods*, pp. 1–15. Springer-Verlag. [696](#)
- D. Walker-Morgan. 2011. Cree.py application knows where you've been. The H Security. <http://www.h-online.com/security/news/item/Cree-py-application-knows-where-youve-been-1217981.html>. [668](#)
- Warren and Brandeis. 1890. The right to privacy. *Harvard Law Review*, IV(5). [664](#)
- Wikipedia. 2013. Secure Communication. [http://en.wikipedia.org/wiki/Secure\\_communication](http://en.wikipedia.org/wiki/Secure_communication). [666](#)
- X. Xiao and Y. Tao. 2007. *m*-Invariance: Towards privacy preserving re-publication of dynamic datasets. *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 689–700. ACM Press. DOI: [10.1145/1247480.1247556](https://doi.org/10.1145/1247480.1247556). [669](#)
- G. Zhong, I. Goldberg, and U. Hengartner. 2007. Louis, lester and pierre: Three protocols for location privacy. In *Proceedings of the Privacy Enhancing Technologies Symposium*. Ottawa, Canada. [668](#)
- T. Zhu, P. Xiong, G. Li, and W. Zhou. 2015. Correlated differential privacy: Hiding information in non-IID data set. *IEEE Transactions on Information Forensics and Security*, 10(2): 229–242. DOI: [10.1109/TIFS.2014.2368363](https://doi.org/10.1109/TIFS.2014.2368363). [673](#)

# Index

- 3D recordings privacy concerns, 676  
6DOF  
    defined, 395  
    TUM Kitchen dataset, 411
- Abstract User Interface (AUI)  
    declarative languages, 360  
    defined, 348
- Abstraction level relationships in declarative languages, 361
- Acceptance level of action in speaking, 93–94
- Accessibility in information processing, 666
- Accuracy in late fusion movement modeling, 311
- Acoustic noise, privacy concerns for, 675
- Acted data collection and annotation in nonverbal behavior, 226–227
- Actions  
    engagement, 117–119  
    perception, 131–135
- Active BCI  
    defined, 528  
    embedded multimodal interfaces, 536–537, 541
- Active exoskeletons, 544
- Active states  
    defined, 528  
    embedded multimodal interfaces, 524, 535–536  
    exoskeletons, 566–567  
    state machines, 376
- AD (Alzheimer Disease), 446, 458
- Adam optimizer for movement modeling, 337
- ADAPT platform, 239–240
- Adaptive probabilistic approach for multimodal fission, 152–153
- Adaptive responses in self-adaptation, 557
- Adaptivity in situation-adaptive behavior, 177
- Addressee detection  
    defined, 79  
    human-robot interaction, 86  
    turn-taking, 92
- AE-DMP approach in movement modeling, 318, 320–321, 324–325, 327
- Aeronautics, 621–625
- AEs (autoencoders)  
    defined, 307  
    movement modeling, 324–327, 335–337  
    variational, 11
- Affective Computing  
    defined, 395  
    overview, 396–398
- Agents in SAIBA framework, 228–229
- Aggregation  
    defined, 663  
    privacy issues, 563, 665
- Agreement in simultaneous interpretation, 599
- AI. *See* Artificial intelligence (AI)
- AIBO Robot, 395

- AIBO robot database
  - characteristics, 404
  - data collection, 401–402
  - interface realism, 400
  - recordings, 398
  - resource location, 406
- AKT cognitive assessment, 451, 456
- Alarms for in-car interaction, 492–493
- Alexa assistant
  - social robot, 633
  - VDA, 180
- Alexa Presentation Language (APL), 639
- Alexa Skills Kit, 382
- Alice ECA, 241–244
- Alignment
  - audio, 384
  - defined, 222
  - nonverbal behavior, 225, 236
  - speech and grammar, 44–45, 49, 51, 57
- Allocation of logical tasks, 359
- Alzheimer Disease (AD), 446, 458
- Amazon Echo Show, 638–639
- American Automobile Association report for multimodal in-car interaction, 502–503
- AMI database
  - characteristics, 404
  - description, 402–403
  - resource location, 406
- Anaphoric referring expressions, 162
- Android Auto system, 505
- Annotated corpora for nonverbal behavior, 230–231
- Annotation for nonverbal behavior, 226–227
- Anomaly treatment for MCPS, 429
- Anonymization
  - defined, 663
  - human-computer interaction, 564–565
  - privacy concern threats, 661, 667–668
- Anonymous social media accounts, 682–684
- Anthropometric posture prediction models, 280
- Anvil tool, 128
- ANVIL Video Annotation Research Tool, 407
- AnyBody software, 288
- APIs (Application Programming Interfaces)
  - defined, 348
  - proprietary, 350
- APL (Alexa Presentation Language), 639
- Apple Siri VDA, 180
- Application domains
  - defined, 427
  - MCPS, 430
- Application layers in dialogue systems, 147–148
- Application Programming Interfaces (APIs)
  - defined, 348
  - proprietary, 350
- Application-specific safety level
  - defined, 528
  - inherently safe robots, 531
- AR (augmented reality), 644–646
- Archival ink applications, 367
- ARGs (Attributed Relational Graphs)
  - defined, 26
  - multimodal reference resolution, 51
- ARIA-Valuspa project, 242–244
- Arms, prosthetic, 629–631
- Arousal
  - DEAP database, 403–404
  - defined, 395
  - MAHNOB-HCI database, 408
- Articulated Social Agents Platform (ASAP), 235–237
- Artificial intelligence (AI)
  - early attempts, 105
  - frames, 28
  - human-robot interaction, 559–560
  - medical and health systems, 426, 446
  - natural language processing, 504
- Artificial skins for robots, 531–532
- ASAP (Articulated Social Agents Platform), 235–237
- ASD (autism spectrum disorders)
  - joint attention, 95
  - social robots for, 635
- ASR. *See* Automatic speech recognition (ASR)

- Assignment
  - defined, 158
  - multimodal input, 157
- Assist-as-needed approach, 541
- Assistant agent, 134
- Assistants, 638–642
- Assistive robots, 635
- Asymmetry in vocal navigation systems, 503
- AT&T speech mashup architecture, 164–165
- ATIS project, 105
- ATNs (Augmented Transition Networks)
  - defined, 26
  - incremental multimodal integration, 49
  - multimodal fusion, 32
- Attacks on privacy, 676–677
- Attention Flow module, 89
- Attention mechanisms in machine translation, 586
- Attention persistence in multimodal fusion, 124
- Attributed Relational Graphs (ARGs)
  - defined, 26
  - multimodal reference resolution, 51
- Audio
  - human-robot interaction, 84–87
  - multimodal in-car interaction feedback, 494
  - nonverbal behavior, 232
- Audio-visual speech recognition, 24–25
- Augmented reality (AR), 644–646
- Augmented Transition Networks (ATNs)
  - defined, 26
  - incremental multimodal integration, 49
  - multimodal fusion, 32
- AUI (Abstract User Interface)
  - declarative languages, 360
  - defined, 348
- Autism spectrum disorders (ASD)
  - joint attention, 95
  - social robots for, 635
- Autoencoders (AEs)
  - defined, 307
  - movement modeling, 324–327, 335–337
  - variational, 11
- Automatic expressive behavior understanding, 436
- Automatic image recognition, 439–442
- Automatic SAL scenario, 410
- Automatic speech recognition (ASR)
  - defined, 578
  - error detection, 605
  - human-robot interaction, 86–87
  - machine translation, 582–583
  - medical and health systems, 440
- Automatic video blurring, 674
- Automotive multimodal human-machine interface
  - challenges and opportunities, 481–486
  - focus questions, 514–515
  - gaze, 510–512
  - gestures, 495–497
  - glossary, 478–479
  - handwriting recognition, 497–498
  - haptic controls, 487–493
  - HMI evolution, 477, 479–481
  - introduction, 477
  - overview, 486–487
  - references, 515–522
  - secondary displays, 507–510
  - summary and outlook, 512–514
  - touch screens, 493–495
  - voice, 498–507
- Autonomous cars, 625–628
- Autonomous Sensor Management for F-35 helmet, 623
- Auxiliary information
  - defined, 663
  - privacy issues, 668
- Ava avatar, 642
- Avatars, 642–643
- Awareness
  - defined, 528
  - human-robot interaction, 524, 526
- Back-channeling
  - defined, 158
  - SmartKom system, 167

- Backchannels
  - defined, 79
  - feedback, 94
  - multimodal fusion, 124
  - turn-taking, 92–93
- Background checks, privacy concerns for, 677
- Barriers for car controls, 489
- BCI. *See* Brain-computer interfaces (BCIs)
- BDI (beliefs, desires, and intentions)
  - models, 156
- Beamforming, 79
- BEAT (Behavior Expression Animation Toolkit), 239–240
- Beat gestures
  - defined, 79, 222
  - nonverbal behavior, 246
  - speech, 81–82
- Behavior coherency in nonverbal behavior, 246
- Behavior Expression Animation Toolkit (BEAT), 239–240
- Behavior Lexicon for nonverbal behavior, 233
- Behavior Markup Language (BML)
  - defined, 222, 348
  - multimodal fission and media synchronization, 373, 376
  - nonverbal behavior, 233–235, 238
  - SAIBA framework, 229–230
  - VIB, 243–244, 249–251
- Behavior Planners in SAIBA framework, 230
- Behavior sets in VIB, 243
- Belfast Driving simulator data, 408
- Beliefs, desires, and intentions (BDI)
  - models, 156
- Benchmark databases, 394, 396
- Bilateral awareness in robotic applications, 545–546
- Binary questions in discomfort questionnaires, 270
- Biocybernetic controls
  - defined, 528
  - embedded multimodal interfaces, 537
- Biofeedback, 437
- Biomarkers, 437
- Biomechanical distractions in in-car interaction, 496
- Biomechanical simulation in ergonomics, 275, 285–289
- Biometric liveness measures, 624
- Biometric spoofing
  - defined, 624
  - minimizing, 637
- Biometrics, 636–638
- Bionic hands, 630
- Biosignals
  - Bitalino platform, 437
  - defined, 427
  - MCPS, 429
  - multisensor interfaces, 432
  - robotic systems, 553
- Bixby Home assistant, 638–639
- Bixby Vision assistant, 640
- Blackmail, privacy concerns, 678–679
- Blood pressure (BP) as physical ergonomics measurement, 273
- BML. *See* Behavior Markup Language (BML)
- BNNs in nonverbal behavior, 232
- Bodies
  - nonverbal behavior, 220–225
  - SAIBA framework, 228–229
- Boeing 737 Max 8 aircraft crashes, 623–624
- BOLT program, 606
- Borg scales for discomfort questionnaires, 270
- Bottleneck CNN in movement modeling, 318
- BP4D database
  - characteristics, 404
  - description, 403
  - resource location, 406
- Brain-computer interfaces (BCIs)
  - active, 528
  - embedded multimodal interfaces, 536–541
  - hybrid, 529
  - passive, 530

- prosthetic arms, 630–631
- reactive, 530
- Brain readings.* *See* Embedded brain readings
- Brake reaction time in multimodal in-car interaction, 498–499
- Breeno assistant, 641
- Broad differences and techniques in nonverbal behavior, 230–231
- Burglary, privacy concerns for, 680–681
- Bystanders, turn-taking by, 92
- C-STAR Systems, 590
- CALO system, 105
- Cameras
  - autonomous cars, 626
  - dialogue systems, 149, 167
  - embedded multimodal interfaces, 534
  - ergonomics, 274, 283
  - existing databases, 408–411
  - joint attention, 95–96
  - modality fusion, 372
  - modeling human-robot interaction, 85–86, 88
  - motion models, 312, 315
  - privacy concerns, 665, 674–676
  - robots, 533
- Canal 9 database
  - characteristics, 404
  - description, 403
  - resource location, 406
- CARE properties
  - abstraction level relationships, 361
  - multimodal input, 157
- CarPlay system, 505
- Cars
  - autonomous, 625–628
  - human-machine interface. *See* Automotive multimodal human-machine interface
- Case studies
  - digital pen interface, 445–453
  - multimodal dialogue system, 438–445
- multimodal-multisensor framework, 453–458
- privacy concerns, 677–684
- Casing in spoken language translation, 579
- Castaway Reality TV Database, 408
- Categorical questions in discomfort questionnaires, 270
- Category-Ratio 10 (CR10) scale for discomfort questionnaires, 270
- CAVEs for robotic applications, 549
- CDTs (clock drawing test), 446, 451, 454–456
- Center stacks in in-car interaction, 486
- Central executive in cognitive load, 482
- CERAD cognitive assessment, 447, 451, 453
- Cerebella system, 237–238
- CFs (Communicative Functions)
  - Cerebella system, 237
  - fusion, 157, 159–162
- Challenge discussion
  - chatbot dialogues, 194
  - collaborative dialogues, 193–194
  - conversation, 196–216
  - discussion questions, 193–196
  - future directions, 213–216
  - hybrid dialogue reasoning, 195–196
  - introduction, 191–193
  - machine learning approaches, 205–208
  - multimodal dialogues, 195
  - opportunities, 196
  - references, 216–217
  - transfer learning, 196, 209–213
- Chameleon effect
  - defined, 222
  - nonverbal behavior, 225
- Chart parsing
  - DAGs, 27
  - defined, 26
  - multimodal fusion, 32
  - unification-based multimodal fusion, 38
- Chatbot dialogues
  - challenge discussion, 194
  - overview, 181–182
  - transactional dialogues, 200–205
- Chit-chat bots, 181

- Chunking in finite-state approaches, 44
- CKY chart parsing algorithm, 38
- Classifiers for gestures, 54
- Clinical medical and health systems
  - description, 426
  - multimodal interfaces, 431–432
  - multisensor interfaces, 432–433
  - overview, 429–430
- Clock drawing test (CDTs), 446, 451, 454–456
- Clock synchronization, 418
- Closed-loop control and design
  - defined, 528
  - reflexive adaptation, 560
  - self-adaptation, 556–557
- Closed-loop fusion system
  - defined, 624
  - F-35 helmet, 623
- CMC (computed muscle control) in biomechanical simulation, 288
- CMU MMAC database
  - characteristics, 404
  - description, 403
  - resource location, 406
- CNN (convolutional neural network)
  - architecture in movement modeling, 321
- Co-reference resolution
  - defined, 26
  - interactive conversational systems, 25
- Code switching
  - defined, 578
  - simultaneous interpretation, 600
- Cognitive assessments
  - description, 446
  - digitalization, 452
  - multisensor digital pen interface case study, 445–453
- Cognitive attention capture in multimodal in-car interaction, 508–509
- Cognitive load
  - automobiles, 480
  - categories, 481–484
  - cognitive resource awareness, 178–179
- defined, 478
- driver distraction, 482–483
- multimodal in-car interaction, 502–503
- robotic applications, 549
- Cognitive resource awareness in SiAM-dp platform, 178–179
- Collaboration, defined, 192
- Collaborative dialogues
  - challenge discussion, 193–194
  - in domain-independent manner, 197–200
- Combined approaches for nonverbal behavior, 232
- Combined modalities in fusion, 157
- Commercialization
  - aeronautics, 621–625
  - assistants, 638–642
  - avatars, 642–643
  - biometrics, 636–638
  - emotion recognition, 650–651
  - field force automation, 646–647
  - focus questions, 651
  - glossary, 624–625
  - insurance, 647–648
  - introduction, 621
  - personal care products, 648–650
  - product search, 643–644
  - references, 652–658
  - robotics, 625–636
  - summary, 651
  - virtual and augmented reality, 644–646
- Commission of the European communities guidelines for driver distraction, 483
- Common ground in robots, 79, 93
- Communication
  - nonverbal behavior, 227–228
  - SAIBA framework functions, 228–229
  - situated interaction, 110
- W3C MMI Architecture, 378–380
- Communicative Functions (CFs)
  - Cerebella system, 237
  - fusion, 157, 159–162
- Communicative intentions in nonverbal behavior, 221

- Communicator project, 105  
 Companion product, 651  
 COMPI arm for robots, 532  
 Complementary modalities  
     defined, 158  
     dialogue platforms, 170  
     multimodal input, 157  
 Composite states in dialogue management  
     architectures, 155  
 Compound control modes for car controls, 491  
 Compound states in dialogue management  
     architectures, 155  
 Compound words in simultaneous  
     interpretation, 598–599  
 Computed muscle control (CMC) in  
     biomechanical simulation, 288  
 Computer-assisted surgery, 433  
 Computer vision (CVMC) in ergonomics  
     measurements, 272–274  
 Concept accuracy  
     defined, 26  
     finite-state approaches, 46  
 Concrete User Interface (CUI)  
     contextual adaptation, 362  
     declarative languages, 361  
     defined, 348  
 ConcurTaskTrees language, 359  
 Conditional Random Fields (CRFs)  
     gesture classifiers, 54  
     machine learning, 53  
 Connection events in multimodal fusion, 124  
 Consecutive interpretation  
     defined, 578  
     machine translation, 579–580, 591  
     mobile systems, 592–593  
     telephony, 594–596  
 Consent issues in embedded multimodal  
     interfaces, 562–565  
 Constant force for car controls, 490  
 Constraint-based motion optimization  
     movement modeling, 314, 316  
     after PCA, 318  
 Constraints on references, 163–164  
 Contact level of action in speaking, 93–94  
 Content selection and structuring in  
     presentation planning, 151–153  
 Context  
     automatic speech recognition, 583  
     multimodal reference resolution, 50  
 Context-free grammars, 43, 45  
 Context of interaction, self-adaptation to, 557–558  
 Contextual adaptation in declarative  
     languages, 361–363  
 Contextual cues in situated interaction, 112  
 Contextual integrity  
     defined, 663  
     privacy approaches on structured data, 673  
 Control models in nonverbal behavior, 226–228  
 Conversational adjacency pairs for  
     multimodal fusion, 124  
 Conversational systems  
     interactive. *See* Interactive conversational  
     systems  
     with robots. *See* Robots and robotics  
 Convolutional neural network (CNN)  
     architecture in movement  
     modeling, 321  
 Coordination in situated interaction, 110  
 Coping  
     defined, 222  
     Virtual Human Toolkit, 235  
 Core layers in dialogue systems, 147  
 Corruptibles privacy concerns, 677–679  
 Cortana VDA, 180  
 Cost functions in movement modeling, 314  
 Court rulings on privacy issues, 664–665  
 Covert measures  
     defined, 528  
     embedded multimodal interfaces, 535–  
     536  
     robotic applications, 544  
 CPS (cyber-physical system) controllers, 429  
 Cree.py application, 668

- CRFs (Conditional Random Fields)
  - gesture classifiers, 54
  - machine learning, 53
- CRMs (customer relationship management systems), privacy concerns, 682
- Cross-lingual communication in machine translation, 579–580
- Cross-lingual subtitling
  - defined, 578
  - machine translation, 580, 608
- Cross-modal reference resolution
  - reference constraints, 163–164
  - referring expressions, 162–163
- Cross-modal repair in multimodal translingual communication, 606
- Crowdsourcing, 231, 245
- CueMe platform, 169–170, 381
- Cues in turn-taking, 91
- CUI (Concrete User Interface)
  - contextual adaptation, 362
  - declarative languages, 361
  - defined, 348
- Cultural factors in mental state mapping, 224
- Customer relationship management systems (CRMs), privacy concerns, 682
- Cyber-physical environments
  - defined, 158
  - multimodal dialogue in, 173–174
- Cyber-physical system (CPS) controllers, 429
- Cybercasing
  - from data brokers, 679–682
  - defined, 663
  - privacy concerns, 661, 674
- CyberGlove (CG), 272
- da Vinci surgical robot, 628–629
- DAGs (Directed Acyclic Graphs)
  - defined, 27
  - multimodal reference resolution, 51–52
- DAMSL (Dialogue Act Markup using Several Layers), 161
- Dashboards in in-car interaction, 486
- Data analytics, privacy concerns, 660–661
- Data brokers, cybercasing from, 679–682
- Data collection
  - nonverbal behavior, 226–227
  - physical ergonomics, 269–278
- Data-driven approaches for nonverbal behavior, 230–231
- De-anonymization
  - defined, 663
  - privacy concern threats, 667
- De-identification
  - defined, 663
  - privacy concerns, 674
- De-identified database, privacy concerns, 667
- De-noising filters, privacy concerns, 674
- DEAP database
  - characteristics, 404
  - description, 403, 407
  - resource location, 406
- Decision fusion, 160
- Decision level in movement modeling, 308
- Decision-theoretic methodology
  - defined, 113
  - perception, 132, 135
- Declarative languages
  - Abstract User Interface level, 360
  - abstraction level relationships, 361
  - Concrete User Interface, 361
  - contextual adaptation, 361–363
  - digital ink, 364–368
  - emotions, 368–369
  - importance, 357–358
  - logical tasks, 359–360
  - model-based specifications, 358
  - voice standards, 363–364
- Decluttering of situational displays
  - defined, 624
  - F-35 helmet, 623
- Deep integration in embedded multimodal interfaces, 555–560
- Deep-learning models
  - defined, 307

- movement modeling, 320, 333–334
- Deep neural networks (DNNs)
  - autonomous cars, 626–627
  - machine learning, 56–57
- Degrees of freedom (DOF)
  - defined, 267
  - movement modeling, 314, 317
- Deictic expressions for referring expressions, 162
- Deictic gestures
  - defined, 79
  - speech, 82
- Deictic references
  - defined, 26
  - multimodal reference resolution, 50
- Deixis expressions for referring expressions, 162–163
- Delivery Context component
  - assistants, 640
  - machine translation, 580–581
- Dementia in multisensor digital pen interface case study, 445–453
- Demographics in multimodal database design, 418
- Dempster-Shafer Theory (DST)
  - defined, 624
  - F-35 helmet, 623
- DemTest cognitive assessment, 451, 456
- Depth-based systems for optical motion capture, 283
- Derived-from links for modal inputs, 371
- Design for ergonomics. *See* Ergonomics
- Destructive unification
  - defined, 26
  - multimodal reference resolution, 52
- Detents for car controls, 489–490
- Development databases, 394
- Dialogue context
  - defined, 113
  - situated interaction, 106
- Dialogue OS dialogue system, 166–167
- DialogFlow system
  - multimodal design guidelines, 639
  - natural language systems, 382
- DIALOGMASCHINE (DIANE) dialogue system, 166
- Dialogue act annotation for fusion, 161–162
- Dialogue Act Markup using Several Layers (DAMSL), 161
- Dialogue applications
  - defined, 158
  - software platforms, 146
- Dialogue architecture trends, 180–182
- Dialogue definition projects, 173
- Dialogue Flow module, 89
- Dialogue management
  - defined, 158
  - SiAM-dp platform, 172
- Dialogue management architectures
  - finite state-based, 154–155
  - frame state-based, 155–156
  - information state-based, 156
  - plan-based dialogue agents, 156
  - probabilistic, 156–157
  - software platforms, 154–157
- Dialogue platforms
  - AT&T speech mashup architecture, 164–165
  - CueMe, 169–170
  - defined, 158
  - Dialogue OS, 166–167
  - DIANE, 166
  - ODP, 168–169
  - SmartKom, 167–168
  - software platforms, 146
  - WAMI toolkit, 165–166
- Dialogue purposes, defined, 192
- Dialogue strategy module for SiAM-dp platform, 172
- Dialogue systems
  - architecture, 146–154
  - defined, 158
  - input/output processing layers, 148–149
  - meta dialogue management, 150–151
  - middleware, 154
  - multimodal fission and presentation planning, 151–153

- Dialogue systems (*continued*)
  - multimodal fusion and discourse processing, 149–150
  - SiAM-dp platform offline evaluation, 179–180
  - software platforms, 146
- DIANE (DIAlogmaschiNE) dialogue system, 166
- DIANEXML dialogue system, 166
- Differential privacy
  - defined, 663
  - privacy issues, 666–667, 671–672
- Digital ink, 364–368
- Digital pen interface case study
  - background, 446
  - lessons learned, 451–453
  - overview, 445–446
  - problem description, 446–448
  - solution, 448–451
- Dimensions in movement modeling, 335–336
- Direct health risk estimation models, 278–280
- Direct measurement methods for physical ergonomics, 272–278
- Direct touch selection concept for in-car interaction, 509
- Directed Acyclic Graphs (DAGs)
  - defined, 27
  - multimodal reference resolution, 51–52
- Directed gaze events, 124
- Directions Robot system, 134
- Discomfort questionnaire methods, 270–271
- Discourse
  - defined, 158
  - dialogue systems, 149–150
- Disfluencies in spoken language translation, 579
- Distant speech input in translingual communication, 609–610
- Distribution
  - multimodal database design, 416–417
- W3C multimodal standards, 352
- DIT++ scheme, 161–162
- DMP (dynamic movement primitive) systems
  - defined, 307
  - movement modeling, 315–316, 320, 325–326, 337
- DMs (dynamometers), 273
- DNNs (deep neural networks)
  - autonomous cars, 626–627
  - machine learning, 56–57
- Document Object Model (DOM), 362–363
- DOF (degrees of freedom)
  - defined, 267
  - movement modeling, 314, 317
- DOM (Document Object Model), 362–363
- Domain-limited research systems in machine translation, 588–591
- Drawings for interactive conversational systems, 31
- DRIVAWORK simulator, 408
- Driver distraction cognitive load actor, 482–483
- Driver seat in in-car interaction, 491–492
- DROCK framework, 559
- Dropout
  - defined, 307
  - movement modeling, 332
- DST (Dempster-Shafer Theory)
  - defined, 624
  - F-35 helmet, 623
- DTask dialogue planner, 238–239
- Ducking in machine translation, 595
- Dutch Musculoskeletal Survey, 270
- Dyadic dialogues, defined, 192
- Dyadic interactions
  - defined, 113
  - situated interaction, 111
- Dynamic aspects in musculoskeletal models, 284
- Dynamic movement primitive (DMP) systems
  - defined, 307

- movement modeling, 315–316, 320, 325–326, 337
- Dynamic programming approach
  - defined, 26–27
  - machine learning, 58
- Dynamic Time Warping, 58
- Dynamometers (DMs), 273
- Dynaspeak in F-35 helmet, 623
- Early fusion
  - defined, 307
  - dialogue systems, 149
  - movement modeling, 328–333
  - overview, 159–161
- Early integration for movement modeling.
  - See* Movement modeling in latent spaces
- Earplugs
  - defined, 578
  - multimodal translational communication, 612
- ECAs. *See* Embodied Conversational Agents (ECAs)
- ECG (electrocardiography), 273
- Echo cancellation
  - defined, 79
  - human-robot interaction, 85
- Echo Show assistant, 638–639
- EDA (electrodermal activity), 273
  - multimodal-multisensor framework case study, 453–458
  - non-clinical medical and health systems, 437
- Edit machines
  - defined, 27
  - finite-state approaches, 46
- Edit transducers
  - defined, 27
  - finite-state approaches, 46
- Education, privacy concerns in, 686–692
- EEG. *See* Electroencephalography (EEG)
- Effective communication in machine translation, 580–581
- Efficiency in late fusion movement modeling, 311
- EG (electronic goniometry) for physical ergonomics measurements, 272
- EHRs (electronic health records)
  - interaction logs in non-clinical medical and health systems, 435
  - patterns in, 425
  - speech-based answering systems, 434
- EJA (Ensure Joint Attention) for robots, 96
- ELAN tool, 128
- Electrical processes in physical ergonomics measurements, 273
- Electrocardiography (ECG), 273
- Electrodermal activity (EDA), 273
  - multimodal-multisensor framework case study, 453–458
  - non-clinical medical and health systems, 437
- Electroencephalography (EEG)
  - BCI, 536
  - defined, 395
  - embedded multimodal interfaces, 539
  - exoskeleton mode changes, 567–568
  - exoskeletons, 546–547, 552–555
  - movement modeling. *See* Movement modeling in latent spaces
  - physical ergonomics measurements, 273
  - robotic applications, 544, 546
- Electromagnetic (EMMC), 272
- Electromyography (EMG)
  - defined, 267, 578
  - movement modeling. *See* Movement modeling in latent spaces
  - physical ergonomics measurements, 273, 276–277
  - robotic applications, 552–555
- Electronic Form-Filling in ink standard, 366
- Electronic goniometry (EG) for physical ergonomics measurements, 272
- Electronic health records (EHRs)
  - interaction logs in non-clinical medical and health systems, 435

- Electronic health records (EHRs) (*continued*)  
 patterns in, 425  
 speech-based answering systems, 434
- Electronic medical records (EMRs)  
 defined, 427  
 patterns in, 425
- Eliza text-based dialogue systems, 105
- Ellipsis expressions, 163
- ELUs (exponential linear units) in movement modeling, 336
- Embedded brain readings  
 defined, 528  
 example, 539  
 exoskeletons, 544, 565–569  
 limitations, 540  
 robotic applications, 550–552  
 spontaneously evoked activity, 541–542
- Embedded multimodal interfaces  
 consent and data privacy issues, 562–565  
 defined, 528  
 definition and relevance, 534–542  
 exoskeleton mode changes, 565–569  
 focus questions, 569–570  
 future trends, 554–565  
 glossary, 528–530  
 human-robot interaction, 527  
 inherent self-adaptation and deep integration, 555–560  
 inherently safe robots, 527, 531–534  
 introduction, 523–527  
 references, 570–576  
 robotic applications, 542–555  
 societal implications, 560–565  
 trust issues, 560–562
- Emblem gestures  
 defined, 79  
 speech, 82
- Embodied Cognition paradigm for nonverbal behavior, 224–225
- Embodied Conversational Agents (ECAs)  
 BEAT, 240  
 defined, 222  
 faces for, 82  
 LiteBody tool, 238
- nonverbal behavior, 226  
 SAIBA framework, 228–229  
 VIB, 241
- Embodiment in nonverbal behavior, 220–225
- Emerging trends and applications  
 overview, 3–4  
 preview, 11–18
- EMF (Eclipse Modeling Framework), 172–173
- EMG. *See* Electromyography (EMG)
- EMMA. *See* Extensible Multimodal Annotation (EMMA)
- EMMC (electromagnetic), 272
- EmoTABOO, 408
- Emotion, faces for, 81
- Emotion Markup Language (EmotionML), 356–357  
 defined, 348  
 description, 354  
 implementations, 385  
 nonverbal behavior, 226–227  
 overview, 368–369  
 vocabularies, 384
- Emotion recognition systems  
 commercialization, 650–651  
 Mabu robot, 633
- EmoTV Corpus database  
 characteristics, 404  
 description, 407  
 naturalistic datasets, 408  
 resource location, 406
- Empathy  
 defined, 222  
 nonverbal behavior, 225
- Employee monitoring, 564
- EMRs (electronic medical records)  
 defined, 427  
 patterns in, 425
- Encapsulation  
 clinical medical and health systems, 431  
 W3C multimodal standards, 352
- Encoder-decoder models in machine translation, 585–586

- Encoding dictionaries
  - defined, 113
  - machine learning approaches, 126
- End-to-end learning in movement
  - modeling, 334
- Engagement
  - defined, 113
  - machine learning approaches, 127
  - situated interaction, 110
- Engagement in social interaction
  - overview, 116–117
  - perception, 119, 121–135
  - representation, 117–121
- Ensure Joint Attention (EJA) for robots, 96
- Enterprise Virtual Assistant (EVA), 640
  - conversational systems, 24
  - graphical and spoken input, 350
- Environmental acoustic noise, privacy concerns for, 675
- Equivalence
  - defined, 158
  - multimodal input, 157
- Ergonomics
  - defined, 267
  - focus questions, 291
  - generic design process, 266, 268
  - glossary, 267
  - introduction, 265–266
  - motion capture-based biomechanical simulation, 282–288
  - physical. *See* Physical ergonomics
  - references, 291–304
  - summary and future research directions, 288–290
- ERP (evoked-cortical potential), 536
- Errors and error handling
  - biometrics, 637
  - machine translation, 581
  - multimodal translingual communication, 604–607
  - spoken language translation, 579
- eTFSs (extended Typed Feature Structures)
  - in ODP, 169
- Ethical issues
  - data collection, 226, 412
  - database design, 415
  - embedded multimodal interfaces, 564–565
- EU-BRIDGE program, 596, 601–602
- European Parliament, simultaneous interpretation in, 602–603
- EVA Café, 640–641
- EVA (Enterprise Virtual Assistant), 640
  - conversational systems, 24
  - graphical and spoken input, 350
- Event logic in incremental multimodal integration, 47–50
- Event processing in W3C MMI Architecture, 378–380
- Everyday emotion vocabulary, 369
- Evidential reasoning algorithms for autonomous cars, 626
- Evoked-cortical potential (ERP), 536
- Exclusion
  - defined, 663
  - privacy issues, 666
- Exophoric references, 163
- Exoskeletons
  - embedded multimodal interfaces, 540–541
  - mode changes, 565–569
  - robotic applications, 544–556
- Exotic sensors, privacy concerns, 676
- Expert-designed features in movement modeling, 318–319
- Expert Mode in vocal navigation systems, 506–507
- Explicit control
  - defined, 529
  - human-robot interaction, 524
- Explicit interaction
  - defined, 529
  - between humans and robots, 523–524
- Exponential linear units (ELUs) in movement modeling, 336
- Exposure-effect models, 280
- Expressivity parameters in VIB, 244

- Extendable systems in machine translation, 581
- Extended Kalman Filters
  - autonomous cars, 626
  - defined, 625
- extended Typed Feature Structures (eTFSs)
  - in ODP, 169
- Extensibility in W3C multimodal standards, 352
- eXtensible Markup Language (XML)
  - AT&T speech mashup architecture, 165
  - BEAT, 239
  - BML. *See* Behavior Markup Language (BML)
  - child content, 349
  - container elements, 349
  - defined, 349
  - description, 357–358
  - DIANFXML dialogue system, 166
  - GRXML and SCXML, 173
  - multimodal user input, 370
  - voice, 363
  - XML-based pipeline, 239
- Extensible Multimodal Annotation (EMMA)
  - AT&T speech mashup architecture, 165
  - assistants, 640
  - CueMe platform, 169–170
  - defined, 348
  - description, 354
  - EmotionML, 369
  - fusion approach, 370–372
  - implementations, 384
  - InkML, 365
  - multimodal fission and media synchronization, 372–377
  - multimodal inputs, 355–356
- Extension notification in W3C MMI
  - Architecture, 380–381
- Extraneous cognitive load, 482
- Eye contact in interaction, 81
- Eye tracking
  - medical and health systems, 450
  - robotic applications, 549
  - virtual reality, 460–461
- F-35 helmet, 622–623
- F-formations
  - defined, 113
  - situated interaction, 110
- Facebook wit.ai, 382
- Faces and face detection
  - human-robot interaction, 87–88
  - in interaction, 80–82
  - privacy concerns, 675
  - robots, 82–83
  - spoofing, 637
- Facial expression analysis in medical and health systems, 450
- Facial feature tracking in human-robot interaction, 88
- Facilitator agents in OAA, 351
- FACS schema
  - nonverbal behavior, 226–227
  - SEMAINE database, 410
- Feature engineering in machine learning, 126
- Feature level
  - early fusion, 159–160
  - movement modeling, 308
- Feature space in movement modeling, 317
- Feature structures
  - defined, 27
  - unification-based multimodal fusion, 33–39
- Federal Trade Commission calls for privacy, 662
- Feedback
  - exoskeletons, 546
  - mental states, 225
  - multimodal integration, 47, 49
  - robots, 78, 93–95
  - surgical robots, 629
- Field-adaptable systems in machine translation, 581
- Field force automation, 646–647
- Field Programmable Gate Arrays (FPGAs), 559
- Fingerprint readers, spoofing, 637
- Finite-state acceptors, defined, 27

- Finite-state approaches for multimodal grammars, 41–47
- Finite-state automatons, defined, 27
- Finite state-based dialogue management architectures, 154–155
- Finite-state composition
  - defined, 28
  - multimodal grammars, 43–46
  - multimodal integration, 63
- Finite-state multimodal understanding mechanism
  - destructive unification for contextual resolution, 64–66
  - finite-state language processing cascade on multimodal inputs, 64
  - representing specific content, 63–64
  - simulating 3-tape finite-state device with transducers, 62–63
- Finite-state transducers
  - defined, 28
  - multimodal grammars, 44–45
- Firmware Version 7.0 for multimodal in-car interaction, 494
- Fitness portals, 434
- Flexi-modal communication, 607–613
- Flipper library, 235–237
- Floors in in-car interaction, 486
- FMA (Foundational Model of Anatomy) ontology, 439–440
- FML
  - SAIBA framework, 230
  - VIB, 243–244
- FML-APML (FML-Affective Presentation Markup Language), 233–234, 242
- Force-controlled robots, 532–533
- Force sensors (FS) in physical ergonomics measurements, 273
- Force transducers (FT), 272
- Forces on exoskeletons, 540
- Forecasting
  - defined, 115
  - machine learning approaches, 128–129
- Foundational Model of Anatomy (FMA) ontology, 439–440
- Foundational technologies
  - clinical systems, 429
  - defined, 427
- FPGAs (Field Programmable Gate Arrays), 559
- Frame-based systems
  - finite state charts, 155–156
  - mixed initiative dialogue, 192
- Frames
  - defined, 28
  - multimodal fusion, 33, 39–41
- Framework for Protecting Privacy, 665
- Frameworks in OAA, 350
- FS (force sensors) in physical ergonomics measurements, 272–273
- Full user support mode
  - exoskeletons, 546
  - VI-Bot exoskeleton, 565–569
- Function Markup Language, defined, 222
- Furhat robot, 83–85, 633
- Fused input in semantics, 371
- Fusion
  - defined, 307
  - dialogue systems, 149
  - movement modeling, 308–311, 328–333
  - overview, 159–161
- Fusion and communicative functions
  - dialogue act annotation, 161–162
  - fusion level, 159–161
  - multimodal input, 157
  - software platforms, 157, 159–162
  - temporal relationship and synchronization, 157, 159
- Fusion and discourse processing module in SiAM-dp platform, 171–172
- Galaxy Communicator, 378
- Galaxy phones, 639
- Gas pedal in in-car interaction, 491
- Gaussian process dynamical models (GPDMs)
  - defined, 307
  - movement modeling, 318, 320, 324

- Gaussian process latent variable models (GPLVMs)
  - defined, 307
  - movement modeling, 318–320, 324
- Gaussian processes (GPs), defined, 307
- Gaze
  - faces for, 81
  - human-robot interaction, 88
  - joint attention, 95–96
  - medical and health systems, 450–451
  - multimodal fusion, 124
  - turn-taking, 91–92
- GDPR (General Data Protection Regulation), 562–564
- GECA (Generic Embodied Conversational Agent), 241
- GECA Scenario Mark-up Language (GSML), 241
- GEMEP (GEneva Multimodal Emotion Portrayals)
  - characteristics, 404
  - description, 407–408
  - resource location, 406
- General Data Protection Regulation (GDPR), 562–564
- Generalization in machine learning approaches, 129–131
- Generators in SiAM-dp platform, 172
- Generic Embodied Conversational Agent (GECA), 241
- GEneva Multimodal Emotion Portrayals (GEMEP)
  - characteristics, 404
  - description, 407–408
  - resource location, 406
- GenieMD, 434
- Geo-location information privacy concerns, 668, 685
- Geotagging privacy issues, 666
- German language in simultaneous interpretation, 598–599
- Germane cognitive load, 482
- Gesticons in SAIBA framework, 230
- Gesture phases
  - defined, 222
- nonverbal behavior, 246
- VIB, 244
- Gesture recognition
  - defined, 28
  - virtual reality, 460–461
- Gestures
  - defined, 28
  - finite-state approaches, 45–46
  - incremental multimodal integration, 48
  - interactive conversational systems, 24, 31
  - machine learning, 54
  - multimodal fusion, 33
  - multimodal grammars, 43
  - multimodal in-car interaction, 495–497
  - multimodal reference resolution, 51
  - nonverbal behavior, 232
  - speech, 81–82
  - unification-based multimodal fusion, 34–36, 39
- Goals for multimodal database design, 417
- Goggles for speech translation, 611–612
- Goniometers
  - defined, 267
  - ergonomics, 269, 272
- Google Assistant VDA, 180
- Google Dialogflow, 180
- Google Glass, 612–613
- Google Home devices, 436, 633
- Google Pixel Buds, 612
- Gorilla arm effect, 289
- GPDMs (Gaussian process dynamical models)
  - defined, 307
  - movement modeling, 318, 320, 324
- GPLVMs (Gaussian process latent variable models)
  - defined, 307
  - movement modeling, 318–320, 324
- GPs (Gaussian processes), defined, 307
- Grammar XML (GRXML), 173
- Graphical user interfaces (GUIs) for automobiles, 480
- Graphs for multimodal reference resolution, 51–52
- Gravity compensation for exoskeletons, 540

- Green Persuasive Dataset, 408  
 Greta platform, 233  
**Grounding**  
 defined, 79  
 robots, 78, 93–95  
 situated interaction, 110–111  
**GRXML (Grammar XML)**, 173  
**GSML (GECA Scenario Mark-up Language)**, 241  
**GUIs (graphical user interfaces) for automobiles**, 480
- Hair Coach product, 649  
**Hairbrushes**, 648–649  
**Hand gestures in virtual reality**, 460–461  
**Hands, bionic**, 630  
**Handwriting**  
 multimodal in-car interaction, 497–498  
 multimodal-multisensor framework case study, 453–458  
 multimodal translational communication, 609  
 virtual reality, 460–461
- Haptic feedback**  
 hairbrushes, 648  
 prosthetic arms, 630  
 surgical robots, 628–629  
 wearable devices, 173
- Haptic objects and controls**  
 mental mapping, 479  
 multimodal in-car interaction, 487–495  
 robots, 531  
 touch screens, 493  
 virtual reality, 461–462
- Hard sensors in CueMe platform, 170  
**Hardware solutions for embedded multimodal interfaces**, 558–560
- Harel statecharts, 80  
**Head-down-displays (HDDs) in multimodal in-car interaction**, 508
- Head-pose tracking in human-robot interaction**, 88
- Head-up-displays (HUDs) in multimodal in-car interaction**, 508–510
- Headsets for virtual and augmented reality, 644–646  
**Health systems. *See* Medical and health systems**  
**Heart rate (HR) in physical ergonomics measurements**, 273  
**Helmets, F-35**, 622–623  
**HeroSurg surgical robot**, 629  
**Heuristics, multimodal fusion via**, 123–125  
**Hidden cameras, privacy issues for**, 665  
**Hidden Markov Models (HMMs)**  
 machine learning, 53  
 nonverbal behavior, 245  
**Hidden space in movement modeling**, 317  
**Hierarchical description in logical tasks**, 359  
**High confidence medical device software development**, 429  
**High-level control**  
 defined, 529  
 inherently safe robots, 533  
**High-school students privacy concerns**, 687  
**Hill-type models for musculoskeletal models**, 284  
**Hills for car controls**, 490–491  
**HMC (human motion capture)**, 272  
**HMI (human-machine interfaces) in cars. *See* Automotive multimodal human-machine interface**  
**HMMs (Hidden Markov Models)**  
 machine learning, 53  
 nonverbal behavior, 245  
**Holding turns in human-robot interaction**, 91  
**Hololens 2 headset**, 644  
**Home scenario in SMARTKOM project**, 411  
**Hover and click concept in multimodal in-car interaction**, 509  
**HR (heart rate) in physical ergonomics measurements**, 273  
**HTML (Hypertext Markup Language)**  
 defined, 348  
 multimodal systems, 350  
**HTTP (Hypertext Transfer Protocol)**  
 for compatibility, 350, 352

- HTTP (Hypertext Transfer Protocol)
  - (*continued*)
  - defined, 348
- HUDs (head-up-displays) in multimodal in-car interaction, 508–510
- HUMAINE database
  - characteristics, 404
  - description, 407–408
  - resource location, 406
- Human-computer interaction. *See* Ergonomics
- Human-Computer SAL data, 408
- Human-machine interfaces (HMIs) in cars. *See* Automotive multimodal human-machine interface
- Human motion capture (HMC), 272
- Human neural system in ergonomics measurements, 276
- Human-robot cooperation
  - challenges and opportunities, 523
  - defined, 529
- Human-robot interaction
  - auditory channel, 84–87
  - defined, 529
  - embedded multimodal interfaces, 523–527
  - overview, 83–84
  - situation modeling, 88–89
  - visual channel, 87–88
- Hybrid BCIs
  - defined, 529
  - embedded multimodal interfaces, 537–538
- Hybrid dialogue reasoning, 195–196
- Hypertext Markup Language (HTML)
  - defined, 348
  - multimodal systems, 350
- Hypertext Transfer Protocol (HTTP)
  - for compatibility, 350, 352
  - defined, 348
- Hypothesis testing databases, 394
- iADLs (instrumental activities of daily living), 436
- IBM Watson Conversations, 180–181
- Iconic gestures
  - defined, 79
  - speech, 82
- Ideational units
  - defined, 222
  - nonverbal behavior, 246
- Identification
  - defined, 663
  - privacy concerns, 665, 674
- Identifiers
  - defined, 663
  - privacy concerns, 667, 674
- Identity, faces for, 80
- iEMG (intramuscular electromyography), 273, 276–277
- IJA (Initiate Joint Attention) with robots, 96
- IM (Interaction Manager), 351
  - assistants, 640–641
  - defined, 348
  - purpose, 353
- Image semantics in medical and health systems, 439
- Imitation
  - defined, 222
  - nonverbal behavior, 225
- Imitation learning
  - defined, 529
  - robotic systems, 559
- Immersion TouchSense PR-1000 device, 488–489
- Implicit interaction
  - defined, 529
  - between humans and robots, 523–524
- Improvisation in self-adaptation, 557
- Impulsion Social AI engine, 240
- iMRK interface, 543
- IMU (inertial measurement unit)-based systems for optical motion capture, 283
- IMUMC (inertial motion unit motion capture), 272
- In the wild research, 394, 396

- Increased accessibility
  - defined, 664
  - privacy issues, 666
- Incremental interaction management, 25, 29
- Incremental multimodal integration in event logic and visual state charts, 47–50
- Incremental processing, defined, 29
- Incremental unfolding in nonverbal behavior, 247
- Independent modalities in fusion, 157
- Industrial design
  - defined, 267
  - interaction design, 265
- Inertia matrices
  - defined, 267
  - musculoskeletal models, 284
- Inertial measurement unit (IMU)-based systems for optical motion capture, 283
- Inertial motion unit motion capture (IMUMC), 272
- Inference chains, privacy concerns, 677–679
- Inference models in machine learning approaches, 129
- Information collection
  - defined, 664
  - privacy issues, 665
- Information dissemination
  - defined, 664
  - privacy issues, 666
- Information leaks, privacy concerns, 685
- Information processing
  - defined, 664
  - privacy issues, 665
- Information state-based dialogue management architectures, 156
- Informed consent issues, 562–565
- Infotainment systems
  - defined, 478
  - interaction with, 479
- Inherent self-adaptation, 555–560
- Inherently safe robots, 527, 531–534
- Initiate Joint Attention (IJA) with robots, 96
- Initiative, defined, 192
- Ink Archiving and Retrieval standard, 366
- Ink input for interactive conversational systems, 23–24
- Ink Markup Language (InkML)
  - defined, 348
  - description, 354
  - digital ink, 364–368
- Ink Synchronized Multimedia Integration Language Messaging, 365
- Input in machine translation, 581
- Input/output processing layers in dialogue systems, 148–149
- Input space in movement modeling, 317
- Instrumental activities of daily living (iADLs), 436
- Instrumented treadmills (IT), 273
- Insurance, 647–648
- Integration for interactive conversational systems. *See* Interactive conversational systems
- Intelligent tutoring systems for medical and health systems, 459
- Intentions in engagement, 117–118
- Inter-pausal units (IPUs), 90
- Interact system, 23
- Interaction Manager (IM), 351
  - assistants, 640–641
  - defined, 348
  - purpose, 353
- Interaction space
  - automobiles, 479
  - defined, 478
- Interaction style in machine translation, 579–580
- Interaction tasks in engagement in social interaction, 117
- Interactive conversational systems
  - conclusion, 59–62
  - early approaches to multimodal fusion, 32–33
  - finite-state multimodal understanding mechanism, 62–66

- Interactive conversational systems  
*(continued)*  
 focus questions, 66–67  
 glossary, 26–30  
 incremental multimodal integration, 47–50  
 introduction, 23–25  
 machine learning for multimodal integration, 53–59  
 motivations for multimodal input, 25, 31–32  
 multimodal grammars and finite-state approaches, 41–47  
 multimodal reference resolution and multimodal dialogue, 50–53  
 references, 67–76  
 unification-based multimodal fusion, 33–41
- Interactive unfolding in nonverbal behavior, 247
- Interactivity in machine learning approaches, 131
- Interakt project, 446
- Internal forces in ergonomics measurements, 275–276
- Internal states of robot  
 defined, 529  
 inherently safe robots, 532
- International Classification of Diseases (ICD-10), 440
- Internet-of-things (IoT) devices for health systems, 437
- Internet-scale multimedia retrieval capabilities, privacy concerns, 677
- Interpreter's Cruise Control, 603
- Interpreters for SiAM-dp platform, 172
- Interstimulus intervals (ISIs) in robotic applications, 550–552
- Intramuscular electromyography (iEMG), 273, 276–277
- Intrinsic cognitive load, 481–482
- Invasions  
 defined, 664  
 privacy issues, 665
- Inverse dynamics  
 biomechanical simulation, 286–287  
 defined, 267
- Inverse kinematics  
 biomechanical simulation, 286  
 defined, 267
- Invisibility in vocal navigation systems, 503
- IoT (Internet-of-things) devices for health systems, 437
- iPal robot, 632–633
- IPUs (inter-pausal units), 90
- IrisFlow dialogue model, 89
- IrisTK card-sorting game, 84–85
- ISIs (interstimulus intervals) in robotic applications, 550–552
- ISO DIS 24617–2, 226–227
- IT (instrumented treadmills), 273
- JANUS systems  
 defined, 578  
 first demonstrations, 588  
 machine translation, 590
- Jargon in simultaneous interpretation, 599–600
- Java Speech Grammar Format (JSGF), 165
- JavaScript Object Notation (JSON) format, 165
- JavaScript programming language  
 defined, 348  
 multimodal fission, 373
- Jibbigo machine translation system, 592–594
- Jibo robot, 631
- Joint attention  
 defined, 79  
 robots, 78, 95–97
- Joint inference  
 defined, 113  
 machine learning approaches, 129
- Joint spatial orientations in situated interaction, 110
- Joint Strike Fighter, 621–625
- JSGF (Java Speech Grammar Format), 165
- JSON (JavaScript Object Notation) format, 165

- k-anonymity
  - defined, 664
  - privacy concerns, 669–671
- Kalman filtering
  - autonomous cars, 626
  - defined, 625
- Karlsruhe Institute of Technology (KIT),
  - simultaneous interpretation at, 596–597
- Kinect sensor in human-robot interaction, 85, 87
- Kinematic aspects in musculoskeletal models, 284
- Kinematics measurements, 273–275
- Kismet robot head, 82–83
- Knowledge acquisition in situation-adaptive behavior, 176
- Kognit project, 446, 459–460
- Labeling data in machine learning approaches, 127–128
- Lane Departure Warning System, 492
- Languages
  - declarative. *See* Declarative languages
  - markup. *See* Markup languages
- Late fusion
  - defined, 307
  - movement modeling, 308–311, 321–322
  - overview, 159–161
- Latent representation
  - defined, 308
  - movement modeling, 321
- Latent spaces, movement modeling in.
  - See* Movement modeling in latent spaces
- Lattices
  - defined, 29
  - finite-state approaches, 46
  - multimodal grammars, 41–43
- Leakage in nonverbal behavior, 221–223
- Learning methods in embedded multimodal interfaces, 558–559
- Learning words in multimodal translingual communication, 606–607
- Lecture Translator, 601–602
- Lectures
  - multimodal translingual communication, 609–610
  - simultaneous interpretation, 595–599, 601–602
- Leggings, 649–650
- Levels of action in robots, 93–94
- Lexicons in VIB, 243–244
- Lidar (Light Detection and Ranging) in autonomous cars, 625–626
- Life Cycle Events
  - EMMA, 356–357
  - MCs, 353
- LifeModeler software, 288
- Light Detection and Ranging (Lidar) in autonomous cars, 625–626
- Light Head robot, 83
- Likert scales in discomfort questionnaires, 270
- Linguistic scalability/portability
  - defined, 578
  - simultaneous interpretation, 601
- Linkage attacks
  - defined, 664
  - privacy concern threats, 667
- Linked Open Drug Data (LODD), 431
- LiteBody tool, 238–239
- Liveness measures in biometrics, 638
- LMM (Lumbar Motion Monitor), 272
- Location estimation, privacy concerns, 675
- Lockout mechanisms in exoskeletons, 546, 567
- LODD (Linked Open Drug Data), 431
- Logical tasks in declarative languages, 359–360
- Logics, privacy concerns, 673
- Logistics in multimodal database design, 413
- Low-level control
  - defined, 529
  - inherently safe robots, 532
- LUIS VDA, 180
- LUKE arm, 630
- Lumbar Motion Monitor (LMM), 272

- M3L (Multimodal Markup Language), 167
- Mabu robot, 632–633, 636
- Machina Arte Robotum Simulans (MARS) software, 549
- Machine learning
  - challenge discussion, 205–208
  - machine translation, 587–588
  - multimodal integration, 53–59
  - nonverbal behavior, 232, 245
  - perception, 125–131
  - privacy concerns, 685
  - self-adaptation, 557
- Machine translation
  - automatic speech recognition, 582–583
  - conclusion, 613–614
  - consecutive interpreting telephony, 594–596
  - deployments and services, 591–603
  - finite-state approaches, 44
  - first demonstrations, 588
  - focus questions, 614–615
  - glossary, 578
  - introduction, 577, 579–581
  - mobile consecutive interpretation systems, 592–593
  - multimodal translingual communication, 604–612
  - overview, 583–587
  - references, 615–620
  - research systems and prototypes, 588–591
  - simultaneous interpretation, 595–603
  - system prototypes and deployments, 588–603
  - technology, 581–588
- Machines with agendas, 561
- Macro gestures in multimodal in-car interaction, 495–496
- Magic Leap One headset, 645
- MAHNOB-HCI database
  - characteristics, 404
  - description, 408
  - resource location, 406
- MAHNOB-Laughter database
  - characteristics, 404
- description, 408
- resource location, 406
- Main Application Server, 164
- Main platform component in SiAM-dp platform, 171
- Mapping mental state for nonverbal behavior, 221, 223–224
- MapTask corpus, 54
- MARC (Multimodal Affective and Reactive Characters toolkit), 241
- Marker adjustment in biomechanical simulation, 285–286
- Markerless computer vision systems, 283
- Markerless motion capture for ergonomics, 290
- Markup languages
  - conclusion, 385
  - declarative. *See* Declarative languages
  - defined, 348
  - focus questions, 385–386
  - future of standardized representations, 381–383
  - glossary, 348–349
  - implementations, 384–385
  - introduction, 347, 349–351
  - lessons learned, 377–381
  - modality fusion and media synchronization, 369–372
  - model-based specifications, 358–369
  - multimodal fission and media synchronization, 372–377
  - open challenges, 383–384
  - references, 386–392
  - standards complements, 353–357
  - W3C MMI standards-based implementation, 351–353
- XML. *See* eXtensible Markup Language (XML)
- MARS (Machina Arte Robotum Simulans) software, 549
- Mary TTS system implementations, 385
- Match score level in movement modeling, 308
- MATCH system
  - finite-state approaches, 46

- late fusion, 160
- multimodal grammars, 42
- prototype system, 23
- Math Data Corpus database**
  - characteristics, 404
  - description, 409
  - resource location, 406
- Max 8 aircraft crashes, 623–624
- Maximum Entropy (MaxEnt) approach**
  - gesture classifiers, 54
  - machine learning, 53
- Maximum voluntary contraction (MVC), 276
- McGurk effect, 81
- MCI (Mild Cognitive Impairment), 458
- MCPS (medical cyber-physical systems)**
  - defined, 427
  - overview, 429–431
- MCs (Modality Components)**
  - defined, 348
  - purpose, 353
- Mean Pupil Diameter Change (MPDC) in cognitive load, 484–485
- Mechanical motion capture (MMC)**, 272, 276
- Mechanical processes in physical ergonomics measurements**, 272–278
- MEDCAP (Medical Civil Action Program)**, 593
- Media synchronization**
  - markup languages, 369–372
  - and multimodal fission, 372–377
- Medical and health systems**
  - clinical systems, 429–433
  - conclusion, 462–464
  - focus questions, 464
  - future directions, 458–462
  - glossary, 427–428
  - introduction, 425–426
  - multimodal dialogue system case study, 438–445
  - multimodal-multisensor combinations, 459–460
  - multimodal-multisensor framework case study, 453–458
- multisensor digital pen interface case study**, 445–453
- non-clinical systems**, 433–438
- references**, 465–476
- virtual reality**, 460–462
- Medical Civil Action Program (MEDCAP)**, 593
- Medical cyber-physical systems (MCPS)**
  - defined, 427
  - overview, 429–431
- Medical decision support systems**
  - defined, 428
  - virtual reality, 460
- Medium for information storage**, 355
- Medium values in modal inputs**, 371
- Member-Team-Committee (MTC)**
  - architecture, 55–56
- Mental models**
  - automobiles, 480
  - defined, 478
- Mental states in nonverbal behavior**
  - feedback, 225
  - mapping, 221, 223–224
- Meta dialogue management**
  - defined, 158
  - overview, 150–151
- Metaphoric gesture in nonverbal behavior**, 224–225
- Metaphors in nonverbal behavior**, 224–225
- Metria Innovation system**, 283
- MHAD (Multimodal Human Action Database)**
  - characteristics, 404
  - description, 409
  - resource location, 406
- mHealth**
  - defined, 428
  - health systems, 431
- MHI-Mimicry database**
  - characteristics, 404
  - description, 409
  - resource location, 406
- Micro gestures in multimodal in-car interaction**, 495–496

- Microphone arrays
  - defined, 79
  - human-robot interaction, 84–85
- Microphones
  - echo cancellation, 79
  - embedded multimodal interfaces, 534–535
  - hairbrushes, 648
  - multimodal translingual communication, 609–610
  - simultaneous interpretation, 601
- Microsoft Kinect system
  - defined, 395
  - optical motion capture, 283
- Middleware in dialogue systems, 154
- Mild Alzheimer Disease, 458
- Mild Cognitive Impairment (MCI), 458
- Mimicry
  - defined, 222
  - nonverbal behavior, 225
- Minds
  - nonverbal behavior, 220–225
  - SAIBA framework, 228–229
- MiroSurge surgical interface, 433
- Mirroring
  - defined, 223
  - nonverbal behavior, 225
- Mitigation research in privacy concerns, 685–686
- Mixed-initiative systems
  - defined, 158
  - dialogue systems, 150–151
  - engagement in social interaction, 116
- Mixed initiatives, defined, 114
- Mixed latent representation in movement modeling, 330–331
- MMC (mechanical motion capture), 272, 276
- MMGL (multimodal grammar model), 58
- MMI (Multimodal Architecture)
  - defined, 348
  - implementations, 384
- MMIs. *See* Multimodal, Multisensor Interfaces (MMIs)
- MMSE cognitive assessment, 451, 456
- Mobile consecutive interpretation systems, 592–593
- Mobile medical sensor architecture, 437
- Mobile multimodal system, 55
- Mobile scenario in SMARTKOM project, 411
- MoCA cognitive assessment, 451
- Modal inputs, modality information from, 371
- Modalities of interaction
  - defined, 192
  - human-computer interaction, 191
- Modality, 478
- Modality Components (MCs)
  - defined, 348
  - purpose, 353
- Modality fission, defined, 158
- Modality fusion
  - defined, 158
  - markup languages, 369–372
- Modality information from modal inputs, 371
- Modality selection in presentation planning, 151–153
- ModDrop method in movement modeling, 331–332
- Mode values in modal inputs, 371
- Model scaling in biomechanical simulation, 285
- Modularity in W3C multimodal standards, 352
- Mona Lisa effect
  - defined, 79
  - faces for, 82–83
- Morphology in finite-state approaches, 44
- Motion capture-based biomechanical simulation
  - musculoskeletal models, 284–285
  - optical motion capture, 283–284
  - overview, 282–283, 285–288
- Motion Capture (Mo-Cap) data
  - movement modeling, 330
  - nonverbal behavior, 231, 245
  - physical ergonomics measurements, 274

- Motion graphs in movement modeling, 314, 316
- Motivations for multimodal input, 25, 31–32
- Movement modeling in latent spaces
  - application, 306, 308
  - basic principles, 313–315
  - benefits, 310–312
  - conclusion, 339–340
  - early multimodal integration, 322–334
  - focus questions, 340
  - glossary, 307–308
  - introduction, 305–306
  - late multimodal fusion, 321–322
  - references, 340–345
  - sensor fusion levels, 308–310
  - spaces, 315–317
  - state of the art, 313–322
  - substance, 312
  - use cases, 312–313, 334–339
- MPDC (Mean Pupil Diameter Change) in cognitive load, 484–485
- MS-MINs (Multi-State Mutual Information Networks), 57–58
- MTC (Member-Team-Committee) architecture, 55–56
- MuDiS system, 160
- Multi-device applications, 382–383
- Multi-party dialogues, defined, 192
- Multi-party interaction with robots, 78
- Multi-State Mutual Information Networks (MS-MINs), 57–58
- Multimodal Affective and Reactive Characters toolkit (MARC), 241
- Multimodal and cross-modal reference resolution
  - reference constraints, 163–164
  - referring expressions, 162–163
- Multimodal Architecture (MMI)
  - defined, 348
  - implementations, 384
- Multimodal behavior overview, 2–3, 9–11
- Multimodal databases
  - creating, 412–419
  - data need, 394, 396–398
- design tips, 417–419
- ethics, 415
- existing, 398–411
- experimental design, 412–414
- glossary, 395
- introduction, 393
- overview, 412
- pilot studies, 415–416
- references, 419–421
- storage and distribution, 416–417
- Multimodal development tools, 641
- Multimodal dialogue system case study for medical and health systems
  - background, 438–439
  - problem description, 439
  - solution, 439–440
- Multimodal dialogues
  - challenge discussion. *See* Challenge discussion
  - in cyber-physical environments, 173–174
  - interactive conversational systems, 50–53
- Multimodal fission
  - defined, 349
  - and media synchronization, 372–377
  - and presentation planning, 151–153
- Multimodal fusion, 23
  - defined, 29, 349
  - dialogue systems, 149–153
  - early approaches to, 32–33
  - unification-based, 33–41
  - via heuristics, 123–125
- Multimodal grammar model (MMGL), 58
- Multimodal grammars
  - defined, 29
  - finite-state approaches, 41–47
  - unification-based multimodal fusion, 37
- Multimodal Human Action Database (MHAD)
  - characteristics, 404
  - description, 409
  - resource location, 406
- Multimodal in-car interaction
  - gaze, 510–512
  - gestures, 495–497

- Multimodal in-car interaction (*continued*)
  - handwriting recognition, 497–498
  - haptic controls, 487–493
  - overview, 486–487
  - secondary displays, 507–510
  - touch screens, 493–495
  - voice, 498–507
- Multimodal inference challenge, perception as, 121–123
- Multimodal input classification, 157
- Multimodal integration, 23
  - defined, 29
  - inputs, 157
  - interactive conversational systems. *See* Interactive conversational systems
- Multimodal interaction management, 47
- Multimodal interaction managers, defined, 29
- Multimodal Interaction Working Group, 351, 377–378
- Multimodal interactions in situated interaction, 112, 115
- Multimodal interfaces
  - clinical medical and health systems, 431–432
  - non-clinical medical and health systems, 435
- Multimodal Markup Language (M3L), 167
- Multimodal-multisensor combinations in medical and health systems, 459–460
- Multimodal-multisensor framework case study
  - background, 454
  - overview, 453–454
  - problem description, 454
  - solution, 454–458
- Multimodal, Multisensor Interfaces (MMIs)
  - benchmark databases, 396
  - categorization, 398
  - datasets, 399
- description, 393
- hypothesis testing databases, 397
- in the wild deployment, 396
- Multimodal parsing in unification-based multimodal fusion, 39
- Multimodal reference resolution
  - defined, 26
  - interactive conversational systems, 50–53
- Multimodal synopsis in movement modeling, 331–332
- Multimodal translingual communication
  - error handling, 604–607
  - flexi-modal communication, 607–613
  - overview, 604
- Multimodality, defined, 478
- Multiparty interactions
  - defined, 113
  - situated interaction, 112
- Multiplatform architecture in SmartKom, 167
- Multiple input streams in interactive conversational systems, 25
- Multisensor digital pen interface case study
  - background, 446
  - lessons learned, 451–453
  - overview, 445–446
  - problem description, 446–448
  - solution, 448–451
- Multisensor interfaces
  - clinical medical and health systems, 432–433
  - non-clinical medical and health systems, 436–438
- MUMIN schema, 226–227
- Muscle control in biomechanical simulation, 288
- Muscle memory
  - automobiles, 479
  - defined, 478
- Muscular aspects in musculoskeletal models, 284
- Muscular fatigue models, 281
- Musculoskeletal models, 284–285
- Mutual disambiguation, 170

- Mutual gaze events in multimodal fusion, 124
- Mutual information in machine learning, 57
- MVC (maximum voluntary contraction), 276
- Myoelectricity
  - defined, 625
  - prosthetic arms, 630
- MyRoom Video Chat product, 595
- N-best lists
  - defined, 29
  - machine learning, 56
  - multimodal grammars, 41–43
  - unification-based multimodal fusion, 34
- Nadi XTM leggings, 649–650
- Naïve Bayes gesture classifiers, 54
- NAO robot, 82–83
- National Highway Traffic Safety Administration (NHTSA)
  - driver distraction, 483
  - vocal navigation commands, 503
- Natural interaction, defined, 478
- Natural language processing (NLP), 504
- Natural language systems, 180, 382
- Natural language understanding (NLU)
  - dialogue systems, 149–150
  - natural language systems, 180
  - vocal navigation systems, 504
- Naturalistic data collection and annotation
  - for nonverbal behavior, 226–227
- Naturalistic vs. effective nonverbal performance, 248–249
- Navigation
  - multimodal in-car interaction, 488
  - vocal systems, 503–507
- Neo Smartpens, 456–458
- Nesting components, 352
- Network Time Protocol
  - clock synchronization, 418
  - defined, 395
- Networked system development in MCPS, 430
- Neural machine translation, 585–586
- Neural machine translation (NMT) networks
  - defined, 578
  - vs. SMT, 586
- Neural networks in machine translation, 587–588
- Neurophysiological measures
  - defined, 530
  - embedded multimodal interfaces, 535
- NHTSA (National Highway Traffic Safety Administration)
  - driver distraction, 483
  - vocal navigation commands, 503
- Night-vision systems
  - autonomous cars, 627
  - F-35 helmet, 623
- NIOSH questionnaire, 270
- NLP (natural language processing), 504
- NLU (natural language understanding)
  - dialogue systems, 149–150
  - natural language systems, 180
  - vocal navigation systems, 504
- NMT (neural machine translation) networks
  - defined, 578
  - vs. SMT, 586
- Noise
  - cameras, 674
  - multimodal in-car interaction measurements, 502
  - privacy concerns, 675
  - simultaneous interpretation, 601
- Noise suppression
  - defined, 79
  - human-robot interaction, 85–86
- Nomadic devices
  - cognitive load, 482
  - defined, 478
- Non-clinical medical and health systems, 426
  - multimodal interfaces, 435
  - multisensor interfaces, 436–438
  - overview, 433–435
- Non-intrusive sensors in health systems, 436–437

- Non-verbal aspects
  - defined, 79
  - robots, 78
- Non-verbal Behavior Generator (NVBG), 234–235
- Nonverbal behavior
  - Articulated Social Agents Platform, 235–237
  - behavior coherency, 246
  - Behavior Expression Animation Toolkit, 239–240
  - behavior generation example in VIB, 241–244, 249–251
  - broad differences and techniques, 230–233
  - Cerebella system, 237–238
  - communication, 227–228
  - conclusion and future trends, 245–249
  - control models, 226–228
  - data collection and annotation, 226–227
  - DTask dialogue planner, 238–239
  - embodiment, 220–225
  - focus questions, 249
  - glossary, 222–223
  - influence of, 247
  - interactive and incremental unfolding, 247
  - introduction, 219–220
  - LiteBody tool, 238–239
  - mental state feedback, 225
  - mental state mapping, 221, 223–224
  - miscellaneous systems, 239–241
  - naturalistic vs. effective nonverbal performance, 248–249
  - references, 252–262
  - SAIBA framework, 228–230
  - shared representations, metaphor, and metaphoric gesture, 224–225
  - synchronization, 228
  - timing and synchronization, 246
  - Virtual Human Toolkit, 234–237
  - Virtual Interactive Behavior platform, 233–234
  - Note phones, 639
- NoXi database
  - characteristics, 404
  - description, 409–410
  - resource location, 406
- Nuance Mix system, 180, 382
- Nuance Nina IQ Studio, 180
- NVBG (Non-verbal Behavior Generator), 234–235
- OAA (Open Agent Architecture), 350–351
- Object detection, privacy concerns, 675
- Object references in situated interaction, 112
- OCRA Index, 271
- Oculus Rift headset, 645
- Oculus Rift-integrated binocular eye tracking system, 460
- Odometric systems in autonomous cars, 626
- ODP (Ontology-based Dialog Platform) framework, 168–169
- Offline processing for simultaneous interpretation, 598
- One-dimensionality in vocal navigation systems, 503
- One-stage approach in multimodal grammars, 41
- One-way phrasebooks for machine translation, 589
- Ontology-based Dialog Platform (ODP) framework, 168–169
- OOVs (Out-of-Vocabulary Words)
  - defined, 578
  - multimodal translingual communication, 604–607
- Open Agent Architecture (OAA), 350–351
- OpenBionic arms, 630
- OpenCV toolkit, 88
- OpenFace toolkit, 450–451
- OpenSim software, 288
- openSMILE toolkit, 451
- Openstream MAM Server, 170
- Optical motion capture, 283–284
- Optical sensors in F-35 helmet, 623

- Optimization  
 biomechanical simulation, 286–290  
 decision-theoretic, 132  
 embedded multimodal interfaces, 558  
 exoskeleton mode changes, 567  
 joint movement, 267  
 movement modeling, 314, 316, 318–319
- Optional information in frame-based dialogue management architectures, 156
- OptiTrack system, 283
- Ordinal questions in discomfort questionnaires, 270
- OSPAÑ tasks, 499
- Out-of-Vocabulary Words (OOVs)  
 defined, 578  
 multimodal translingual communication, 604–607
- Output coordination in presentation planning, 151–153
- Output in machine translation, 581
- Output space in movement modeling, 317
- Ovako Working posture Analysis System (OWAS), 271
- Over-dimensionalization in multimodal database design, 418
- Over-generate-and-filter technique for nonverbal behavior, 232
- Overhearers in turn-taking, 92
- Overlays  
 defined, 29  
 unification-based multimodal fusion, 36
- Overt measures  
 defined, 530  
 embedded multimodal interfaces, 535–536  
 robotic applications, 544
- OWAS (Ovako Working posture Analysis System), 271
- Ownership of data issues, 562–565
- Parametrized behavior trees (PBTs), 240
- Parsing  
 finite-state approaches, 44
- unification-based multimodal fusion, 38–40
- Partially observable Markov decision processes (POMDPs), 182
- Participation status  
 defined, 114  
 situated interaction, 111
- Particle filtering for autonomous cars, 626
- Partners in multimodal database design, 417–418
- Passive BCIs  
 defined, 530  
 embedded multimodal interfaces, 537
- Passive input modes in clinical systems, 429
- Passive states  
 defined, 530  
 embedded multimodal interfaces, 535–536  
 human-robot interaction, 524
- Pattern mining methods in health systems, 435
- Pattern noise in cameras, 674
- PBTs (parametrized behavior trees), 240
- PCA (principal component analysis) in movement modeling, 319, 326
- Pen Input and Multimodal Systems, 366
- Pen interface case study  
 background, 446  
 lessons learned, 451–453  
 overview, 445–446  
 problem description, 446–448  
 solution, 448–451
- PEO (Portable Ergonomics Observation) method, 271
- Pepper robot, 631–633, 636
- Perception  
 actions, 131–135  
 engagement in social interaction, 119, 121–135  
 machine learning approaches, 125–131  
 multimodal fusion via heuristics, 123–125  
 as multimodal inference challenge, 121–123

- Perception (*continued*)  
 speaking level, 93–94
- Perceptual tunneling in multimodal in-car interaction, 508–509
- Perfect Squat Challenge app, 650
- Periphery in in-car interaction, 486
- Person detection, privacy concerns, 675
- Personal care products, 648–650
- Personal digital assistants, 180–182
- Persuasive technologies, defined, 428
- Pervasive distributed computing, defined, 479
- PhaseSpace system, 283
- Phonology in finite-state approaches, 44
- Physical context  
 defined, 114  
 situated interaction, 106, 111–112
- Physical ergonomics  
 anthropometric posture prediction models, 280  
 data collection overview, 269–270  
 defined, 267  
 direct health risk estimation models, 278–280  
 direct measurement methods, 272–278  
 discomfort questionnaire methods, 270–271  
 experimental model overview, 278  
 exposure-effect models, 280  
 muscular fatigue models, 281  
 physiologic measures, 281  
 posture-based skeletal load prediction models, 280–281  
 posture observation methods, 271–272
- Physical factors in mental state mapping, 224
- Physiologic measures, 281
- Physiological computing  
 defined, 530  
 embedded multimodal interfaces, 537  
 robotic applications, 543–544
- Physiological measures  
 defined, 530  
 embedded multimodal interfaces, 558
- Physiological processes for ergonomics measurements, 273
- Pillo robot, 632–633
- Pilot product, 612
- Pilot studies for multimodal database design, 415–416
- Pixel-buds  
 defined, 578  
 multimodal translingual communication, 612
- Place deixis expressions, 162–163
- Plan-based dialogue agents, 156
- Platform for Situational Intelligence (PSI), 641
- PLIBEL questionnaire, 270
- PLS (Pronunciation Lexicon Specification), 354–355, 364
- Point devices on headsets, 645
- POMDPs (partially observable Markov decision processes), 182
- Portable Ergonomics Observation (PEO) method, 271
- Position-controlled robots, 532
- Post-stroke hold phase for gestures, 222
- Post-traumatic stress disorder (PTSD), 460
- Posture-based skeletal load prediction models, 280–281
- Posture distribution-based RULA, 271
- Posture observation methods for physical ergonomics, 271–272
- Prediction in machine learning approaches, 128
- Preparation gesture phases in VIB, 244
- Presentation planning  
 dialogue systems, 151–153  
 SiAM-dp platform, 172
- Pressure sensors (PrS) for physical ergonomics measurements, 273
- PresTK platform, 153
- Prestroke hold phase for gestures, 222
- Prevention  
 clinical systems, 429  
 defined, 428
- Primary prevention  
 clinical systems, 429

- defined, 428
- Primary tasks in cognitive load, 482
- Principal component analysis (PCA) in movement modeling, 319, 326
- Principle of upward completion for robots, 93–94
- Privacy concerns
  - anonymous social media accounts, 682–684
  - attacks, 676–677
  - calls for privacy, 662
  - case studies, 677–684
  - corruptibles, 677–679
  - cybercasing from data brokers, 679–682
  - differential privacy, 671–672
  - education, 686–692
  - embedded multimodal interfaces, 562–565
  - focus questions, 692–696
  - future directions, 685–691
  - glossary, 663–664
  - introduction, 659–662, 664–665
  - k-anonymity, 669–671
  - logics, 673
  - medical and health systems regulations, 452
  - mitigation research, 685–686
  - notable works, 673–674
  - references, 696–704
  - risks, 674–676
  - structured data, approaches to, 668–673
  - structured data, threats on, 667–668
- Pro Eye headset, 644–645
- Probabilistic architectures based on POMDPs, 182
- Probabilistic dialogue management architectures, 156–157
- Probabilistic models for nonverbal behavior, 232
- Processing times in movement modeling, 310
- Product search, commercialization of, 643–644
- Pronouns in simultaneous interpretation, 600
- Pronunciation in automatic speech recognition, 583
- Pronunciation Lexicon Specification (PLS), 354–355, 364
- Prosody
  - defined, 80
  - human-robot interaction, 87
  - spoken language translation, 579
- Prosthetic arms, 629–631
- Proxemics
  - defined, 114
  - perception, 121
- Proximity in multimodal fusion, 124
- PrS (pressure sensors) for physical ergonomics measurements, 273
- PSI (Platform for Situational Intelligence), 641
- Psychophysiological measures in human-robot interaction, 526
- Psychophysiological techniques in human-robot interaction, 535
- PTSD (post-traumatic stress disorder), 460
- Public scenario for SMARTKOM project, 411
- Punctuation in spoken language translation, 579
- Pupil size (PS) in physical ergonomics measurements, 273
- Pupillometry for cognitive load, 484–485
- Push-to-Talk buttons in multimodal in-car interaction, 500–501
- Put-that-there system, 32
- pySPACE framework, 558
- Quantified-self
  - defined, 428
  - medical and health systems, 426
- Quasi-identifiers
  - defined, 664
  - privacy concerns, 669–670
- Question Answering in vocal navigation systems, 504
- Question answering (QA) bots, 181
- Questionnaire methods in physical ergonomics, 270–271

- Quick Exposure Checklist, 271
- QuickSet system  
late fusion, 160  
machine learning, 56  
prototype system, 23  
unification-based multimodal fusion, 34, 36
- Radar (Radio Detection and Ranging) in autonomous cars, 626
- Radio Frequency Identification (RFID)  
defined, 395  
TUM Kitchen dataset, 411
- RadLex ontology, 440
- RadSem tool, 439–441
- Radspeech system  
architecture, 445  
mutual disambiguation of errors, 432
- Rank level in movement modeling, 308
- Rapid Entire Body Assessment (REBA)  
physical ergonomics, 278–279  
posture observation methods, 271–272
- Rapid Upper Limb Assessment (RULA), 271–272
- Rapport  
defined, 223  
nonverbal behavior, 225
- Raspberry PI computers, defined, 395
- Ratings of Perceived Exertion (RPE) scale, 270
- Raw data in movement modeling, 328
- RDF (Resource Description Framework)  
defined, 428  
repository, 431
- Re-identification  
defined, 664  
privacy concerns, 674
- Reactive BCI  
defined, 530  
description, 536–537
- Readability in simultaneous interpretation, 600
- REBA (Rapid Entire Body Assessment)  
physical ergonomics, 278–279  
posture observation methods, 271–272
- Receptionist system  
detailed trace of interaction, 120–122  
situated interaction, 106–109
- Recognition errors in multimodal in-car interaction, 502
- RECOLA database  
characteristics, 405  
description, 410  
resource location, 406
- Recruitment plans in multimodal database design, 414
- Rectified linear units (ReLUs) in movement modeling, 336–337
- Recurrent Neural Networks (RNNs)  
machine learning, 53  
machine translation, 585–586
- Recursive Transition Networks (RTNs),  
defined, 30
- Redundancy  
defined, 158  
multimodal input, 157
- Reem robot, 633
- Reem-C robot, 633
- Reference resolution  
defined, 26  
interactive conversational systems, 25  
multimodal, 50–53, 162–164
- References to collections, 163
- Referents in referring expressions, 162
- Referring expressions, 162–163
- Reflexive adaptation  
defined, 530  
embedded multimodal interfaces, 557
- Relational agents in nonverbal behavior, 238–239
- Relative timing in modality fusion, 371–372
- Relaxation phase for gestures, 222
- Reliability in movement modeling, 310–311
- ReLU (rectified linear units) in movement modeling, 336–337
- Remote tactile feedback in multimodal in-car interaction, 494–495
- Repair in multimodal translingual communication, 605–607
- Repetitive strain injury (RSI), defined, 267

- Repetitive tasks in multimodal in-car interaction, 488
- Representation
  - clinical medical and health systems, 431
  - engagement in social interaction, 117–121
- Representational State Transfer (REST)
  - API simplification, 352
  - defined, 349
  - medical and health systems, 444–445
- Required information in frame-based dialogue management architectures, 156
- Research systems and prototypes for machine translation, 588–591
- Resolved semantic content in semantic dialogue act model, 175
- Resource Description Framework (RDF)
  - defined, 428
  - repository, 431
- Resource Manager in W3C Multimodal Architecture, 383
- reSPACE framework, 558
- Respiratory measurement (RM) for physical ergonomics, 273
- Respond to Joint Attention (RJA) for robots, 96
- REST (Representational State Transfer)
  - API simplification, 352
  - defined, 349
  - medical and health systems, 444–445
- Restricted Boltzmann machines, 324
- RFID (Radio Frequency Identification)
  - defined, 395
  - TUM Kitchen dataset, 411
- RGB video in movement modeling, 312–315, 328–330, 335, 338
- Risks, privacy, 674–676
- RJA (Respond to Joint Attention) for robots, 96
- RNNs (Recurrent Neural Networks)
  - machine learning, 53
  - machine translation, 585–586
- Road sign translators, 608
- RoboHelper project, 54
- Robot Construction Kit (ROCK), 559
- Robot CPS surgery systems, 433
- Robot Operating System (ROS), 84
- Robots and robotics, 527, 531–534
  - actions, 134
  - auditory channel, 84–87
  - autonomous cars, 625–628
  - commercialization, 625–636
  - conclusions, 97–98
  - embedded multimodal interface applications, 542–555
  - embedded multimodal interfaces, 534–542
  - faces, importance, 80–82
  - faces, providing, 82–83
  - glossary, 79–80
  - grounding and feedback, 93–95
  - human-robot interaction, 83–89
  - introduction, 77–78
  - joint attention, 95–97
  - prosthetic arms, 629–631
  - references, 98–104
  - situation model, 88–89
  - social robots, 631–636
  - surgical robots, 628–629
  - turn-taking, 89–93
  - visual channel, 87–88
- Robustness in machine learning approaches, 129–130
- ROCF cognitive assessment, 451
- ROCK (Robot Construction Kit), 559
- ROS (Robot Operating System), 84
- RPE (Ratings of Perceived Exertion) scale, 270
- RSI (repetitive strain injury), defined, 267
- RTF (runtime framework) in W3C MMI Architecture, 380–381
- RTNs (Recursive Transition Networks), defined, 30
- RULA (Rapid Upper Limb Assessment), 271–272
- Rule-based approaches
  - automatic speech recognition, 584
  - Cerebella system, 237–238
  - nonverbal behavior, 231–232

- Run-time dialogue model in SiAM-dp platform, 172
- Runtime framework (RTF) in W3C MMI Architecture, 380–381
- Safe robots, 527, 531–534
- Safety belts in in-car interaction, 492
- Safety by design
  - defined, 530
  - robots, 527
- SAIBA framework, 228–230
- SAL scenario for SEMAINE database, 410–411
- SantosHuman software, 288
- SC (stressor characteristics) in medical and health systems, 449
- SCR (skin conductance response) in medical and health systems, 454
- SCXML. *See* State Chart XML (SCXML)
- SDSs (spoken dialogue systems)
  - defined, 479
  - multimodal in-car interaction, 498–507
- SE (squared exponential covariance) in movement modeling, 319
- Search engines, privacy threats from, 668
- Second-order adaptation in embedded multimodal interfaces, 557–558
- Secondary displays in multimodal in-car interaction, 507–510
- Secondary prevention
  - clinical systems, 429
  - defined, 428
- Secondary tasks in cognitive load, 482
- Secondary uses of data
  - defined, 664
  - privacy issues, 665–666
- Self-adaptation in embedded multimodal interfaces, 555–560
- SEMAINE database
  - characteristics, 405
  - description, 410–411
  - resource location, 406
- SEMAINE System for virtual listener agents, 240–241
- Semantic constraints, 164
- Semantic dialogue act model, 174–176
- Semantic Interpretation for Speech Recognition (SISR), 354–355, 364
- Semantic level in late fusion, 160
- sEMG (surface electromyography), 273, 277
- Semi-automatic SAL scenario for SEMAINE database, 410
- Sense-think-act loops in engagement in social interaction, 119
- Sequence tagging in machine learning, 53
- Shared representations in nonverbal behavior, 224–225
- ShopAssist system, 160
- Shoptalk system, 32
- SHRLDU text-based dialogue systems, 105
- SI (Strain Index), 271–272
- SIAM. *See* Situation-Adaptive Multimodal Dialogue Platform (SiAM-dp)
- Side participants in turn-taking, 92
- Silent speech input in multimodal translingual communication, 610–611
- SIMM software for biomechanical simulation, 288
- Simplicity in late fusion movement modeling, 311
- Simultaneous interpretation
  - defined, 578
  - machine translation, 580, 591, 595–603
- Siri assistant for non-clinical medical and health systems, 436
- SISR (Semantic Interpretation for Speech Recognition), 354–355, 364
- Situated interaction
  - conclusions, 135–137
  - defined, 114
  - engagement. *See* Engagement in social interaction
  - focus questions, 137–138
  - glossary, 113–115
  - introduction, 105–110
  - references, 138–143
  - spoken language, 110–112, 115, 126

- Situation-Adaptive Multimodal Dialogue Platform (SiAM-dp)  
 basic architecture, 170–173  
 cognitive resource awareness, 178–179  
 description, 170  
 dialogue system offline evaluation, 179–180  
 multimodal dialogue in cyber-physical environments, 173–174  
 semantic dialogue act model, 174–176  
 situation-adaptive behavior, 176–178
- Situation modeling  
 defined, 80  
 human-robot interaction, 88–89
- Sketch-Thru-Plan (STP) system, 24, 34
- Skills in VDA, 181
- Skin conductance for emotional activation, 536
- Skin conductance response (SCR) in medical and health systems, 454
- Skin360 product, 649–650
- Skype Translator, 595–596
- SLM (statistical language model), 46
- Slot-filling systems, 192–193
- Smart devices in medical and health systems, 434
- Smart Display, 638
- Smart hairbrushes, 648–649
- SmartBody platform for nonverbal behavior, 235
- SMARTKOM database  
 characteristics, 405  
 description, 411  
 resource location, 406
- SmartKom system  
 dialogue platform, 167–168  
 late fusion, 160  
 multimodal reference resolution, 52  
 ODP concepts from, 169  
 presentation planning, 153  
 prototype system, 23
- Smartphones  
 machine translation, 592–593
- non-clinical medical and health systems, 436
- SMIL (Synchronized Multimedia Integration Language)  
 defined, 349  
 ink standard, 365–366
- Smoothing  
 defined, 115  
 machine learning approaches, 128
- SMS (Speech Mashup Server), 164
- SMT (statistical machine translation)  
 defined, 578  
 description, 585  
 limitations, 586
- Social bots, 181
- Social companions in non-clinical medical and health systems, 437–438
- Social force in nonverbal behavior, 232
- Social media accounts, privacy concerns for, 682–684
- Social robots, 631–636
- Social situational factors in mental state mapping, 224
- Societal implications in embedded multimodal interfaces, 560–565
- Soft sensors in CueMe platform, 170
- Software agents in OAA, 350
- Software frameworks for embedded multimodal interfaces, 558–560
- Software platforms and toolkits  
 definitions, 145–146  
 dialogue management architectures, 154–157  
 dialogue systems, 146–154  
 existing dialogue platforms, 164–170  
 focus questions, 182–183  
 fusion and communicative functions, 157, 159–162  
 glossary, 158–159  
 introduction, 145  
 multimodal and cross-modal reference resolution, 162–164  
 references, 183–190  
 SiAM-dp platform, 170–180

- Software platforms and toolkits (*continued*)
  - trends in dialogue architectures, 180–182
- Solid SAL scenario for SEMAINE database, 410
- Sound source localization, 79
- Spaces in movement modeling, 315–317
- SPARQL query standard, 431, 443
- Speak-what-you-see concept, 501, 511
- Speak4It system
  - conversational systems, 23
  - machine learning, 55
  - multimodal reference resolution, 52
- Speaker role in turn-taking, 92
- Spectacles product, 612
- Speech and spoken language
  - finite-state approaches, 44–46
  - gestures, 81–82
  - human-robot interaction, 84–87
  - interactive conversational systems, 31
  - machine learning, 54
  - machine translation, 579
  - medical and health dialogue systems, 438–439
  - multimodal fusion, 32
  - multimodal grammars, 41–43
  - multimodal in-car interaction, 498–507
  - predicting, 129
  - situated interaction, 106, 110–112, 115
  - unification-based multimodal fusion, 34–36, 39
  - visual articulation, 81
- Speech confusions in multimodal translational communication, 604
- Speech inputs
  - interactive conversational systems, 24
  - multimodal in-car interaction, 495
  - multimodal translational communication, 609–610
- Speech Mashup Client, 164
- Speech Mashup Server (SMS), 164
- Speech Recognition Grammar Specification (SRGS), 165, 354–355, 364
- Speech synthesis
  - defined, 578
  - machine translation, 582, 587
- Speech Synthesis Markup Language (SSML)
  - AT&T speech mashup architecture, 165
  - defined, 349
  - voice standard, 354–356, 364
- Speech-to-text (STT) machine translation, 579
- Speech translation
  - defined, 578
  - goggles, 611–612
  - machine. *See* Machine translation
- Spoken dialogue systems (SDSs)
  - defined, 479
  - multimodal in-car interaction, 498–507
- Spontaneous speech
  - machine translation, 588–591
  - simultaneous interpretation, 600–601
- Spoofing biometrics
  - defined, 624
  - minimizing, 637
- Squared exponential covariance (SE) in movement modeling, 319
- SRGS (Speech Recognition Grammar Specification), 165, 354–355, 364
- SRI International Artificial Intelligence Center, 350
- SSML (Speech Synthesis Markup Language)
  - AT&T speech mashup architecture, 165
  - defined, 349
  - voice standard, 354–356, 364
- SSVEP (steady-state visually evoked potential), 536–537
- Stability in multimodal fusion, 124
- Standardised Nordic Questionnaire, 270
- Standardized representations
  - future of, 381–383
  - open challenges, 383–384
- Standards in clinical medical and health systems, 431
- Stare in the crowd-effect, 81
- State Chart XML (SCXML)
  - assistants, 640
  - defined, 349
  - IM standard, 354–356

- multimodal fission and media synchronization, 375–377
- SiAM-dp platform, 173
- State space in movement modeling, 316
- Statecharts
  - defined, 30, 80
  - human-robot interaction, 89
  - interactive conversational systems, 25
- States
  - engagement, 117–118
  - exoskeletons, 546
  - human-robot interaction, 524
  - mental state mapping, 223–224
- Static optimization
  - biomechanical simulation, 287–290
  - defined, 267
- Statistical disclosure prevention
  - defined, 664
  - description, 666
  - privacy approaches on structured data, 668
- Statistical language model (SLM), 46
- Statistical machine translation (SMT)
  - defined, 578
  - description, 585
  - limitations, 586
- Statistics in machine translation, 587–588
- Steady-state visually evoked potential (SSVEP), 536–537
- Steering wheels in in-car interaction, 486, 491, 496–497
- Storage in multimodal database design, 416–417
- STP (Sketch-Thru-Plan) system, 24, 34
- Strain Index (SI), 271–272
- Streaming ink applications, 367
- Stressor characteristics (SC) in medical and health systems, 449
- Strokes
  - defined, 30
  - gestures, 28, 222
  - multimodal grammars, 42–43
- Structured data
  - privacy concern approaches on, 668–673
  - privacy concern threats on, 667–668
- STT (speech-to-text) machine translation, 579
- Study-Model-Build-Test development cycle in nonverbal behavior, 245
- Sub-symbolic level for early fusion, 159–160
- Subcategorization frames
  - defined, 30
  - unification-based multimodal fusion, 40–41
- Substitution errors in multimodal translingual communication, 607
- Subtitling in machine translation, 580
- Sufficiency of sensing
  - defined, 625
  - F-35 helmet, 623
- Support Vector Machines (SVMs), 53
- Surface electromyography (sEMG), 273, 277
- Surgical robots, 433, 628–629
- SVMs (Support Vector Machines), 53
- Symbol grounding, defined, 79
- Symbolic level in late fusion, 160
- Symmetric multimodality in SmartKom, 167
- Synchronicity in late fusion movement modeling, 311
- Synchronization
  - defined, 223
  - fusion, 157, 159
  - multimodal database design, 418
  - nonverbal behavior, 225, 228, 246
- Synchronized Multimedia Integration Language (SMIL)
  - defined, 349
  - ink standard, 365–366
- Syntactic constraints, 164
- System-initiative dialogue systems
  - defined, 159
  - description, 150
- System initiatives, defined, 114
- Tablet devices for medical and health systems, 450

- Tactile feedback in multimodal in-car interaction, 491–492, 494–495
- Tags
  - defined, 349
  - XML, 357–358
- Tailorability in late fusion movement modeling, 311
- Talos robot, 633
- Tangible user interfaces
  - automobiles, 480
  - defined, 479
- Targeted audio
  - defined, 578
  - multimodal translational communication, 610–611
- Task-independent motion models, 334
- Task loads in robotic applications, 550
- Task-oriented bots, 181
- Task-oriented spoken dialogue systems
  - defined, 114
  - situated interaction, 105
- Task space in movement modeling, 316
- Tasks in cognitive load, 482
- TBIs (traumatic brain injuries), 453
- Teachers' Resources for Online Privacy Education (TROPE), 687–688, 692
- Teaching Privacy Project (TPP), 687–688
- Technical multimodal database design, 412–413
- Technical terms in simultaneous interpretation, 599–600
- Technology for machine translation, 581–588
- Telemedicine
  - clinical medical and health systems, 432
  - defined, 428
- Teleoperation mode
  - exoskeletons, 546
  - VI-Bot exoskeleton, 565–567
- Temporal cascaded approach
  - defined, 530
  - embedded multimodal interfaces, 551
- Temporal relationships
  - fusion, 157, 159
- logical tasks, 359
- Ten Principles for Online Privacy, 687–692
- Tertiary prevention
  - clinical systems, 429
  - defined, 428
- Tertiary tasks in cognitive load, 482
- Tesla Model S, 627
- Text-based dialogue systems
  - defined, 115
  - situated interaction, 105
- Text-to-speech synthesis (TTS)
  - defined, 578
  - machine translation, 587
  - medical and health systems, 440
- Text to speech (TTS) engine in SRGS, 364
- Therapist interface, 449–450
- Thermal imaging privacy issues, 665
- Tiago robot, 633
- Time deixis expressions, 163
- Time estimation privacy concerns, 675
- Time markers in VIB, 244
- Timed Text Markup Language (TTML), 373
- Timing in nonverbal behavior, 246
- TMT cognitive assessment, 451, 456
- Touch screens for in-car interaction, 487, 493–495
- Touchsense 8.4" LCD touchscreen, 494
- TPP (Teaching Privacy Project), 687–688
- Tracking
  - defined, 115
  - machine learning approaches, 128
  - movement modeling, 335
- Tracking device privacy issues, 665
- Tracks
  - defined, 625
  - F-35 helmet, 623
- Trait factors in mental state mapping, 223–224
- Transactional chatbot dialogues, 200–205
- Transcode
  - defined, 625
  - surgical robots, 629
- Transfer learning in challenge discussion, 196, 209–213

- Transformations in movement modeling spaces, 316–317
- Transience in vocal navigation systems, 503
- Transient mode
  - exoskeletons, 546
  - VI-Bot exoskeleton, 565–569
- Transition-Relevance Places (TRPs), 90
- Translation errors in multimodal translational communication, 605
- Transparency
  - embedded multimodal interfaces, 560
  - robotic applications, 544–545
- Traumatic brain injuries (TBIs), 453
- TRIPS system, 105
- TROPE (Teachers' Resources for Online Privacy Education), 687–688, 692
- TRPs (Transition-Relevance Places), 90
- Trust issues
  - embedded multimodal interfaces, 560–562
  - human-robot interaction, 526
- TTML (Timed Text Markup Language), 373
- TTS (text-to-speech synthesis), 364
  - defined, 578
  - machine translation, 587
  - medical and health systems, 440
- TUM Kitchen database
  - characteristics, 405
  - description, 411
  - resource location, 406
- Turn management functions
  - defined, 159
  - VHT, 234
- Turn-taking
  - defined, 80
  - robots, 78, 89–93
- Tutoring systems for medical and health systems, 459
- Two-layered latent movement modeling representation, 330–331
- Typed feature structures
  - defined, 30
  - multimodal reference resolution, 52
- unification-based multimodal fusion, 34–35
- UCD (user-centered design), 266, 268
- Ultrasound sensors in autonomous cars, 626
- UML (Unified Modeling Language) for SCXML, 375
- Uncanny valley phenomenon, 83
- Understanding level of action in speaking, 93–94
- Unfolding in nonverbal behavior, 247
- Unification
  - defined, 30
  - interactive conversational systems, 25
  - overview, 33–41
- Unified Modeling Language (UML) for SCXML, 375
- Unique identifiers (URIs) for EmotionML, 368–369
- Unknown words in simultaneous interpretation, 599–600
- Unresolved semantic content in semantic dialogue act model, 175
- Unscented Kalman Filters
  - autonomous cars, 626
  - defined, 625
- URIs (unique identifiers) for EmotionML, 368–369
- Use case for movement modeling, 312–313, 334–339
- User-centered design (UCD), 266, 268
- User-initiative dialogue systems
  - defined, 159
  - description, 150
- User initiatives
  - defined, 114
  - situated interaction, 134
- User input in semantics, 370–371
- Utah study
  - driver distraction, 485
  - recognition errors, 502
- Utterances in human-robot interaction, 90, 498–500

- VAD (voice activity detection)
  - defined, 80
  - human-robot interaction, 86
- VAE-DMP in movement modeling, 326–327, 337
- VAEs (variational autoencoders)
  - defined, 308
  - description, 11
  - movement modeling, 324–327, 335–338
- Valence
  - defined, 395
  - embedded multimodal interfaces, 561
- VAM database
  - characteristics, 405
  - description, 411
  - resource location, 406
- Variational autoencoders (VAEs)
  - defined, 308
  - description, 11
  - movement modeling, 324–327, 335–338
- VDA (Virtual Digital Assistant), 180–181
- Verbal communication
  - defined, 80
  - robots, 78
- VERBMOBIL system, 590
- Verbsurgical platform, 433
- Verrotouch Medical surgical robot, 629
- VHT (Virtual Human Toolkit), 234–236
- VI-Bot interface
  - exoskeleton mode changes, 565–569
  - robotic applications, 545, 547–550
- VIB (Virtual Interactive Behavior platform)
  - behavior generation example, 241–244, 249–251
  - overview, 233–234
- Vibrations in in-car interaction, 492
- Video
  - latent representations, 328–329
  - privacy concerns, 674
  - robotic applications, 549–550, 553
- Virginia Technology Transportation Institute study of driver distraction, 483
- Virtual animated characters, faces for, 82
- Virtual assistant-based insurance solutions, 647–648
- Virtual assistants, 180–181, 638–642
- Virtual companions for non-clinical medical and health systems, 437–438
- Virtual Digital Assistant (VDA), 180–181
- Virtual health agents, 435
- Virtual Human Toolkit (VHT), 234–236
- Virtual Humans
  - defined, 395
  - depression assessment, 437
  - interaction with, 399
- Virtual Interactive Behavior platform (VIB)
  - behavior generation example, 241–244, 249–251
  - overview, 233–234
- Virtual nurse interface, 435
- Virtual People Factory, 239
- Virtual reality (VR) systems
  - commercialization, 644–646
  - medical and health systems, 460–462
- Visual appearance of motion, 305
- Visual articulation in faces, 81
- Visual channels in human-robot interaction, 87–88
- Visual demand in multimodal in-car interaction, 503
- Visual desynchronization in nonverbal behavior, 231
- Visual state charts in incremental multimodal integration, 47–50
- Viterbi search in machine learning, 58
- Voice activity detection (VAD)
  - defined, 80
  - human-robot interaction, 86
- Voice Browser Working Group
  - SCXML, 375–376
  - voice standards, 351
- Voice in multimodal in-car interaction, 498–507
- Voice Search navigation systems, 504–505
- Voice standards
  - declarative languages, 363–364
  - initial work, 351

- VoiceXML markup language
  - defined, 349
  - description, 363
  - standard, 354
- VR (virtual reality) systems
  - commercialization, 644–646
  - medical and health systems, 460–462
- Vue.ai product search, 643–644
- Vuzix Blade augmented reality headset, 645
- W3C MMI Architecture, 378
  - communication and event processing, 378–380
  - Resource Manager, 383
  - runtime framework and extension notification, 380–381
  - standards-based implementation, 351–353
- W3C Multimodal Interaction Working Group, 364
- W3C (WorldWideWeb Consortium), defined, 349
- Wait-vs.-act tradeoff
  - actions, 133–135
  - defined, 115
- WAMI (Web-Accessible Multimodal Interfaces) toolkit, 165–166
- WatsonPaths system, 434
- Wearable devices
  - embedded multimodal interfaces, 535
  - medical and health systems, 434
- Web-Accessible Multimodal Interfaces (WAMI) toolkit, 165–166
- Web Ontology Language, 383
- Web Services Description Language, 383
- Weka package, 54
- What You See Is What You Touch (WYSIWYT) concept, 509
- Wickens model of attention management, 179
- Will avatar, 642
- Windshields in in-car interaction, 486
- wit.ai VDA, 180
- Wizard of Oz methodology and studies
  - data collection, 399
  - defined, 395, 479
  - multimodal reference resolution, 51
  - recognition errors, 502
  - robots, 97
  - SMARTKOM project, 411
- WorldWideWeb Consortium (W3C), defined, 349
- WYSIWYT (What You See Is What You Touch) concept, 509
- XML. *See* eXtensible Markup Language (XML)
- Yielding turns in human-robot interaction, 91
- Yoga leggings, 649
- Zeno robot, 632, 635



# Biographies

## Editors

**Philip R. Cohen** (Monash University) is Director of the Laboratory for Dialogue Research and Professor of Artificial Intelligence on the Faculty of Information Technology at Monash University. His research interests include multimodal interaction, human-computer dialogue, and multi-agent systems. He is a Fellow of the American Association for Artificial Intelligence, past President of the Association for Computational Linguistics, recipient (with Hector Levesque) of the Inaugural Influential Paper Award by the International Foundation for Autonomous Agents and Multi-Agent Systems, and recipient of the 2017 Sustained Achievement Award from the International Conference on Multimodal Interaction. He was most recently Chief Scientist for Artificial Intelligence and Senior Vice President for Advanced Technology at Voicebox Technologies, and was founder of Adapx Inc. Prior to Adapx, he was a Professor at Oregon Health and Science University and Program Director of Natural Language in the Artificial Intelligence Center at SRI International. Cohen has published more than 150 articles and has 5 patents. He co-authored the book *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces* (2015, Morgan & Claypool Publishers) with Sharon Oviatt. (Contact: [philip.cohen@monash.edu](mailto:philip.cohen@monash.edu))

**Antonio Krüger** (Saarland University and DFKI GmbH) is Professor of Computer Science and Director of the Media Informatics Program at Saarland University, as well as Scientific Director of the Cognitive Assistants Department at the German Research Center for Artificial Intelligence (DFKI). His research areas focus on intelligent user interfaces, and mobile and ubiquitous context-aware systems. He has been General Chair of the Ubiquitous Computing Conference and Program Chair of MobileHCI, IUI, and Pervasive Computing. He is also on the Steering Committee of the International Conference on Intelligent User Interfaces (IUI) and an Associate Editor of the journals *User Modeling and User-Adapted Interaction* and *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*. (Contact: [krueger@dfki.de](mailto:krueger@dfki.de))

**Sharon Oviatt** (Monash University) is internationally known for her multidisciplinary work on multimodal and mobile interfaces, human-centered interfaces, educational interfaces, and learning analytics. She was a recipient of the inaugural ACM-ICMI Sustained Accomplishment Award, National Science Foundation Special Creativity Award, and ACM-SIGCHI CHI Academy award. She has published over 160 scientific articles in a wide range of venues and is an Associate Editor of major journals and edited book collections in the field of human-centered interfaces. Her other books include *The Design of Future Educational Interfaces* (2013, Routledge) and *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces* (2015, Morgan & Claypool Publishers). (Contact: [sharon.ovviatt@monash.edu](mailto:sharon.ovviatt@monash.edu))

**Gerasimos Potamianos** (University of Thessaly) is Associate Professor and Director of Graduate Studies in Electrical and Computer Engineering. His research spans multisensory and multimodal speech processing and scene analysis, with applications to human-computer interaction and ambient intelligence. He has authored over 120 articles and has 7 patents. He received a Diploma degree from the National Technical University of Athens, and a M.Sc. and Ph.D. from Johns Hopkins University, all in electrical and computer engineering. In addition to his academic experience, he has worked at AT&T Research Labs, IBM Thomas J. Watson Research Center (US), and at the FORTH and NCSR ‘Demokritos’ Research Centers in Greece. (Contact: [gpotam@ieee.org](mailto:gpotam@ieee.org))

**Björn Schuller** (University of Augsburg and Imperial College London) is currently ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at University of Augsburg and Professor of Artificial Intelligence at Imperial College. He is best known for his work on multisensorial/multimodal intelligent signal processing for affective, behavioral, and human-centered computing. In 2015 and 2016, he was honored by the World Economic Forum as one of 40/50 extraordinary scientists under age 40. In 2018, he was elevated to Fellow of the IEEE and Senior Member of the ACM. He has published over 800 peer-reviewed scientific contributions across a range of disciplines and venues, and is Editor-in-Chief of *IEEE Transactions on Affective Computing*. His books include *Intelligent Audio Analysis* (2013, Springer) and *Computational Paralinguistics* (2013, Wiley). (Contact: [bjoern.schuller@imperial.ac.uk](mailto:bjoern.schuller@imperial.ac.uk))

**Daniel Sonntag** (German Research Center for Artificial Intelligence, DFKI GmbH) is a Principal Researcher and Research Fellow. His research interests include multimodal and mobile AI-based interfaces, common-sense modeling, and explainable machine learning methods for cognitive computing and improved usability. He has

published over 130 scientific articles and was the recipient of the German High Tech Champion Award in 2011 and the AAAI Recognition and IAAI Deployed Application Award in 2013. He is the Editor-in-Chief of the *German Journal on Artificial Intelligence (KI)* and Editor-in-Chief of Springer's Cognitive Technologies book series. Currently, he leads both national and European projects from the Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and Horizon 2020. (Contact: daniel.sonntag@dfki.de)

### **Authors and Challenge Discussants**

**James Allen** is the John H. Dessauer Professor of Computer Science at the University of Rochester and Associate Director of the Institute for Human and Machine Cognition in Pensacola, Florida. He is a Founding Fellow of the American Association for Artificial Intelligence (AAAI) and a Fellow of the Cognitive Science Society. He was Editor-in-Chief of the journal *Computational Linguistics* from 1983–1993 and authored the well-known textbook *Natural Language Understanding*. His research concerns defining computational models of intelligent collaborative and conversational agents, with a strong focus on the connection between knowledge, reasoning, and language comprehension and dialogue.

**Elisabeth André** is a full professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University in Germany. She has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing, and social signal processing. In 2010, André was elected a member of the prestigious Academy of Europe and the German Academy of Sciences Leopoldina. To honor her achievements in bringing AI techniques to human-computer interaction, she was awarded a EurAI fellowship (European Coordinating Committee for Artificial Intelligence) in 2013. In 2017, she was elected to the CHI Academy, an honorary group of leaders in the field of human-computer interaction. Since January 2019, she has served as the Editor-in-Chief of the *IEEE Transactions on Affective Computing*.

**Myroslav Bachynskyi** is a human-computer interaction researcher at the Faculty for Mathematics, Physics, and Computer Science of Bayreuth University. His previous research positions were at Computer Science Faculties of Aarhus University and the University of Glasgow. He received his Ph.D. in Computer Science and Human-Computer Interaction at the University of Saarland and the Max Planck Institute for Informatics. He received his M.Sc. in Computer Science at the University of Saarland and his B.Sc. in Computer Science at the National University of “Kyiv-Mohyla

Academy.” His research focuses on development and application of data-driven methods to improve novel interaction techniques with a large space of alternative designs. He adopts optical motion capture, movement dynamics modeling, biomechanical modeling, and simulation besides standard performance measurement methods to uniformly cover the whole design space, formalize its performance and ergonomics properties in a mathematical model, and find the optimal solution.

**Jonas Beskow** (KTH) is a professor in speech communication with a focus on multimodal speech technology. In particular, his research interests include modeling and generating verbal and non-verbal communicative behavior as well as embodied conversational agents or social robots that use speech, gesture, and/or other modalities in order to accomplish human-like interaction. He is a cofounder of Furhat Robotics, a startup developing an innovative social robotics platform based on KTH research.

**Dan Bohus** is a Principal Researcher at Microsoft Research. His work focuses on the study and development of computational models for physically situated spoken language interaction and collaboration. Prior to Microsoft Research, Dan has obtained his Ph.D. from the Computer Science Department at Carnegie Mellon University, working on error handling in spoken dialogue systems.

**Angelo Cafaro** (Amadeus IT Group) was a postdoctoral researcher in the CNRS-ISIR laboratory of Sorbonne University. In the past years he conducted research in the area of embodied conversational agents with emphasis on social interaction, first impressions, and multimodal expression of interpersonal attitudes. He obtained his Ph.D. from Reykjavik University in 2014 with a dissertation about analysis and modeling of human nonverbal communicative behavior exhibited by a virtual agent in a first greeting encounter with the user. He contributed to a proposal for a SAIBA-compliant agent platform featuring a unified specification for the Function Markup Language (FML). More information is available on his personal webpage at [www.angelocafaro.info](http://www.angelocafaro.info).

**Nutan Chen** received his Ph.D., M.Eng., and B.Eng. from Technische Universität München, National Universitz of Singapore, and Dalian University of Technology, respectively. Since 2016, he has been a researcher with AI research, Volkswagen Data Lab. His main areas of interest are machine learning, robotics, and movement modeling.

**Deborah Dahl** (Conversational Technologies) is an independent consultant in speech recognition and natural language technologies. Her focus is on implementations of practical and scalable spoken dialogue systems at the boundary between

theory and applications. In addition to her technical work, Dr. Dahl has participated in the development of speech and multimodal standards at the World Wide Web Consortium since 1999. Her most recent book is *Multimodal Interaction with W3C Standards* (Springer, 2016).

**Stephen H. Fairclough** obtained his undergraduate degree in Psychology from the University of Central Lancashire in 1985. He received his Ph.D. from Loughborough University in 2000 where he worked as a Research Fellow at Human Sciences and Advanced Technology (HUSAT) Research Centre. He is currently a Professor of Psychophysiology in the School of Natural Sciences and Psychology at Liverpool John Moores University (LJMU). He has been involved in applied psychophysiological and neuroscientific research for over 25 years and is active in the area of physiological computing. His main areas of methodological expertise cover EEG, fNIRS, and cardiovascular in both laboratory and field settings. He is currently program co-chair for the International Conference on Physiological Computing, a director on the Executive of the Human Factors and Ergonomics Society (European Chapter) and a member of the editorial board for IEEE Transactions of Affective Computing. He has co-edited two collections on physiological computing research (both published by Springer-Verlag) and special issues of *Applied Ergonomics*, *ACM Transactions on Computer-Human Interaction*, *Interacting with Computers*, and *IEEE Computer*. His work has been published in *Nature*, *Biological Psychology*, *International Journal of Psychophysiology*, *Journal of Psychosomatic Health*, *Human Factors*, and *Interacting with Computers*.

**Michael Feld** is a Senior Researcher at the German Research Institute for Artificial Intelligence (DFKI GmbH). He is leading research and industry projects in the context of personalized and multimodal human-machine interaction in application domains including Advanced Driver Assistance Systems (ADAS), chatbots, and cyber-physical industry environments. His other focus topic is applied machine learning. He received his Ph.D. in 2011 on speaker classification for non-intrusive user modeling and is currently head of the DFKI ADAS living lab in Saarbrücken.

**Gerald Friedland** is Principal Data Scientist at Lawrence Livermore National Laboratory and also Adjunct Professor at the University of California, Berkeley. He leads a group of privacy and multimedia researchers, mostly focusing on acoustic analysis, methods for large-scale video retrieval, privacy concerns, and education. Dr. Friedland has published more than 200 peer-reviewed articles in conferences, journals, and books. He also authored a textbook on multimedia computing together with Dr. Ramesh Jain. He served as Associate Editor for *IEEE Multimedia Magazine* and

*ACM transaction on Multimedia.* He is the recipient of several research and industry recognitions, among them the European Academic Software Award and the Multi-media Entrepreneur Award by the German Federal Department of Economics. Dr. Friedland received his doctorate (summa cum laude) and Master's in Computer Science from Freie Universitaet Berlin, Germany, in 2002 and 2006, respectively.

**Joakim Gustafson** (KTH) is a professor in speech technology and head of the department of Speech, Music, and Hearing. He has been a prolific researcher and active systems developer of spoken and multimodal dialogue systems since 1993. He has an industrial background from Telia Research, where he led the research work of the EU project NICE, which developed a computer game where kids could interact with animated 3D characters using a combination of speech and gestures. He currently has three research projects where social robots act as third-hand helpers in assembly, as social skills coaches for autistic children, and as companions for the elderly with the task of detecting early signs of dementia.

**Dilek Hakkani-Tür** is a senior principal scientist at Amazon Alexa AI focusing on enabling natural dialogues with machines. Prior to joining Amazon, she was leading the dialogue research group at Google (2016–2018), a principal researcher at Microsoft Research (2010–2016), International Computer Science Institute (ICSI, 2006–2010), and AT&T Labs Research (2001–2005). She received her B.Sc. from Middle East Technical University in 1994, and M.Sc. and Ph.D. from Bilkent University, Department of Computer Engineering, in 1996 and 2000. Her research interests include conversational AI, natural language and speech processing, spoken dialogue systems, and machine learning for language processing. She holds over 70 patents and has co-authored more than 200 papers in natural language and speech processing. She is the recipient of three best paper awards for her work on active learning for dialogue systems, from the IEEE Signal Processing Society, ISCA, and EURASIP. She was a member of the IEEE Speech and Language Technical Committee (2009–2014), area editor for speech and language processing for Elsevier's *Digital Signal Processing Journal* and *IEEE Signal Processing Letters* (2011–2013), and currently serves on the ISCA Advisory Council (since 2015). She is the Editor-in-Chief of the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, and a fellow of the IEEE and ISCA.

**Alexis Heloir** is an assistant professor at the Université Polytechnique des Hauts de France and an affiliated researcher at the LAMIH laboratory UMR CNRS 8201 in Valenciennes, France. In 2004, he obtained his Master's in Computer Science from the University of Lille I. He received his Ph.D. from the University of Southern Brittany in 2008 on the specification and design of intelligible signing avatars.

He has been collaborating with the German Research Center for Artificial Intelligence (DFKI GmbH) since 2008 and became an independent group leader at the Multimodal Computing Excellence Cluster in late 2012. His current research interests are collaborative tangible user interfaces, gesture-based input, embodied conversational agents, as well as signing avatars.

**Rachel Hornung** received her M.Sc. in Robotics, Cognition, Intelligence—an interdisciplinary program combining engineering and cognitive science—from Technische Universität München in 2012. Since 2012 she has been a research assistant at the German Aerospace Center. Her research interests include adaptable assistive robotics, EMG-analysis, and machine and deep learning.

**Eric Horvitz** is a Technical Fellow and Director of Microsoft Research Labs. His research interests include multimodal inferences and systems, human-AI collaboration, conversational dialogue, and models of attention and engagement and their use in physically situated interactive systems. He earned his Ph.D. at Stanford University, focusing on computation and action under bounded resources.

**Michael Johnston** (Interactions Corporation) is Director of Research and Innovation at Interactions, where he manages research programs in conversational AI and machine learning. He has over 25 years of experience in AI, with a focus on applications to multimodal interaction, dialogue, natural language understanding, and human-assisted AI. Before joining Interactions, he held positions at AT&T Labs Research, Oregon Graduate Institute, and Brandeis University. Michael holds 60 U.S. patents, has published over 80 technical papers, and served as editor and chair of the W3C EMMA Multimodal Standard.

**Ronald Kaplan** is an Adjunct Professor of Linguistics at Stanford University. He has served as Chief Scientist for Amazon Search and Director of the Natural Language and Artificial Intelligence Laboratory at Nuance Communications, with a focus on dialogue and the conversational user interface. For many years he directed the Natural Language Theory and Technology research group at the Xerox Palo Alto Research Center. He created the modular architecture of Lexical Functional Grammar and introduced many of its formal devices for linguistic description. His research centers on the mathematical and computational properties of the LFG formalism and its ability to support well-motivated accounts for a wide range of linguistic phenomena. He is a past President and Fellow of the Association for Computational Linguistics, a co-recipient of the 1992 Software System Award of the Association for Computing Machinery, a Fellow of the ACM, and a Fellow of the Cognitive Science Society.

**Elsa A. Kirchner** graduated with a degree in Biology from the University of Bremen, Germany in 1999. Her diploma thesis was the result of a cooperation between the Brain Research Institute I: Behavioral Physiology and Developmental Neurobiology of the University of Bremen and the Department of Epileptology at the University Hospital of Bonn, Germany. From 1997–2000, she was a fellow of the Studienstiftung des Deutschen Volkes. During her studies, she received several other scholarships. With the help of the Stiftung Familie Klee award she was able to work as a guest researcher in the Department of Brain and Cognitive Sciences at MIT, Boston, MA, from 1999–2000. Since 2005, she has been a staff scientist at the Robotics Lab at the University of Bremen, Germany, leading the Brain and Behavioral Labs. Since 2008, she has led the team Interaction and since 2016 she has led the extended team Sustained Interaction and Learning at the Robotics Innovation Center of the German Research Center for Artificial Intelligence (DFKI GmbH) in Bremen, Germany. In 2014, she graduated (Dr. rer. nat.) in Computer Science from the University of Bremen. Her scientific interests focus on human-machine interaction, cognitive architectures, neuropsychology, and electrophysiological methods. She has supervised several Diploma, B.Sc., and M.Sc. students and served as a reviewer of international journals and conferences. Since 2018, she has been a member of Germany's "Platform for Artificial Intelligence" in the working group 6: Health Care, Medical Technology, Care. She is also author of more than 60 publications in peer-reviewed international journals and conferences as well as 7 book chapters.

**Frank Kirchner** studied computer science and neurobiology at the University of Bonn from 1989–1994, where he received his Diploma (Dipl. Inf.) in 1994 and graduated (Dr. rer. nat.) in Computer Science in 1999. In 1994, he was a Senior Scientist at the Gesellschaft für Mathematik und Datenverarbeitung (GMD) in Sankt Augustin, which became part of the Fraunhofer Society in 1999. In this position, Dr. Kirchner was the head of a team, focusing on cognitive and bio-inspired robotics. Moreover, he worked as a Senior Scientist at the Department of Electrical Engineering at Northeastern University in Boston in 1998. Dr. Kirchner was appointed Tenure Track Assistant Professor at Northeastern University in 1999. In this context, he was cooperating with several U.S. universities and research centers—among them the Massachusetts Institute of Technology (MIT) as well as the National Space Agency (NASA)—on a series of national German and US projects concentrating on autonomous intelligent robots in real-world applications. Since 2002 he has been full professor in the computer science faculty of the University of Bremen and since 2005 he has been CEO of the Robotics Innovation Center at DFKI GmbH. Prof. Kirchner is also the Scientific Director of the Brazilian Institute of Robotics. He is

a leading expert in research on biologically inspired locomotion and behavior of highly redundant, multifunctional robotic systems. He is the author of more than 300 publications, dealing in the field of robotics and AI (including space robotics). Furthermore, he is the principal supervisor for a number of Ph.D. students and regularly serves as a reviewer for a series of international scientific journals and conferences.

**Oliver Lemon** (Heriot-Watt University) is Director of the Interaction Lab and Professor of Computer Science at Heriot-Watt University, Edinburgh. His research focuses on conversational AI, machine learning approaches to spoken and multimodal interaction, NLP, and socially intelligent human-robot interaction. He was previously a research fellow at Stanford and Edinburgh universities and has led several national and international research projects; for example, developing in-car dialogue systems with BMW. He is an Associate Editor of *ACM Transactions on Interactive Intelligent Systems*, and was faculty advisor of Heriot-Watt's Amazon Alexa Challenge team "Alana," which was a finalist in both 2017 and 2018.

**Stacy C. Marsella** is a Professor at Glasgow University's Institute of Neuroscience and Psychology and Director of the Centre for Social, Cognitive, and Affective Neuroscience (cSCAN). He works on the computational modeling of cognition, emotion, and social behavior, both on basic research methodology in the study of human behavior as well as the use of these computational models in applications. Of special interest is the application of these models to the design of social simulations and virtual humans. As part of that work, he has released software in the public domain to lower the barrier of entry to virtual human research and crafting virtual human applications. He received the Association for Computing Machinery's (ACM/SIGART) 2010 Autonomous Agents Research Award.

**Robert Neßelrath** is Head of Interaction Technology at paragon semvox GmbH. He is responsible for human-machine interaction in conversational assistance systems with a focus on the automotive industry. Prior to that he worked at the German Research Center for Artificial Intelligence (DFKI GmbH) at the topic of multimodal dialogue systems in application areas including advanced driver assistance systems (ADAS), cyber-physical industry environments, and the smart home. He received his Ph.D. in 2016 from Saarland University.

**Fabrizio Nunnari** is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI GmbH) in Saabrücke, Germany. In 2001, he obtained a Master's in Computer Science from the University of Torino, Italy. From the same university, in 2005, he received his Ph.D. on the use of 3D data visualization for collaborative

work. Between 2006 and 2012, he worked as a researcher and lead developer at the Virtual Reality and Multimedia Park in Torino, Italy. Since 2013, he has been part of the Sign Language Synthesis and Interaction group at the Multimodal Computing and Interaction Excellence Cluster in Saarbrücken. His current research interests are on the generation of interactive virtual humans and on the synthetic animation of sign language avatars.

**Fabio Paternò** is Research Director at the Italian National Research Council, where he founded and leads the Laboratory on Human Interfaces in Information Systems at the Information Science and Technologies Institute in Pisa. He has been the scientific coordinator of several European Projects. His research interests include ubiquitous interfaces, methods and tools for accessibility and usability evaluation, frameworks for cross-device user interfaces, model-based design of interactive systems, and end-user development. He has published more than 200 papers in refereed international conference proceedings and journals. He was appointed ACM Distinguished Scientist and received the IFIP Silver Core award.

**Catherine Pelachaud** (CNRS-ISIR) is a Director of Research in the ISIR laboratory at Sorbonne University. Her research interests include embodied conversational agents, nonverbal communication (face, gaze, and gesture), expressive behaviors, and socio-emotional agents. With her research team, she has been developing an interactive virtual agent platform, Greta, that can display emotional and communicative behaviors. She has co-edited several books on virtual agents and emotion-oriented systems and has authored more than 200 articles. She was a recipient of the ACM SIGAI Autonomous Agents Research Award in 2015. Her Siggraph '94 paper received the Influential paper Award of IFAAMAS (the International Foundation for Autonomous Agents and Multiagent Systems).

**Dr.-Ing. Stefan Radomski** received his M.Sc. in Computer Science from the Leibniz University of Hanover. Subsequently, he worked at the Telecooperation Lab at the Technical University of Darmstadt where he was awarded with a Ph.D. for his work on multimodal interfaces in pervasive environments and respective description languages. Since 2017, he has been heading the data analytics group at the urban institute (<http://www.ui.city/en/>), where he established and continues to develop the company's Big & Fast data competency.

**Dirk Schnelle-Walka** led the "Talk&Touch" group at the Telecooperation Lab at Technische Universität Darmstadt until the end of 2014. Presently, he is a team lead for speech at HARMAN International Industries (previously Sennheiser) in the automotive industry, where he is taking his research portfolio to an industrial level.

At HARMAN, he works on intelligent personal assistants for the smart car. His personal research focus is on voice-centric multimodal interaction in smart spaces.

**Tim Schwartz** (German Research Center for Artificial Intelligence, DFKI GmbH) is a Senior Researcher and head of the Human-Robot Communications Group at DFKI Saarbrücken. His research interests include multimodal interaction and the coordination of and communication within hybrid human-robot teams in the context of Industrie 4.0. He has published over 30 scientific papers and is currently leading several projects funded by the German Federal Ministry of Education and Research.

**Gabriel Skantze** (KTH) is a professor in speech technology with a focus on dialogue systems. His research interests include real-time dialogue processing, generation and understanding of multimodal communicative behaviors, and modeling of phenomena such as turn-taking and joint attention in dialogue. His research in these areas over the last 15 years has resulted in multiple nominations and awards for best papers at conferences such as SigDial, ICMI, ICSR, and HRI. He is currently leading four research projects in related areas, including participation in the Amazon Alexa Prize challenge on conversational AI. He is also cofounder and software architect at Furhat Robotics, a company developing a platform for conversational robotics.

**David Traum** is the Director for Natural Language Research at the Institute for Creative Technologies (ICT) and Research Professor in the Department of Computer Science at the University of Southern California (USC). He leads the Natural Language Dialogue Group at ICT. (More information about the group can be found at <http://nld.ict.usc.edu/group/>.) Traum's research focuses on dialogue communication between human and artificial agents. He has engaged in theoretical, implementational, and empirical approaches to the problem, studying human-human natural language and multimodal dialogue, as well as building a number of dialogue systems to communicate with human users. Traum has authored over 250 refereed technical articles, is a founding editor of the journal *Dialogue and Discourse*, has chaired and served on many conference program committees, and is a past President of SIGDIAL, the international special interest group on discourse and dialogue. Traum earned his Ph.D. in Computer Science at the University of Rochester in 1994.

**Michael Carl Tschantz** received an Sc.B. from Brown University in 2005 and a Ph.D. from Carnegie Mellon University in 2012, both in Computer Science. Before becoming a researcher at the International Computer Science Institute in 2014, he

did two years of postdoctoral research at the University of California, Berkeley. He uses the models of AI and statistics to solve the problems of privacy and security. His interests also include experimental design, formal methods, and logics. His current research includes automating information flow experiments, circumventing censorship, and securing machine learning. His dissertation formalized and operationalized what it means to use information for a purpose.

**Raj Tumuluri** is the principal of Openstream Inc., one of the industry leaders in developing context-aware multimodal technologies for enhanced mobile user experience. Raj has been a champion of open standards that facilitate interoperability and is a co-author of the W3C MMI Architecture. He contributes to the R&D efforts of several international standards committees at the World Wide Web Consortium (W3C) and IEEE.

**Patrick van der Smagt** is director of the open-source Volkswagen Group AI Research Lab in Munich's DataLab, focusing on probabilistic deep learning for time series modeling, optimal control, reinforcement learning, robotics, and quantum machine learning. He previously directed a lab as a professor for machine learning and biomimetic robotics at the Technical University of Munich while leading the machine learning group at the research institute fortiss, and before that founded and headed the Assistive Robotics and Bionics Lab at the DLR Oberpfaffenhofen. Quite a bit earlier, he completed his Ph.D. and M.Sc. at Amsterdam's universities. Besides publishing numerous papers and patents on machine learning, robotics, and motor control, he has won a number of awards, including the 2013 Helmholtz-Association Erwin Schrödinger Award, the 2014 King-Sun Fu Memorial Award, the 2013 Harvard Medical School/MGH Martin Research Prize, and best-paper awards at machine learning and robotics conferences and journals. He is founding chairman of a not-for-profit organization for Assistive Robotics for tetraplegics and co-founder of various tech companies.

**Michel Valstar** is an Associate Professor in Computer Science at the University of Nottingham, and a member of both the Computer Vision and Mixed Reality Labs. He received his Master's in Electrical Engineering at Delft University of Technology in 2005 and his Ph.D. at Imperial College London in 2008 on the topic of the timing of facial expressions. He is an expert in the fields of computervision and pattern recognition, where his main interest and world-leading work is in automatic recognition of human behavior. Valstar pioneered the concept of Behaviomedics, which aims to diagnose, monitor, and treat medical conditions that alter expressive behavior by employing objective assessment of that behavior. He is the founder of

the facial expression recognition challenges, FERA 2011/2015/2017, and the Audio-Visual Emotion recognition Challenge series, AVEC 2011–2018. He leads the Objective Assessment research area of the Nottingham Biomedical Research Centre and was the coordinator of the EU Horizon 2020 project ARIA-VALUSPA. Valstar is a recipient of Melinda and Bill Gates Foundation funding to help premature babies survive in the developing world. His work has received popular press coverage in *The Guardian*, *Science Magazine*, *New Scientist*, CBC, and on BBC Radio, among others. Valstar is a senior member of the IEEE and published over 90 peer-reviewed articles.

**Dr. Alexander Waibel** is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the Karlsruhe Institute of Technology, Germany. He is the director of the International Center for Advanced Communication Technologies (interACT). Waibel's pioneering work on the "Time-Delay Neural Network" (1987) was the first "convolutional" neural network, a learning approach that is now at the heart of most AI technologies. Waibel founded and served as chairmen of C-STAR, the Consortium for Speech Translation Advanced Research in 1991. He directed many research programs in machine perception, interpretation, and machine learning in the U.S., Europe, and Asia, including EU-IPs EU-Bridge (2012–2015) and CHIL (2004–2007). From 2010–2013, he co-directed the German side of the French-German research initiative Quaero, and IMMI, a French-German joint venture between KIT, CNRS & RWTH, devoted to machine intelligence and multimedia. Dr. Waibel is a member of the National Academy of Sciences of Germany. He was named Honorary Senator of the Hochschulrektorenkonferenz. He is a Fellow of the IEEE and has published extensively (more than 800 publications, more than 28,000 citations, h-index 82) in the field and holds many patents. During his career, Dr. Waibel founded and built 10 successful companies.

**Massimo Zancanaro** is the head of the i3 (intelligent interfaces and interaction) research unit at FBK. His primary research interest is in the field of computer-human interaction and in particular intelligent user interfaces. He co-edited two books and co-authored more than 100 papers in research journals (among them *Artificial Intelligence*, *User Modeling*, *TOCHI*, *IJHCS*) and peer-reviewed conferences (including IJCAI, CHI, IUI, UMAP). He teaches a course on graphical user interface design at the University of Trento and is currently a member of the Faculty of the Doctoral School on Cognitive Science. At present, he serves as chair for the Italian Chapter of the ACM Special Interest Group in Computer-Human Interaction.



## Volume 3 Glossary

**Abstract User Interface (AUI).** User interface description in terms of elements that are independent of the possible interaction modalities.

**Active BCI** is a brain-computer interface that derives its outputs based upon a voluntary act of explicit control from the human, e.g., generates motor imagery consistent with movement of right hand to move cursor to the right.

**Active state** is a psychological state associated with a volitional act or intention generated by the person, e.g., to open the door.

**Addressee detection** is the detection of who a user is addressing when speaking (the system/robot or another user).

**Aggregation** refers to two concepts: (i) the creation of aggregate data and (ii) the combining of information from multiple sources to enable new inferences, similar to those used in *linkage attacks*. Solove [2006] uses the second sense when referring to privacy violations that result from *information processing* that combines diffuse pieces of information together to make new inappropriate inferences possible. Given that aggregate data tends to be less privacy sensitive than microdata, the two uses of the term are almost at odds.

**Alignment** is the coordination along one modality (e.g., verbal modality) that occurs when interlocutors reach a common understanding [Pickering and Garrod 2004].

**Anonymization** modifies data to make it unlinkable to the people measured by the data. Often attempts at anonymization fail and *de-anonymization* is possible.

**Application domains** include serious games, conversational agents, or dialogue systems for healthy behavior promotion; intelligent interactive monitoring of patient's environment and needs; intelligent interfaces supporting access to healthcare services; patient-tailored decision support, explanation for informed consent, and retrieval and summarization of on-line healthcare information;

risk communication and visualization; tailored access to electronic medical records; tailoring health information for low-literacy, low-numeracy, or underserved audiences; virtual healthcare counselors; and virtual patients for training healthcare professionals.

In addition, we address decision support systems especially for the doctor, which model the diagnostic reasoning and decision-making of medical experts, and systems designed to interact directly with patients as healthcare consumers.

**Application Programming Interface (API).** Set of procedures made available by a software application to provide services to external programs.

**Application-specific safety level** describes the concept to include information into the robot's behavioral control that comes from sensors that are not part of the robot itself. This concept uses the information from sensors that are placed outside of the robot to monitor the environment, e.g., a workplace in a production line and which are used in more traditional applications to create strict safety boundaries around the workplace. In more advanced approaches this information is used differently; here it helps to derive contextual information that can be used to adapt the robot's behavior instead of overwriting it. For example, in a more traditional scenario a violation of the safety boundary by a human walking by would result in a full stop of the production line. The more advanced concepts would predict the human's path and instead of stopping the production line would only reduce the speed of the moving robots. This concept therefore modifies the behavior of the robot as commanded by the *high-level control* module instead of overwriting it.

An **Attributed Relational Graph** (ARG) is a graph structure augmented with attributes presenting information regarding nodes in the graph. Attributed relational graphs have been used in computer vision and in work on multimodal reference resolution [Chai et al. 2004] where the graph represents a sequence of possible referents from either gestures or the discourse context and the attributes contain properties of the referents (e.g., color, size).

An **Augmented Transition Network** (ATN) is a *recursive transition network* augmented with a set of procedural operations and variables. ATNs have been applied to numerous tasks in natural language processing including parsing [Woods 1970].

**Assignment** expresses the absence of choice. Exactly one specific modality can be used in order to reach a goal. An example is the steering wheel of a car.

An **autoencoder (AE)** is an unsupervised neural network that has the same output as input. Using a latent layer with fewer dimensions than the original data, the network is forced to compress the data and find a lower-dimensional representation for the data. The net can be arbitrarily deep and transformations between layers can be nonlinear.

**Automatic speech recognition:** the signal spoken in language is recorded by microphone, processed, and converted to text (speech to text).

**Auxiliary information** is side information in addition to supposedly *anonymized* data used in an attempt to *de-anonymize* it. The difficulty of predicting the available auxiliary information makes preventing de-anonymization difficult.

**Awareness** can refer to perceiving sensory stimuli in the environment including other actors, which may be machines or peoples.

**Backchannel** is a brief feedback (a very short utterance like “mhm” or a gesture such as a head nod) that the listener gives without intending to take the floor.

**Beat gesture** is “defined as movements that do not present a discernible meaning, [ . . . ]. They are typically biphasic (two movement components ), small, low energy, rapid flicks of the fingers or hand” [McNeill 1992].

**Behavior Markup Language (BML)**. An XML-based language for describing behaviors that should be realized by animated agents.

**Biocybernetic control** describes a model of *closed-loop control* (negative or positive control) wherein measures are derived from psychophysiological or neurophysiological sources and converted into control input for an adaptive computer system.

**Biometric Liveness measures.** In order to avoid spoofing, a biometric system can sense whether the data it is receiving is in fact coming from a live person (vs. a cadaver, or a photograph). Example measures include heart rate, heat, etc.

**Biometric spoofing.** The provision of replicas of a person’s biometric data, such as a photograph or plaster cast, in order to cause a biometric algorithm to falsely recognize the presence of that person.

**Biosignals** provide information from a person’s biological or physiological structures and their dynamics. Signals measured from the human body typically originate from neural or muscular activity. Neural activity is captured by methods such as EEG, electroencephalogram, a test that detects electrical activity in your brain using small, flat metal electrodes attached to your scalp. Muscular activity is captured by methods such as electromyogram (EMG) electric signals

generated by muscles, or electrocardiogram (ECG) electric signals emitted from the human heart. They are the basis for human computing, physiological computing and affective computing. See also [Silva et al. \[2015\]](#). For applications in human computer interaction (HCI) and intelligent user interfaces (IUI), only surface electrodes are used. Signal processing includes, first, time series analysis and, second, the mapping to physical or physiological states [[Schuller 2018](#), [D'Mello et al. 2018](#), [Wagner and André 2018](#), [Martin et al. 2018](#)] toward cognitive states [[Zhou et al. 2018](#), [Cohn et al. 2018](#), [Oviatt et al. 2018a](#)]. Biosignals of future interest include electric conductance, bioimpedance, and bioacoustic signals.

**Chameleon effect** “refers to non-conscious mimicry of the postures, mannerisms, facial expressions, and other behaviors of one’s interaction partners, such that one’s behavior passively and unintentionally changes to match that of others in one’s current social environment” [[Chartrand and Bargh 1999](#)].

**Chart parsing** is an algorithm for parsing strings with respect to a grammar. It uses a [dynamic programming](#) approach in which partial hypothesized results are stored in a tabular representation for re-use. This avoids re-computation of parses of sub-trees shared among multiple parses and avoids the combinatorial explosion that can result from a brute force approach to application of rules in order to construct a parse.

**Closed-loop control** is a control system that uses the concept of an open-loop system as its forward path but has one or more feedback loops (hence its name) or paths between its output and its input.

**Closed-loop fusion.** Based on mission/pilot objectives, an autonomous sensor manager will analyse fused track information in order to prioritize sensor tasks, including maintaining the sensing of current tracks vs. searching for new ones. Thus, sensor fusion leads to creation and prioritization of new sensor fusion tasks.

**Code switching:** mixing words from different languages, declination rules and compounding.

**Cognitive load** expresses a user’s mental load and mental effort to solve a given problem. Mental load is imposed by problem’s parameters such as task and structure of sequence of information. Mental effort refers to the amount of capacity that is allocated to the problem’s demands. [[Sweller 1988](#), [Paas 1992](#)]

**Collaboration.** Dialogues often are expected to further participants’ plans and goals, rather than just respond literally to the overt utterances. On the other hand, dialogues such as legal proceedings or negotiations can be adversarial

or non-collaborative dialogues, in that some of the participants' goals are not shared, but the parties still obey collaborative conversational rules.

**Complementary** modalities are used in complementary manner within a temporal window for reaching a goal, i.e., both modalities are needed to describe the desired meaning. A speak-and-point system is a classic example of this ("change the color of this (*pointing gesture*) item to blue").

**Concept accuracy** is a measure of the semantic correctness of input processed by an interactive system. This is in contrast to word accuracy and sentence accuracy that are measures of the accuracy of transcription of words by speech or handwriting recognizers. At the utterance level, the interpretation of an input is *concept accurate* if the meaning captured by the system corresponds to the user's intended meaning. Sub-turn measures of concept accuracy involving, for example, measures such as number of individual concepts successfully interpreted, are also used but need to be carefully defined as there is no generally accepted standard for assigning partial credit in determining the interpretation of an utterance.

**Concrete User Interface (CUI)**. User interface description in terms of elements that are modality-dependent but implementation language independent.

**Consecutive interpretation** typically interprets a few sentences, one at a time, before giving the dialogue partner a chance to respond.

**Contextual integrity** is a philosophical theory of privacy concentrating on how privacy norms about information sharing depend upon the social context in which the information is shared [Nissenbaum 2009].

**Coping** is the process of dealing with emotion, either by acting externally on the world (problem-focused coping), or by acting internally to change beliefs or attention (emotion-focused coping).

**Co-reference Resolution or Reference Resolution** is the process of determining for a given word or phrase which other word or phrase, or entity in the application or environment that word or phrase refers to. Reference resolution may apply within an utterance, between utterances, or between a phrase and an entity presented visually in a graphical interface or present within the physical environment of the listener or in which an interactive system operates, the latter case is referred to as *multimodal reference resolution*.

**Covert measure** is a measure of human behavior or performance that cannot be detected based upon human perception, e.g., heart rate, and brain activity.

**Cross-lingual subtitling:** a mixture of consecutive and simultaneous interpretation where interpretation is performed on media content and delivered textually as subtitles.

**Cybercasing** uses technology to determine which buildings are uninhabited and, therefore, prime targets for crime.

**Cyber-physical Environments** are characterized by a large number of individual systems and devices with their sensors and actuators, and shift the interaction paradigm from the user's perspective toward system-environment interaction.

**De-anonymization** takes data thought to be anonymized and reverses its attempted *anonymization* for the purpose of *identification*, that is, to link the data to the people measured by the data. Typically, de-anonymization uses a *linkage attack*, which compares the available information in a record for a person to various records provided as *auxiliary information*, looking for a match.

**Decision-theoretic:** reasoning methodology for selecting ideal actions in accordance with the principles of probability and utility; decision-theoretic reasoning involves guiding actions by expected utility. Expected utilities of actions are computed by coupling probabilities, inferred about current or future states, with considerations of the value of outcomes (see [Horvitz et al. \[1988\]](#)).

**Decluttering of a situational display.** Using multisensory fusion techniques to simplify a display by combining tracks that possibly represent single entities. Decluttering has to accommodate multisensory information from the onboard sensor suite, as well similar information being propagated from other sensing platforms (aircraft, ground stations, etc.).

**Deep learning** refers to a subgroup of machine learning algorithms. It comprises methods that use several processing layers with (non)linear transformations. Neural networks that include at least one hidden layer belong to this group. The methods learn different representations of the data autonomously. These representations differ based on the task they are intended to solve and the architecture selected.

**Degrees of freedom (DOF)** of a mechanical system describes how a normalized mechanical junction can move. It usually defines a set of rotational axes together with rotation boundaries.

**Deictic gestures** are pointing gestures which single out an object of interest (usually with the index finger, but also with gaze). These may be accompanied with a *deictic expression* (or *deixis*), such as "this," "that," "these," or "those."

**Deictic reference** is a subtype of reference where a word or phrase cannot be fully understood without additional contextual information, such as the identity of the speaker (“me”) or their location (“here”) or something they are pointing at (“that dog”).

**De-identification** is a method that attempts the *anonymization* of microdata by removing identifiers from records. De-identification often fails to anonymize data because data fields that do not appear to be identifiers can often, in combination with one another and with *auxiliary information* identify people in surprising ways.

**Dempster-Shafer Theory (DST)** is a mathematical theory of evidence that is often used for sensor fusion [Calderwood et al. 2017, Murphy 1998] because it enables one to combine evidential information from different independent sources without requiring prior probabilities. The theory is based on belief functions (in DST, called “mass functions”) that assign a degree of belief (mass) to *sets* of propositions. DST starts with a “frame of discernment”  $\Omega$ , a set of mutually exclusive and exhaustive possible states of the system being studied. For example,  $\Omega$  might consist of airliner, military jet, helicopter meaning that the system is sensing one of those aircraft. A belief function  $m$  will assign evidential belief mass (perhaps zero) to the elements of the powerset of  $\Omega$ . To continue the example, the mass function may assign evidential mass to airliner, helicopter, meaning that it is assigning evidence to its sensing either an airliner or a helicopter. Evidence from multiple independent mass functions  $m_1, m_2$  (e.g., multiple sensors) may be combined using Dempster’s rule, which sums the product of those mass functions  $m_1(X)m_2(Y)$  evaluated over the non-null intersection of all sets  $X$  and  $Y$  from  $\Omega$ . Dempster’s rule ignores conflicting evidence between belief functions, i.e., where  $X$  and  $Y$  do not overlap, through a normalization factor that has led some to argue that this rule can provide counterintuitive results [Zadeh 1986]. Numerous attempts have been made to overcome these problems with different rules of combination while maintaining the benefits of the approach [Haenni 2004, Sentz and Ferson 2002].

**Destructive unification** is an operation that combines information from two feature structures. In cases where a path in the feature structure does not unify the value from the second feature structure is taken.

**Dialogue Applications** are dialogue systems that are written for specific application use-cases.

**Dialog context:** information contained in previous utterances that is relevant to and can help determine how the conversation progresses.

**Dialogue Management** controls the flow of the dialogue with a computer. It basically updates the context and determines how to react to dialogue acts.

**Dialogue Platforms** are underlying frameworks that are used to execute a dialogue system.

**Dialogue purpose.** Dialogue purposes can range from helping the user or system to perform a task, provide information, to social dialogues, in which the participants are making friendships, building rapport and trust, exploring their likes and dislikes, flirting, arguing, telling jokes, passing the time, etc.

**Dialogue Systems** are software agents that allow users to converse with a machine in a coherent structure.

**Differential privacy** is an approach to *anonymization* and *statistical disclosure limitation*, which ensures that the released data will not depend much on any one person's data [Dwork et al. 2006, Dwork 2006]. Differential privacy does particularly well on aggregate data.

A **Directed Acyclic Graph (DAG)** is a graph consisting of a finite number of nodes and edges, where the edges are directed from one node to another, and such that there is no way to start at a particular node and loop back to the same node by following a sequence of directed edges (that is, there are no cycles in the graph). DAGs can be used to model many different types of information. They provide the formal mathematical underpinning of feature structures.

**Discourse** describes the sequence of dialogue acts, which in turn are a communicative function unit.

**Dropout** is a technique applied in neural network learning. Some of the in- or output neurons are artificially and randomly set to zero. This way the algorithm has to learn to handle corrupted data and prevents coadaptation of neurons, while improving the detection of correlations. Dropout can also be used to learn correlations between different modalities.

**Dyadic:** a term denoting an interaction that involves two participants. Most work in spoken dialog systems has traditionally focused on dyadic settings, where a dialog system interacts with a single human user.

**Dyadic vs. Multi-party dialogues.** In addition to the dyadic (two-party) case, a system may act as a full-fledged dialogue participant in a multi-way conversation,

requiring it to process and respond to human-human dialogues in addition to responding to utterances addressed to it.

**A dynamic movement primitive (DMP)** is a nonlinear dynamic system trained from demonstration of a trajectory. Using a point attractor in a second-order dynamic system a parametrizable description of motion can be obtained.

**Dynamic programming** is an approach to solving a complex computational problem by exploiting the structure of a problem in order to break it down into more simple sub-problems, storing their solutions, so they can be re-used without re-computation. In *chart parsing* for example, the parse trees for sub-constituents of a parse are stored and re-used in multiple different candidate parses of the input.

**Early fusion.** After possibly preprocessing data of different modalities, the data is merged and a model for the mixed data is calculated from it.

**Earplugs and pixel-buds:** a set of ear-plugs provides input and output for a speaker attempting to dialogue with others.

**Echo cancellation** is the signal processing step of removing the system's (robot's) own speech (and its room-acoustic echoes) from the audio that comes into the microphones.

An **Edit Transducer** or **edit machine** is a *finite-state transducer* whose function is to transform a transducer (representing the input) into another transducer (representing the output) through a series of edit operations such as deletion, insertion, and substitution of symbols. These operations are captured using transitions in the edit transducer. For example, deletion of the symbol  $x$  can be represented using the transition  $x : \epsilon$  while insertion of  $x$  can be represented as  $\epsilon : x$ , where  $\epsilon$  is the epsilon transition. Substitution of  $x$  with  $y$  is represented using the transition  $x : y$ . The edit operation is achieved through *finite-state composition* of the input transducer  $I$  with an edit transducer  $E$ , annotated as  $I \circ E$ . Edit operations may be weighted and this captured by associating costs with arcs in the edit transducer.

**Electromyography:** electrodiagnostic medicine technique for recording the electrical activity produced by muscles.

**Electromyography measurements (EMG)** provide quantitative data on real muscle recruitment. They were often used in physical ergonomics assessment of desktop interfaces, in particular mouse and keyboard. However, they are ineffective and inefficient for post-desktop interface evaluation, as they are

limited to only close-to-surface muscles, and they suffer from cross-talk and muscle belly drift in dynamic movements, thus providing unreliable data.

An **electronic medical record** (EMR) is a narrower view of a patient's medical history including laboratory values, for example, while an **electronic health record** (EHR) is a more comprehensive report of the patient's overall health.

**Embedded brain reading** is an approach for user state detection, which is based on the online analysis of brain activity. Brain activity is used which is spontaneously evoked during human-machine interaction. The approach is deeply embedded into the system's control, the context of interaction, and makes use of multimodal data. It is applied for implicit interaction, i.e., to non-intentionally adapt or drive explicit interaction.

**Embedded multimodal interface** is an interface that makes use of **multimodal data** from **multimodal input** and is able to generate **multimodal output**. Its main characteristic is that it is deeply incorporated into the control of the robotic system, and may be subject to complex adaptation mechanisms such as *reflexive adaptation*. While its function might be to gain explicit control of a system, it might be subject to implicit control to be adapted to the human's or system's needs.

**Emblem** is a symbolic gesture, such as the thumbs up gesture, where the gesture bears no direct resemblance to what it signifies.

**Embodied Conversational Agent** is a virtual or robotic human-like character that demonstrates many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication.

**Emotion Markup Language (EmotionML)**. An XML markup language for describing emotion, standardized by the W3C Multimodal Interaction Working Group.

**Empathy** is the capacity to put oneself in the shoes of another one. Emotional empathy refers to feel what the other is feeling while cognitive empathy to understand what the other feels by taking his perspective [Paiva et al. 2004].

**Encoding dictionary**: captures a set of templates, i.e., transformations, which can be applied to a feature to better capture the relationship between this feature and events on other modalities [Morency et al. 2008]. Examples include temporally shifting the location of the feature activation by an offset to enable reasoning about potential coordinative delays between activities on different channels, or the use of ramp functions in cases where the influence on the target variable is expected to be changing over time.

**Engagement:** we adopt here the definition proposed by [Sidner et al. \[2004\]](#) as “the process by which two (or more) participants establish, maintain, and end their perceived connection.” Situated interaction systems generally need to reason about and manage engagement, i.e., who they are interacting with, and when.

**Equivalence** expresses the concept of free choice of modality. Multiple modalities can reach the same goal and it is sufficient to use only one of them without any temporal constraint on them.

**Ergonomics** is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance.

**Exclusion** is the privacy violation that exists when data subjects do not know enough about how their information will be used [[Solove 2006](#)].

**Explicit control** represents a mode of input control where the user intentionally generates a specific behavior in order to achieve a specified goal, e.g., move a cursor upward.

**Explicit interaction** is a mode of human-computer interaction where the human user is fully cognizant of the issuing of commands and receives explicit feedback from the computer.

**Extensible Multimodal Annotation (EMMA)**. An XML markup language for describing the results of multimodal processors such as speech recognition, image recognition, and natural language understanding, standardized by the W3C Multimodal Interaction Working Group.

**F-formation:** a term used to denote the spatial pattern in which participants arrange themselves during interactions. Per [[Kendon 1990b](#)], “*An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.*” The prototypical example is a circular pattern, with participants oriented toward a common center, although other configurations such as L-shaped, V-shaped, side-by-side, and vis-à-vis are common.

**Feature structures** are a formal representation used in a variety of different grammatical formalisms including head-driven phrase structure grammars [[Pollard and Sag 1994](#)] and lexical functional grammar [[Kaplan and Bresnan 1995](#)]. A feature structure consists of a set of attribute-value pairs. The values may be atomic or feature structures themselves. Numerical indices in a feature structure indicate that the values of particular feature paths must be equal, in

the sense that they must be unified (see *unification* below). Feature structures (sometimes called attribute-value matrices) have an underlying interpretation as *directed acyclic graphs* (DAGs) where the arcs are labeled with the features and the nodes contain the values. Use of numerical indices in feature structure notation corresponds to shared structure in the underlying directed acyclic graph.

A **finite-state automaton**, or **finite-state acceptor** is a finite state machine that operates over a single stream of input symbols. A finite state automaton defines a set of strings corresponding to all of the possible paths from the state space it defines. Formally, a finite-state automaton is a 5-tuple  $\langle S, I, B, F, T \rangle$  where  $S$  is a finite set of states,  $I$  is a finite set of symbols,  $B$  is a finite set of beginning states, and  $F$  is a finite set of ending states.  $T$  is a finite set of transitions between states, each labeled with a symbol from  $I$  or the epsilon ( $\epsilon$ ) symbol. Finite-state automata may also be weighted with costs associated with arcs and states.

**Finite-state composition** is an operation on a pair of finite-state transducers. Given as input a finite-state transducer  $T$  with the input symbol alphabet  $I$  and output symbol alphabet  $O$  and a finite-state transducer  $R$  with the input symbols alphabet  $O$  and output symbols  $O_2$ , there is a composition of the two finite-state transducers if there is a string  $y$  drawn from the alphabet  $O$  that  $T$  produces as output and  $R$  accepts as input. Finite-state composition of  $T$  and  $R$  is represented as  $T \circ R$ .

A **finite-state transducer** is a finite-state machine that operates over two streams, an input stream, and an output stream. Formally, a finite-state transducer is a 6-tuple  $\langle S, I, O, B, F, T \rangle$  where  $S$  is a finite set of states,  $I$  is a finite set of input symbols,  $O$  is a finite set of output symbols,  $B$  is a finite set of beginning states and  $F$  is a finite set of ending states.  $T$  is a finite set of transitions between states, each labeled with an input label from  $I$  or the epsilon ( $\epsilon$ ) symbol and an output label from  $O$  or the epsilon ( $\epsilon$ ) symbol. Finite-state transducers may also be weighted with costs associated with arcs and states. In weighted transducers the best path is the path with the lowest overall cost.

**Foundational technologies** are introduced in other chapters in Volumes 1 and 2 of this handbook, namely machine learning [[Panagakis et al. 2018](#), [Baltrušaitis et al. 2018a](#)], deep learning [[Keren et al. 2018](#), [Bengio et al. 2018](#)], and knowledge management [[Alpaydin 2018](#)].

**Frames** are a data structure that have been used in a variety of artificial intelligence applications including natural language processing, computer vision, and

knowledge representation and reasoning. A frame consists of a series of slot names and values that define a stereotypical object. In frame representations, typically the value of a frame slot may itself be another frame. The term *frame* was introduced by Minsky [1974]. Frames are similar to feature structures in that they consist of sets of attributes and values with the potential to embed further sets of attributes and values within values. Feature structures differ from frames in the capability of structure sharing and the underlying mathematical foundation of feature structures as directed acyclic graphs.

**Function Markup Language** is an XML-like mark up language specially suited for representing communicative intentions.

A **Gaussian process (GP)** is a stochastic process defined by its mean and covariance. Assuming that similar inputs behave similarly, test data targets will be similar to closely located input data targets, based on multidimensional Gaussian distributions.

**Gaussian process dynamical models (GPDMs)** are based on the *Gaussian process latent variable models (GPLVM)*. By incorporating temporal dependency, they enforce a smooth latent space.

**Gaussian process latent variable models (GPLVMs)** are based on the *Gaussian process (GP)*. They generate a low-dimensional representation of the input space.

**Gestures** are inputs provided by a user through body movement, usually via hands and/or arms. In some instances, gestures are produced through contact with a display and involve touch or drawing on a screen using a finger or stylus (sometimes called touch gesture, pen gesture, or ink gesture). In other instances, gestures involve hand and arm movement in space and do not involve contact with a surface.

**Gesture phases** is a type of communicative gesture [McNeill 1992]. The optional phase preparation brings the hand into the gesture space. This may be followed by a *pre-stroke hold* where the hand holds its position until the stroke. The *stroke* corresponds to the forceful part of the gesture. It carries the meaning of the gesture and is synchronized with the linguistic segments it “coexpressed.” A *post-stroke hold* phase may follow where the hand remains in its position. Finally, the hand may come to a rest position within the *relaxation* phase.

**Gesture recognition** is the process of automatically classifying and interpreting gestures made to an interactive system. The means of capturing and recognizing a gesture depends on the particular type of gesture. To recognize a drawing made on a display for example the trace made by a finger or stylus can be captured

directly and the resulting series of *strokes* can be fed into a classification algorithm. For other types of gesture without direct physical contact with a touch-sensitive surface, video cameras or other types of sensors, some of which are worn by the user, can be used to track the location of the hand and arm.

**Goniometer.** A goniometer is an instrument which measures an angle.

**Grounding**, as defined by [Clark \[1996\]](#), is the process by which speakers reach a common understanding (adding to their *common ground*). This should not be confused with the notion of *symbol grounding*, which is sometimes used in robotics, and denotes the problem of how words should be linked to objects in the real world.

**High-level control** refers to the specification and feedback control of target positions in 3D space that must be reached by the end effector of a multi-joint robot based on information coming from sensors sampling the robot itself and its environment.

**Human-robot cooperation** is a subfield of human-robot interaction where a human and robot or teams of humans and robots work or act together to reach a shared goal. It often requires direct contact between human and robot or a shared workspace.

**Human-robot interaction** is any interaction between a human and a robot or teams of humans and robots including communication, control, feedback, direct contact, or information exchange.

**Hybrid BCI** describes a brain-computer interface that combines *active BCI* with either *passive* or *reactive BCI* or other measures such as eye movements or heart rate.

**Hypertext Markup Language (HTML)**. A graphical markup language for defining web pages.

**Hypertext Transfer Protocol (HTTP)**. An application protocol commonly used in the World Wide Web. A protocol is a set of rules to exchange data among different applications. HTTP dictates as web browsers and other similar applications can access web content.

**Iconic gesture** is a gesture which bears a resemblance to what it signifies, such as showing a round shape by tracing a circle in the air.

**Ideational unit** a unit that makes up discourse and that may span over several gestures that convey related information.

**Identification**, a form of *information processing*, links information about a person to that person's identity by way of an *identifier*, and can be a privacy violation [Solove 2006].

**Identifiers** are types of data that pick out the person or entity about whom the record is concerned. Sometimes the term is restricted to data fields of a record that obviously does so, as in *De-identification*. However, subtle patterns in data can serve as unintended identifiers, making *de-anonymization* possible.

**Imitation** is the production of a behavior, be it verbal or motor, with a learning or a communicative goal that was perceived earlier [Nadel and Butterworth 1999].

**Implicit interaction** is a mode of human-robot interaction where the human user is not aware of the issuing of (control) commands that may be used for the control of a technical system or adaptation of an interface to the needs of the technical system or user. The user may or may not receive explicit feedback from the computer.

**Increased accessibility** can be a privacy violation when it inappropriately makes information easier to obtain by those who should not know it [Solove 2006].

**Incremental processing** refers to the ability of a system to provide feedback or interpretation on the input that it is receiving before the input has completed. One example is incremental speech recognition, where a visual interface provides feedback on recognized words as the user speaks.

**Industrial design** deals with design and development of hardware and physical input artifacts, for example mouse, keyboard or joysticks, and their appropriate physical ergonomics assessment.

**Inertia matrices** describe inertial properties such as total mass and mass distribution of the rigid segments of the human body, and are involved in computation of joint moments for given kinematics through Newton's law.

**Information collection** can be a privacy violation when it inappropriately makes observations through surveillance or seeks information through interrogation [Solove 2006].

**Information dissemination** shares information with people who might have otherwise not have had access to it, which can violate privacy, for example, through *increased accessibility* [Solove 2006].

**Information processing** converts and combines data in ways that can violate privacy, including by *aggregation* and by *identification* [Solove 2006].

An **infotainment system** combines entertainment and information in cars for drivers and passengers. It originated in car audio systems but now also includes navigation systems, telephony, and comparable functionality.

**Ink Markup Language (InkML).** An XML markup language for describing digital ink traces and their properties, standardized by the W3C Multimodal Interaction Working Group.

**Interaction Manager (IM).** A component in the W3C Multimodal Architecture that coordinates operations among modality components.

**Internal state of a robot** is computed on the basis of all sensor information directly or indirectly available to the robot. Directly available information is all information that comes from the robot's own sensors, while indirectly available information is all information that comes from sensors that are external to the robot but that the robot can access through communication pathways. The set of internal states of a robot is in most cases a finite set of eventually multi-dimensional vectors. Elements of this set are computed through means of clustering that range from simple thresholds to complex statistical methods.

**Invasions** is the privacy violation that happens when people experience inappropriate interference with the private sphere of their lives [[Solove 2006](#)].

**Inverse dynamics** is a method for computing forces and/or moments of force (torques) based on the kinematics (motion) of a body and the body's inertial properties (mass and moment of inertia).

**Inverse kinematics** is an algorithm which resolves joint parameters for a skeletal structure given a specific set of constraints: for example, in the context of biomechanics, joint angles are resolved given kinematics of a set of markers attached to the body segments and recorded by the motion capture system.

**JANUS** system was the first speech translation system presented to the public in the USA and Europe in 1991.

**JavaScript.** A programming language used in conjunction with HTML for defining the behavior of web pages.

**Joint attention** means that the speakers are attending to the same object and are mutually aware of it.

**Joint inference:** an inference model that reasons jointly about multiple entities and that produces a probability distribution over the joint space (cardinal product) of the variables of interest. This stands in contrast to inference models that

reason independently about each entity, and that produce a separate probability distribution for each variable of interest.

***k*-Anonymity** is a syntactic approach to *anonymization* that ensures that no one record is unique in released microdata [Samarati and Sweeney 1998a, 1998b, Sweeney 2002a]. This approach and its extensions has lost ground to *Differential privacy*, at least for aggregate data.

**Kalman Filter** A Kalman filter [Kalman 1960] is a continuously iterative estimation algorithm used to predict future states of a continuously varying system or process from potentially noisy measurements of the current state. Because it naturally incorporates multiple sensors, it is often referred to as a sensor fusion algorithm. The standard Kalman filter predicts the system and minimizes the noise (hence the term “filter”) optimally if the measurement and noise are Gaussian and the sensors are linear. Variants of the Kalman Filter, such as Extended Kalman Filters [Pfeiffer and Franke 2010] and Unscented Kalman Filters [Allodi et al. 2016, Julier and Uhlmann 2004], are used when the system being modelled behaves nonlinearly. Models of the covariance of multiple sensors are used to accomplish sensor fusion.

**Late fusion.** Data of multiple modalities is processed for each modality individually. The decisions for each modality are merged based on their probabilities or other means of comparison or by applying a further learning algorithm for merging. The final result is based on data of all modalities. However, correlations between modalities are not exploited.

**Latent representation** substitutes data in a lower dimensional space than the original data. It is capable of explaining the changes in the data. If the mapping between the independent factors accountable for changes in the observed data, and the observed data is known or can be learned, a latent representation should have that intrinsic dimensionality as it is the minimal description of the data.

A **lattice** is a *finite-state automaton* that captures a range of possible different interpretations of user input. For example, a lattice from a speech recognizer will represent different possible speech recognition results for a particular input. Each path through the lattice from a designated start state to one of more end states is a different potential recognition of the input.

**Linguistic scalability/portability.** Implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet.

**Linkage attacks** are methods of *de-anonymization* that operates on *anonymized* microdata by comparing the available information in a record for a person to records in *auxiliary information* looking for a match.

**Low-level control** refers to the direct feedback control of movements of the motors in the joints of a multi-joint robot using sensor information coming directly from the individual motors to reach a specified position in 3D space.

**Markup /Markup language.** An approach for annotating text in which the annotations are merged within the text (in a way that they are syntactically distinguishable from the text itself); one of the most common markup languages is HTML which is used to describe web pages and (by embedding programming code) as the basis to develop web-based applications.

**Mental models** are a pervasive property of humans. People form internal, mental models of themselves and of the things and people with whom they interact. These models provide predictive and explanatory power for understanding the interaction. Mental models evolve naturally through interaction with the world and with the particular system under consideration. [Norman 1986]

**Medical cyber-physical systems** are real-time, networked medical device systems to improve safety and efficiency in healthcare. The specific advantage of the concepts of cyber-physical systems (CPS) involves the use of both real-time sensor devices (e.g., monitoring devices such as bedside monitors) and real-time actuation devices (such as infusion pumps). In this way, MCPS collect information from the monitoring sensors and actuators by, for example, adjusting the setting of actuation devices, firing an alarm, or providing decision support to caregivers. See [MedicalCPS \[2018\]](#) for intelligent user interface projects that fall into this category.

**Medical decision support** systems are guidance services that predict a patient's health status to influence health choices by clinicians. Other functions can be administrative, but we focus on supporting clinical diagnosis and treatment plan processes by for example proposing medical substances with little adverse effects. Future implementations should be integrated into the clinical workflow, provide decision support such as treatment options at the time and location of care as a MCPS rather than prior to or after the patient encounter, and provide recommendations for care, not just assessments.

**Meta-Dialogue Management** deals with those aspects not related to a particular domain or dialogue application, typically particular types of communicative functions.

**mHealth** includes the use of mobile devices in collecting aggregate and patient-level health data.

**Microphone array** is a technology where several microphones are used, which allows for *sound source localization* (determine where the sound is coming from) and *beamforming* (focusing the signal on sound from a certain direction).

**Mimicry or mirroring** “is behavior displayed by an interaction participant who does what another person does and refers to an automatic tendency to imitate others” ([Van Baaren et al. \[2003\]](#) in [Glas and Pelachaud \[2015\]](#)).

**Modality Component (MC)**. A component in the W3C Multimodal Architecture that processes a certain type of input, for example, a speech recognizer, handwriting recognizer, or natural language understanding system.

**Modality Fission** is a technique where a single semantic content is spread over multiple (complementary or alternative) output channels.

**Modality Fusion** describes the process of resolving the semantic intent of a dialogue act by combining different input modalities.

**Modality**. A modality identifies a perceptual mode of a human or a comparable concept simulated by a computer.

**Modalities of interaction**. Dialogues can take place in a richly coordinated multimodal context involving voice, gesture, vision, touch, etc., which is typical of human-robot or human-avatar interaction. At the other end of the spectrum, conversations can take place in unimodal settings, including so-called typed “chat” interactions, or can be supplemented with graphical user interface elements (e.g., buttons for selecting options).

**Mona Lisa effect** is the phenomenon that when a 3D object (such as a face) is projected on a 2D display, all observers in the room will perceive the object to have the same rotation relative to them, no matter where they are located.

**Multimodal**. A human-computer interface is multimodal if the computer supports two or more perceptual modes (1) to decode information relevant to the interaction and (2) to encode its response.

**Multimodal Architecture (MMI)**. A software architecture for a multimodal system, for example, the W3C Multimodal Architecture, the Open Agent Architecture, or DARPA Communicator.

**Multimodal fission**. The process of splitting a generic meaning into two or more modalities for presentation to a user.

**Multimodal fusion or multimodal integration** is the process of combining content from different modes of input in order to determine their combined meaning.

**Multimodal grammar.** The most common use of the term *grammar* is a formal mechanism for accounting the possible combinations of words. A multimodal grammar in contrast is a formal mechanism that accounts for the possible combinations of words and symbols from other modalities such as gesture.

A **multimodal interaction manager** is a system component in an interactive multimodal system that manages the flow of interaction (performs *multimodal interaction management*). A multimodal interaction manager may operate at the turn level, within the turn (*incremental interaction management*), or both. For example, a multimodal interaction manager may be responsible for application of temporal constraints for consideration of speech and gesture inputs as part of a single turn. An incremental interaction manager may trigger incremental feedback to the user while they are still speaking, such as highlighting an object that has been mentioned by voice or pointed at using hand gestures.

**Multiparty interaction:** an interaction that may involve more than two participants. Situated interaction systems deployed in the open world need to be designed to handle multiparty interactions, as people may often arrive and interact with the system in groups.

**Muscle memory** is a form of procedural memory that involves consolidating a specific motor task into memory through repetition. When a movement is repeated over time, a long-term muscle memory is created for that task, eventually allowing it to be performed without conscious effort.

**Myoelectricity.** The electrical signals that stimulate muscles.

An **N-best list** is a list of length  $N$  of different alternative hypotheses produced by a component of a system such as a speech recognizer, gesture recognizer, or natural language understanding component. Each member of an N-best list represents a different possible result. Generally, they are associated with scores or probabilities assigned by the component generating the result, and the results appear in order with the highest score or highest probability result first. For example, an N-best list from the speech recognizer would represent different possible decodings of the speech signal into strings of words.

**Natural interaction** is a more intuitive way of interacting with a computing device. Recent advancements in HCI have facilitated this kind of interaction and its development is expected to make it easy for users to learn how to use the interface in the quickest possible way. [Villaroman et al. \[2011\]](#)

**Neural machine translation:** greater abstraction and greater ease of integration is obtainable through neural translation approaches, where internal (“hidden”) abstractions are generated as a side-effect of training many layers of neural structures.

**Neurophysiological measures** represents the act of measurement based on physiological activity from cerebral sites in the human brain. These measures may be based upon electrical activity (electrocortical, electroencephalographical) or neurovascular changes (functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS)).

**Noise suppression** is a signal processing method which reduces unwanted sound (such as traffic noise) from the signal.

A **nomadic device** is a device for information including entertainment, and/or communication that can be used outside of the vehicle and inside the vehicle while driving. It is not supplied or installed by the vehicle manufacturer. [Kulmala and Mäuerer \[2005\]](#)

**Non-verbal** communication is mediated by other signals than words. In spoken language, this refers to aspects such as prosody, breathing, and laughter. In face-to-face interaction, non-verbal signals are also conveyed in the visual channel, such as facial expressions and gaze.

**Out-of-vocabulary words (OOVs):** when words are missing in the pronunciation dictionary of a recognizer, leading to one or more substitution errors. Named entities and specialty terms are particularly prone to this type of problem.

**Overlay** of feature structures is a type of [unification](#) operation similar to [destructive unification](#). In overlay and destructive unification, if there are clashing values as the two feature structures A, B are combined the values in B appear in the result.

**Overt measure** is a measure of human behavior or performance that can be detected based upon human perception, e.g., voice commands, gestures.

**Participation status:** characterizes the alignment between people and a particular interaction. At the high level, people in a scene can be divided into participants (those who are involved in the interaction) and non-participants (those who are not). Participants can be further divided into speaker (the producer of an utterance), addressees (the people being addressed), and side-participants (the people that are not being addressed by the current utterance but are still ratified participants). Non-participants include bystanders (people nearby that hear the utterance but do not participate in the interaction and are known to the

speaker), and eavesdroppers (other people that are listening, but are not known to the speaker) (see [Goffman \[1979, 1981, Clark \[1996\]\]](#)).

**Passive BCI** is a brain-computer interface that derives its outputs from arbitrary brain activity without the purpose of voluntary control.

**Passive state** is a spontaneous psychological state that arises during behavior without an intention on the part of the person, e.g., fatigue, frustration.

**Persuasive technologies** focus on the design, development, and evaluation of interactive technologies aimed at changing users' attitudes or behaviors through persuasion, but not through coercion or deception. In general, persuasive technologies are used to change people's behavior. The persuasion approach we support is that choices are not blocked, fenced off, or significantly burdened. The influence on people's behavior in order to make their lives longer, healthier, and better should be subtle. For example, displaying nutrition information at eye-level is a subtle persuasion technology.

**Pervasive distributed computing** promises a computing infrastructure that seamlessly and ubiquitously aids users in accomplishing their tasks and that renders the actual computing devices and technology largely invisible. The basic idea behind pervasive computing is to deploy a wide variety of smart devices throughout our working and living spaces. These devices coordinate with each other to provide users with universal and immediate access to information and support users in completing their tasks. [Grimm et al. \[2001\]](#)

**Physical context:** information contained in the physical environment that is relevant to and helps determine how the conversation progresses. Examples include information about people, such as how many people are around, their location and body pose, head and hand gestures, eye gaze and attention; information about the presence, location, and affordances provided by task-relevant objects; the overall topology and structure of spaces, etc.

**Physical ergonomics** is a subfield of ergonomics which considers human anatomy, physiology, and biomechanics in relation to physical activity.

**Physiological computing** refers to a field of research in human-computer interaction wherein **Physiological measures** derived from the human user are used as a source of input control for a computer system or interface.

**Physiological measures** describes the act of measurement based on processes related to human physiological functions.

**Proxemics:** a term coined by [Hall \[1966\]](#) denoting “the interrelated observations and theories of man's use of space as a specialized elaboration of culture.”

Proxemic (spatial) information is an important ingredient in reasoning about conversational engagement, participation status, and other communicative processes in physically situated interaction.

**Psychophysiological measures** describes the act of measurement and inference wherein psychological processes and concepts are inferred on the basis of physiological measurements from the autonomic nervous system.

**Prevention (primary, secondary, and tertiary)** covers several prevention methods. Primary prevention aims to prevent disease or injury before it occurs and includes education about health risk factors. Secondary prevention aims to reduce the impact of a disease or injury that has already occurred and addresses an existing disease prior to the appearance of symptoms. Examples include regular exams and screening tests to detect disease in its earliest stages (e.g., mammograms to detect breast cancer) or diet programs to prevent further heart attacks. Tertiary prevention aims to soften the impact of an ongoing illness or injury that has lasting effects. Examples are cardiac or stroke rehabilitation programs and chronic disease management programs (e.g., diabetes). New approaches to improve prevention-related user interaction include *persuasive technologies*.

**Prosody** refers to the elements of speech that are not the individual phonetic segments (that make up syllables and words). The three main prosodic elements are intonation (fundamental frequency), energy (loudness), and duration. Prosody may reflect many different things, including the emotional state of the speaker, the form of the utterance (statement or question), emphasis, contrast, and focus.

**Quantified self** is a term used to describe data acquisition on aspects of a person's daily life, e.g., incorporating self-monitoring and self-sensing, which combines wearable *biosignals* sensors and wearable computing.

**Quasi-identifiers** are attributes that can be used, often in combination, to pick out a single person from a data set. They can enable *linkage attacks*.

**Rapport** is a feeling of mutual attentiveness, positivity and connection with another [Zhao et al. 2014, Huang et al. 2011].

**Reactive BCI** is a brain-computer interface that derives its outputs from brain activity arising in reaction to external stimulation, e.g., a visual stimulus or sound.

**Redundancy** is present when two modalities have the same expressive power, but are both required to be used within a temporal window in order to reach a goal. Redundancy can be important for safety relevant functions.

**A Recursive Transition Network** (RTN) is a graphical construct that can be used to capture the legal strings of a language. An RTN consists of a series of nodes, transitions between those nodes, a defined start state and a defined set of end states. The transitions are labeled with symbols from the language and paths through the network from the start state to an end state enumerate the legal strings of the language. Critically in an RTN, the label on a transition may also be a reference to another RTN, enabling re-use of sub-networks and recursive description of the language. RTNs are generally implemented as pushdown automata and have equivalent computational power to context-free grammars.

**Reflexive adaptation** refers to a second-order process of adaptation whereby the computer makes an autonomous response and subsequently monitors the response of the human user to that response in order to inform future responses.

**Re-identification** is another term for *de-anonymization*, but emphasizes that the data was merely *de-identified* and never really *anonymized* in the first place.

**Repetitive strain injury** (RSI) is an “injury to the musculoskeletal and nervous systems that may be caused by repetitive tasks, forceful exertions, vibrations, mechanical compression, or sustained or awkward positions”: definition taken from [van Tulder et al. \[2007\]](#).

**Representational State Transfer** (REST). An architecture based on the HTTP protocol commonly used in web-based applications.

The **Resource Description Framework** (RDF) is a family of World Wide Web Consortium (W3C) specifications for metadata. It is used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats.

**Safety by design** refers to the fact that next-generation technical systems for human-robot cooperation will include, e.g., a compliant element in their actuators that absorbs energy. Thereby safety is an integral part of the mechanical construction of the system.

**Secondary uses** of data are uses for purposes other than the one for which the data was collected, which can violate privacy [[Solove 2006](#)].

**Simultaneous interpretation** attempts to recognize and translate spoken language in parallel to the input speech without making the speaker pause.

**Situated (spoken language) interaction:** an area of research investigating the development of computer systems that can reason about their surroundings and interact (via spoken language) in a more natural manner in open-world, physically situated settings. Such systems generally integrate information from speech and vision, reason about both verbal and non-verbal behaviors of potentially multiple participants, and leverage both the dialog and *physical context* when making decisions.

**Situation modeling** is the modeling of the current physical situation: which speakers are involved and which roles they have, as well as any objects or spatial elements that might be of interest.

**Speech synthesis:** text in the target language is output in spoken language (text to speech).

**Speech Synthesis Markup Language (SSML).** An XML markup language for defining how a text should be pronounced, standardized by the W3C Voice Browser Working Group.

**Speech translation goggles** translated output from a simultaneous (lecture) translation system delivered textually via heads-up display goggles.

A **spoken dialog system** (SDS) converses with the user via voice. It consists of a number of components to function successfully: Speech recognition to recognize the words a user said, natural language understanding to assign meaning to these words, a dialog manager to decide how the utterance fits into the dialog so far and decide which actions to perform next, external communication to access external data, a response generation to choose the words and phrases to be used in a response and a speech synthesizer to actually speak the response. [McTear \[2004\]](#)

**Statecharts** first presented by [Harel \[1987\]](#), extend the concept of state machines and state diagrams with concepts supporting hierarchy, concurrency, and communication, enabling the description of complex event-driven real-time applications systems including the operation of devices, human computer interfaces, and communication protocols. Examples of applications include the SCXML standard, developed by W3C (see also Chapter 1 and Chapter 9), as well as the IrisTK framework [[Skantze and Moubayed 2012a](#)].

**State Chart XML (SCXML).** An XML language based on Harel State Charts, used for describing reactive processes, standardized by the W3C Voice Browser Working Group.

**Static optimization** is a method to resolve the joint moments computed by inverse dynamics to forces and activations of individual muscles.

**Synchronization** can be defined as “the dynamic and reciprocal adaptation of the temporal structure of behaviours between interactive partners” [Delaherche et al. 2012].

**Statistical machine translation:** greater speed of learning and better performance and generalization to broader topics, but still requires collections of large parallel corpora.

**Statistical disclosure prevention** attempts to prevent all disclosures about individuals from releasing data. In general, this goal is impossible; a motivation for *Differential privacy*. “Statistical disclosure limitation/control” sometimes is used to mean limited attempts to prevent statistical disclosures [Dalenius 1977].

**Strokes** are the individual components of a gesture made by a user. For example, in order to make an “X” gesture the user might make two strokes one from the top left to the bottom right, then raise the finger or stylus and make a second stroke from the top right to the bottom left.

A **subcategorization frame** is a construct, generally a list, which captures a series of phrases that a word or phrase needs to combine with. For example, the verb “give” in English subcategorizes for a noun phrase (NP) and a prepositional phrase (PP), e.g., “give [the book]NP [to john]PP.” In the grammatical theory Head-driven Phrase Structure Grammar [Pollard and Sag 1994], a subcategorization frame is a list structure represented as feature structure which captures the words or phrases that a word or phrase needs to combine with. In Johnston [1998], the concept of a subcategorization frame is extended to multimodal inputs and a spoken phrase such as “put that there” is said to “subcategorize” for two gestures, the first corresponding to “that” and the second corresponding to “there” in the spoken input.

**Sufficiency of Sensing.** Rather than fusing multiple sensors to determine the most precise estimate of a track’s properties (e.g., range, bearing, identity, etc.), modern aeronautical information fusion algorithms may attempt to derive an estimate sufficient only for the pilot’s understanding or mission success. Once that is accomplished, sensing resources can be redeployed to analyse other tracks.

**Synchronized Multimedia Integration Language (SMIL).** An XML-based language for describing interactive multimedia presentations.

**System initiative, User initiative, Mixed initiative:** Terms that refer to the nature of decisions about action or initiative in human-computer interaction. With system initiative, the computer system takes primary control of the flow of the activity, including conversation or other kinds of collaborations. With user initiative, the human takes primary control of the flow of the activity. In mixed-initiative systems, the computer and human can each take primary control of one or more steps of the activity. Initiative in a mixed-initiative system can be guided by simple fixed-policies, heuristic procedures, or probabilistic and decision-theoretic inference (see [Horvitz \[1999\]](#)).

**Tag.** An XML annotation that defines metadata for the content that it surrounds; for example, “<sentence>I would like to go from Philadelphia to San Francisco</sentence>” expresses the fact that “I would like to go from Philadelphia to San Francisco” is a sentence.

**Tangible user interfaces** are concerned with providing tangible representations to digital information and controls, allowing users to quite literally grasp data with their hands. Serving as direct, tangible representations of digital information, these augmented physical objects often function as both input and output devices providing users with parallel feedback loops: physical, passive haptic feedback that informs users that a certain physical manipulation is complete; and digital, visual or auditory feedback that informs users of the computational interpretation of their action. [Shaer and Hornecker \[2010\]](#)

**Targeted audio:** synthetic speech output in a speech translation system delivered selectively by directional loudspeakers.

**Task-oriented spoken dialog system:** a computer system that allows the user to perform a certain task, like booking a flight or checking the weather, via spoken language input and output.

**Telemedicine** subsumes physical and psychological diagnosis and treatments at a distance, including telemonitoring of patient functions.

**Temporal cascaded approach** is an approach of using multimodal data in a timely sequenced fashion where the usage and outcome of analysis of one data type influences the analysis or choice of a second- or higher-order data type.

**Text-based dialog system:** a computer system that interacts with a user via natural language text input and output.

**Text-to-speech synthesis (TTS):** TTS makes translated sentences audible in the target language and thus permits full speech-to-speech dialogues between two participants.

**Track.** In aeronautical parlance, an entity that is being sensed (e.g., by radar).

**Tracking, smoothing, and forecasting:** Terms that refer to different hidden state estimation problems in a dynamical system. Let us assume a system which has an unknown but evolving state over time. The tracking problem (also sometimes referred to as filtering) amounts to estimating the state of the system, at the current time  $t$ , based on observations made up to the current time  $t$ , i.e.,  $P(s_t|o_1, o_2, \dots, o_t)$ . Smoothing refers to estimating the state of the system, at some past time  $p < t$ , based on observations made up to the current time  $t$ , i.e.,  $P(s_p|o_1, o_2, \dots, o_t)$ . Finally, forecasting or prediction refer to the task of estimating the state of the system, at some future time  $f > t$ , based on observations made up to the current time  $t$ , i.e.  $P(s_f|o_1, o_2, \dots, o_t)$ .

**Transcode.** The transformation of one encoding of an entity into another.

**Turn management functions**, such as requesting or assigning a turn, are a family of communicative functions used in multi-party dialogue to mediate between agents.

**Turn-taking** is the process by which speakers take turns (speaking and listening), i.e., sequencing utterances and managing the floor.

**Typed feature structures** are a feature structure representation where each feature structure is assigned a type within a hierarchical structure. When feature structures are unified, the types are required to be in an ancestor relation in the hierarchy and the result is the more specific type. See [Carpenter \[1992\]](#) for more details.

**Unification** of feature structures is an operation which takes two feature structures as input and determines whether or not they are compatible with each other, that is they unify, and if they do returns a feature structure containing the combination of information from the two input feature structures. This operation corresponds to graph unification of the two underlying directed acyclic graphs that the feature structures represent. Feature structure unification is a generalization of term unification, an operation used in logic programming.

**User-initiative Systems** are typically command and control systems.

**Variational autoencoders (VAEs)** are a variational implementation of [autoencoders \(AEs\)](#). They can be trained via stochastic gradient descent.

Verbal communication is mediated by words. For example, a written chat is purely verbal (with the possible exception of emoticons and the like). Spoken language has a verbal and a non-verbal component.

**Voice activity detection (VAD)** is the binary detection of speech activity vs. silence or noise.

**VoiceXML**. An XML markup language for defining form-filling spoken dialogue applications, standardized by the W3C Voice Browser Working Group.

**Wait-vs.-act tradeoff**: denotes a tradeoff that often arises in systems that need to make decisions under uncertainty and latency constraints. In general, by choosing to wait rather than act, a system might accumulate more evidence about important state variables, reduce uncertainty, and therefore make a better decision about which action to take. However, as timing is also important, waiting might also lead to missed opportunities, or to increased costs of action, e.g., due to inappropriateness of delay.

In **Wizard-of-Oz** studies subjects are told that they are interacting with a computer system, although in fact they are not. Instead, the interaction is mediated by a human operator, the wizard, with the consequence that the subject can be given more freedom of expression, or be constrained in more systematic ways, than is the case for existing implementations.

**World Wide Web Consortium (W3C)**. An international organization whose mission is to support interoperability, security and accessibility of web pages by defining standard languages.

**eXtensible Markup Language (XML)**. A markup language standardized by the World Wide Web Consortium and used in numerous applications.

**XML container element**. An XML annotation which surrounds other content with begin and end tokens, for example in “<author>William Shakespeare</author>”, “author” is the container element.

**XML child content**. Material included in an XML container element. “William Shakespeare” in the XML container element example.

# **The Handbook of Multimodal-Multisensor Interfaces, Vol. 3**

## *Language Processing, Software, Commercialization, and Emerging Directions*

Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, Gerasimos Potamianos, Antonio Krüger

The Handbook of Multimodal-Multisensor Interfaces provides the first authoritative resource on what has become the dominant paradigm for new computer interfaces—user input involving new media (speech, multi-touch, hand and body gestures, facial expressions, writing) embedded in multimodal-multisensor interfaces.

This three-volume handbook is written by international experts and pioneers in the field. It provides a textbook, reference, and technology roadmap for professionals working in this and related areas.

This third volume focuses on state-of-the-art multimodal language and dialogue processing, including semantic integration of modalities. The development of increasingly expressive embodied agents and robots has become an active test-bed for coordinating multimodal dialogue input and output, including processing of language and nonverbal communication. In addition, major application areas are featured for commercializing multimodal-multisensor systems, including automotive, robotic, manufacturing, machine translation, banking, communications, and others. These systems rely heavily on software tools, data resources, and international standards to facilitate their development. For insights into the future, emerging multimodal-multisensor technology trends are highlighted for medicine, robotics, interaction with smart spaces, and similar topics. Finally, this volume discusses the societal impact of more widespread adoption of these systems, such as privacy risks and how to mitigate them. The handbook chapters provide a number of walk-through examples of system design and processing, information on practical resources for developing and evaluating new systems, and terminology and tutorial support for mastering this emerging field. In the final section of this volume, experts exchange views on a timely and controversial challenge topic, and how they believe multimodal-multisensor interfaces need to be equipped to most effectively advance human performance during the next decade.

### **ABOUT ACM BOOKS**



ACM Books is a new series of high quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

ISBN 978-1-970001-72-3

9 0 0 0 0



9 781970 001723