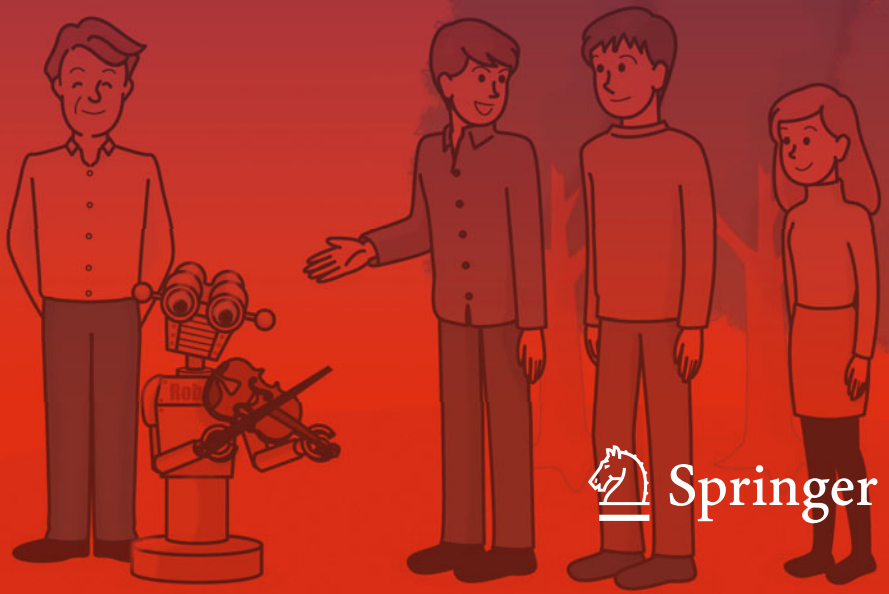
Toyoaki Nishida
Atsushi Nakazawa
Yoshimasa Ohmoto
Yasser Mohammad

Conversational Informatics

A Data-Intensive Approach with Emphasis on Nonverbal Communication



Springer

Conversational Informatics

Toyoaki Nishida · Atsushi Nakazawa
Yoshimasa Ohmoto · Yasser Mohammad

Conversational Informatics

A Data-Intensive Approach with Emphasis on Nonverbal Communication

Springer

Toyoaki Nishida
Atsushi Nakazawa
Yoshimasa Ohmoto
Yasser Mohammad
Graduate School of Informatics
Kyoto University
Kyoto
Japan

ISBN 978-4-431-55039-6 ISBN 978-4-431-55040-2 (eBook)
DOI 10.1007/978-4-431-55040-2

Library of Congress Control Number: 2014942529

Springer Tokyo Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

This important volume brings together perspectives from multiple disciplines to provide an integrated view of an exciting new paradigm that places conversation at the center of an intelligent system. For sure, this text book will become an inspiring resource for students, teachers, and researchers that wish to go beyond traditional dialog systems and investigate the power and potential of social signals in human–computer interaction.

Prof. Dr. Elisabeth André
Professor of Computer Science
Augsburg University

This book offers a comprehensive account of conversational informatics, spanning from its origins to state-of-the-art applications. The data-intensive approach taken makes it a marvelous resource for novice and expert readers interested in the details of designing technical artifacts capable of interacting with humans in a conversational fashion. The valuable connection of the disciplines and perspectives of conversation research and, not least, its nice illustrations make the book enjoyable to read.

Ipke Wachsmuth
Professor of Artificial Intelligence
Bielefeld University

This is a unique book that uses conversation informatics both as a source of inspiration and a rich resource for research in the multifaceted area of emphatic conversation. Recognising the pivotal role of conversation in all walks of life, this book extends beyond the state-of-the art to present long-term vision centered around the magic of conversation. In doing so, this work brings together enabling technologies, including signal processing, artificial intelligence, and computer graphics, the combination of which will ultimately provide harmonious and natural computer-aided assistive devices. The reader will learn how to use signal processing to detect subtle social signals in everyday conversation, to incorporate experience using artificial intelligence, and the role of computer graphics in user-friendly interaction with a synthetic character. This comprehensive big picture

revolves around AI, and includes even the ethical issues for the years to come, such as human-artifact symbiosis. This work is the latest and mature addition to the series of related books organized by Prof. Nishida, and is strongly recommended to researchers active in the field.

Professor Danilo P. Mandic
Imperial College
London

This book opened a new perspective on AI for me. Its authors have a clear and vitally important mission. They report on 10 years of rich research on communicative intelligence. This means the aim is not for AI to outperform humans, but to join forces with them in augmenting conversation. Conversation is meant here in its full richness as a group activity that sustains social ties. In the authors' vision, a "primordial conversation soup" allows humans and agents to meet in conversation, and the conversation allows the soup to evolve. Empathy fiction replaces domination fiction? Only this is not just fiction. The book contains 12 chapters, many of which are filled with detailed description of actual immersive conversation environments.

Besides descriptions of practice, and perhaps even more importantly, the book articulates a clear vision on what the aims of artificial intelligence should be in aiding conversation, in order to serve the needs of society. This vision is informed by a solid grasp of historic developments in AI.

Decidedly, the book shows the way forward for artificial sociality, which is where AI should go. If you like science fiction to be realistic, you should read it.

Dr. Ir. Gert Jan Hofstede
Associate Professor
Social Sciences Group
Wageningen University

Driven by the inspiring vision of a society cohabited by humans and conversational artificial agents, this highly innovative work puts forward an overarching, ambitious, and original research program toward the development of both virtual and robotic conversational agents.

The authors introduce and draw the first comprehensive map of conversational informatics, vividly demonstrating that conversational informatics affords a promising common framework for scientists working in natural language processing and multimodal communication, affective computing, computer vision, machine learning, virtual reality, robotics and situational awareness in multi-agent systems.

At the same time, the authors present a novel approach to the study of conversational interactions, which promises to yield a rich harvest of near-term advances in conversational agent technologies, models, and applications. Most notably, they develop a new data-intensive methodology for the modeling and

building of conversational systems and environments, while vindicating conclusively the central role of unduly neglected forms of nonverbal communication.

Add to this insightful forays into highly relevant neuroscientific, psychological and philosophical work—jointly with in-depth analyses of challenging conceptual, methodological, and even ethical issues—and we are presented with an outstanding work of importance to a wide community of researchers and students, from computer science and robotics to the social sciences and the humanities, who share an interest in understanding or engineering conversational interactions and systems.

Guglielmo Tamburrini
Professor
Philosophy of Science and Technology
Universita' di Napoli Federico II

"What would happen if mankind would have to live without conversation?" starts the journey into the new interdisciplinary field of conversational informatics. The novice reader quickly realizes the importance and complexity of conversation—designated as the air in our intellectual life—reaching far beyond social interaction for information and knowledge exchange, toward the multimodal sharing of emotions, thoughts and wisdom. The purpose of the book is to provide a first reference on conversational informatics, a new paradigm of artificial intelligence for exploiting conversation as a primary principle of building intelligent systems, integrating the authors achievements in definition, knowledge, expertise and research platform.

The book provides a comprehensive approach to the study of conversational informatics, covering the three angles from which conversations can be observed: the transactional angle to capture and manage utterance, narrative and story content; the interactional angle to search for underlying principles and a computational model for the exchange of social signals in conversational interactions; and the cognitive angle to reveal the mental processes and their interaction to form a sophisticated intelligent process. The angles are integrated in an interdisciplinary, data-driven approach to exploit an artificial "primordial soup of conversation," facilitated through an intelligent augmented conversation environment. This allows co-evolution of common ground and communicative intelligence through the accumulation of conversation, focusing on key features such as interactive content, intelligent sensors and objects, and the creation of conversational agents and robots, which turns out the most challenging and ultimate goal in this endeavor.

Through a fascinating and capturing storytelling, the book unifies the developmental paths of artificial intelligence, information technology and society: In the history of AI, the initial success of closed systems achieving impressive results under limited resources is followed by a "winter" period, where astonishment decays and challenges turn out more profound. The evident exponential growth in information content and computation performance recently allows a renaissance

through reformulation as statistical computing. However, this information and technology explosion also leads to novel challenges for computing and society. In this setting, the authors ground the seemingly natural and consequent concept of conversational informatics. On the long-term horizon, the idea to deal with this inevitable singularity is an agent-mediated society, where empathic surrogates act as our proxies in a complex information and technology infrastructure, and help mankind to maintain and improve empathy among people. Apart from this futuristic vision, the book focuses on actual achievements and tackles the question about the lowest hanging fruits. The promising next step may be a minimal system to launch a primordial soup of conversation, covering a broad range of aspects comprising theory, platform, measurement and analysis, model building, content production, application, and evaluation.

The book well captures the interdisciplinary nature of the topic, lines up historic development for a prediction of future challenges, and proposes the concept for this emerging field by covering a variety of aspects. The content ranges from an easily accessible introduction, an overview of scientific aspects of conversation, and a historical and technical survey; through a detailed theoretical framework for collecting and reusing conversational content, and a technical framework for sensing, measurement, analysis, modeling, machine learning, and reproduction of conversation, communicative behavior and underlying cognitive and intelligent processes; toward high-level issues such as social intelligence design, ethics and empathy. Thus, unlike the title suggests, the book aims at a broad interdisciplinary audience reaching beyond artificial intelligence and computer science. It targets the novice as well as the expert and allows for casual as well as focused reading.

Christian Nitschke
Assistant Professor
Graduate School of Informatics
Kyoto University

Preface

One often starts something new solely with a strong belief. Such optimism may bring about unexpected outcomes. On the one hand, one might come across a magnificent landscape consisting of ridges, valleys, and, if the initial intuition is right, a stunning hidden falls or lake, which could not have been visible at the onset. On the other hand, this process might compel us to spend far more energy and cost than our initial estimates, as the goal becomes far more profound than our initial guess. In general, benefit surpasses loss so long as one proceeds in the right direction. The greater and more complex the terrain is, the deeper and more exquisite surprises we experience on this path of adventure. Addressing difficult problems brings about an awareness of a huge accumulation of knowledge and a good community of friends approaching the same goals from different angles.

When we launched our project on conversational informatics more than a decade ago, we did not envision ourselves as reckless artificial intelligence researchers aiming at making computers invisible and sociable without foreseeing the depth and influence that inquiry into conversation would bring about. After much effort and time, the vast landscape gradually showed up out of the mist. We have hence been able to sustain shaping our vision, knowledge, expertise, and research platform by both absorbing abundant nutrition from the fertile soil of accumulated literature and becoming allied with powerful friends who shared our vision.

The purpose of this book is to share what we have built, conversational informatics, a new paradigm of artificial intelligence for exploiting conversation as a primary principle for building intelligent systems. Conversation is a traditional wisdom of mankind for understanding and communicating experiences and knowledge. Endowing intelligent systems the ability to converse with people enables such artificially intelligent systems to coexist as our partners.

In general, there are three angles from which we can observe conversations. First, the transactional angle involves how conversation manifests as utterances are dictated. Each verbatim utterance used in conversational interactions is deemed a main constituent of conversation, which in turn may be captured in a larger discourse of narratives and stories. In this vein, conversation can be seen as an opportunity for shaping and sharing stories by participants, likely with new interpretations and criticisms presented along the way.

Second, the interactional angle focuses on social signals. As pointed out by Herbert H. Clark, conversation is social interaction in nature. Conversation not only consists of joint activities among participants but also is conducted as a part of a larger joint activity. Speech is just one component of interaction. A vast amount of untold nonverbal messages are used to express meaning and intention. Participants brilliantly use nonverbal communication to exchange social signals with one another, thus making conversational interactions efficient and engaging. Unveiling the structure of the flux of social signals exchanged in conversational interactions, we search for the underlying principles and a computational model; such a search has been a key challenge that has fascinated the communication sciences for many years.

Third, the cognitive angle focuses on the underlying cognitive processes; to understand conversation, we need to investigate the phenomena beneath the surface. Conversation is a very interesting phenomenon from the viewpoint of cognitive science and artificial intelligence, as conversation results from a complex interaction of numerous mental processes ranging from the sensorimotor level to the deliberation level. A key challenge in cognitive science is to uncover and describe the mental processes involved and how they interact with one another. From the viewpoint of artificial intelligence, the challenge lies in how these numerous intellectual processes combine to comprise an overarching sophisticated intelligent process.

In an artificial intelligence, conversation is not a topic often focused on, while dialog has been much more popular as a research subject. As to why this is the case, we suspect that dialog is more thematic, oriented toward a problem that participants would like to address. Evaluation criteria for dialog systems are more clearly specified in terms of how well the system can reason about the goals and plans of the human user to provide the user a proper answer possibly after a series of clarification dialogs or repairs. Dialog might have been a feasible target for artificial intelligence researchers to integrate intelligent computing, natural language understanding and generation, dialog and discourse planning and management, social problem solving and interactions, perceptual user interfaces, including speech and vision, and affective and cognitive computing.

In contrast, conversation is more a bottom-up process that consists of spontaneous contributions and efforts to sustain the activity. Dialog is not a dyad interaction and conversation may have two or more participants. There does not seem to be an objective criterion to judge the appropriateness of conversation, as it appears to depend on the background knowledge and personal interests of the participants, all of which tend to be subjective in nature. Conversation may be deemed more useful and effective provided that it provokes more thoughts and stimulates more utterances. Intelligence may not be the sole factor of successful conversation. Content is indispensable; without content, conversation is merely a social event to make social decisions or maintain social relationships. In the long history of artificial intelligence research, content has not received much attention.

Interactional aspects of conversation have received much more attention than transactional aspects with a few exceptions. Again, content is rather subjective and less objective, as its value depends on the background of the producers and consumers.

To put it another way, underlying story structure is important in conversation. In contrast, goal-oriented dialogs involve utterances that are used to achieve social goals, such as requesting information or making reservations. In conversation, participants often collaboratively create a joint story. It is critical for participants to find a relevant story and relate what is told to their knowledge. Rather than identifying social activities behind the utterances, the addressee contributes to joint storytelling by asking questions or making comments and learning from the responses.

An interdisciplinary approach is indispensable to integrating the transactional, interactional, and cognitive angles of conversation; doing so supports the pursuit of theory and practice of conversational informatics. Analysis is mandatory to gain a solid understanding of the phenomenon, while synthesis allows for the combining of parts to envision the whole. We take a data-intensive approach. We believe that it incorporates both analysis and synthesis. Data-intensive understanding brings about quantitative understanding, permitting us to turn a great accumulation of keen observations into a pile of computational models. It also greatly helps to build conversational agents by virtue of recent progress in machine learning and data mining, learning by imitation in particular. Rather than merely addressing the construction of conversational agents or conversational content, we aim to build an intelligent smart environment that implements the idea of primordial soup of conversation, in which both conversational agents and conversational content co-evolve.

In the long run, we aim at building an empathic agent that can not only understand the emotion and intention of people but also induce a user's positive feeling of partnership with the agent, and therefore trigger spontaneous activities toward the agent. We base our approach on the sharing hypothesis, which states that the more common ground is shared, the more empathy there is to be gained. Recent advancements in networks and sensing technologies enable us to share with conversational agents the universe of discourse, knowledge and skills, situation, and a first-person view of the world. This in turn suggests that the more data accumulated into conversational agents, the more empathy there is to be achieved.

The idea of a primordial soup of conversation is implemented as a shared conversation space built on the spectrum of cyber-physical space for augmented conversation, ranging from augmented reality to virtual reality. In the augmented conversation space based on augmented reality, people move around to talk with one another in reference to objects and events in the space. Conversely, we use virtual reality to realize an immersive conversation space in which each participant is completely surrounded by large screens, and the behavior of the participant is sensed such that it may be reflected in his or her self-image in the shared conversation space. Our immersive interaction environment allows participants from spatially distributed locations to participate in interactions in the shared

conversation space. Both enable the grounding of a conversation in a given situation; conversational content is significantly more valuable if it is coupled with the situation it refers to.

How will the primordial soup of conversation evolve? In the beginning, conversation participants in the primordial soup may only consist of avatars backed by real people. Avatars may be characterized as mechanized humans (MHs), whose conversational behaviors are somewhat constrained by the interface. Communication behaviors of avatars are recorded and accumulated for reuse or for creating behaviors of conversational agents or human-like machines (HMs), which imitate communication patterns demonstrated by MHs. Gradually, the population of HMs may increase, increasing the size and scope of the primordial soup, from which more content will be generated for further evolution.

Our immersive interaction environment works not only as a telepresence facility that permits users to incarnate as embodied robots that move around and interact with partners in a remote place, but also as a very powerful tool for collecting data in a fashion called immersive Wizard of Oz, which entails how people behave in a situated conversation environment. We provide a theoretical framework for collecting and reusing conversational content, survey a computer vision technique for sensing and reproducing human communicative behaviors, and present a suite of machine learning methods for imitating the communicative behaviors of people.

For a newcomer to explore new horizons in conversational informatics, we provide an introduction to major work on scientific aspects of conversation, conversational analysis in particular, followed by a brief history of conversational system development and an overview of existing technologies for building conversational systems.

Finally, we thank our colleagues and friends who have helped and supported us pursue our adventure. Our thanks go to the following long list of supporters: Elisabeth André, Subhash Bhalla, Aleksandra Cerekovic, Renate Fruchter, Benjamin Alexander Hacker, Gert Jan Hofstede, Hung-Hsuan Huang, Rajiv Khosla, Takuya Izukura, Lakhmi Jain, Kenichi Kanda, Misao Kataoka, Kazumi Kinoshita, Andrey Kiselev, Hidekazu Kubota, Divesh Lala, Danilo P. Mandic, Loic Merckel, Takashi Miyake, Stuart Moran, Shingo Mori, Yukiko Nakano, Keiichi Nakata, Christian Nitschke, Hiroki Ohashi, Shogo Okada, Igor Pandic, Tomasz M. Rutkowski, Hiroyasu Saiga, Kae Sakamoto, Jordi Vallverdú Segura, Hengjie Song, Yasuyuki Sumi, Guglielmo Tamburrini, Nguyen Ngoc-Thanh, Sutasinee Thovuttikul, Kazuhiro Ueda, Ipke Wachsmuth, Thomas Wankerl, Ryota Yanai, Masaharu Yano, Yukari Okuda, Rei Takimoto, Yuko Sumino, friends, and family members.

Kyoto, March 2014

Toyoaki Nishida
Atsushi Nakazawa
Yoshimasa Ohmoto
Yasser Mohammad

Contents

Chapter 1
Artificial Intelligence and Conversational Intelligence

Abstract Conversation is indispensable in our intellectual life. People may make conversation either to achieve a social goal, to create a joint story, or to just enjoy a language game. Although a conversation has a fairly sophisticated structure and dynamism, people are sufficiently proficient in expressing their thoughts and interpreting utterances of their partners. In this chapter, we will introduce conversational informatics as a new interdisciplinary study that focuses on understanding and augmenting conversations. We will discuss why conversational informatics is important in the history ofartificial intelligence research and show our long-term research strategy.

Keywords Conversational informatics · Artificial intelligence · Empathy · Sharing hypothesis · Engagement · Primordial soup of conversation

1.1 Conversation as a Focus of Interdisciplinary Study

Why do we converse with each other? What would happen if mankind would have to live without conversation? We would instantly find that we would run into trouble if all conversation were to be blocked even for a short while. We would realize how deeply we depend on conversation, not just in sharing information and knowledge or conducting social interactions but also in sharing emotions, thoughts, and wisdom beyond those who originally possessed them.

Conversation is like air in our intellectual life. It looks as if we could not live an intellectual life without conversation, just like we cannot live a physical life without air. From time to time, it appears that people simply enjoy building a joint story to share in a community. Conversation not only allows you to articulate tacit thoughts deeply embedded in the mind to transmit to other people but also permits you to examine others' thought. Even in the simple transactions of just selling or buying train tickets, people may need a reason to understand ongoing daily events, regardless

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_1

of whether they are involved. People will feel extremely happy if they have found themselves a part of an underlying fantastic story beneath the surface.

Participating in a conversation is very much like participating in a game. People are excited and thrilled in playing a role to achieve a better score even when the game is not official. People often appear to simply enjoy a genuine language game just for fun even without achieving any practical benefit.

Conversation is a fairly complex business, comprising a fairly sophisticated structure and dynamism. Among others, Goffman (1963, 1981) provided with numerous brilliant observations in many aspects of communication ranging from unfocused to focused gatherings. Austin (1962) and Searle (1969) originated speech act theory by highlighting how people carry out social interactions in conversation. In his comprehensive theory of language use, Clark (1996) characterized language use as a joint activity consisting of layers of representation devices and intervening processes across multiple levels and tracks. Nonverbal communication introduces additional complexity in conversation resulting from polysemy and polymorphism, as summarized by Richmond et al. (2004) among others. Kendon (2004) characterized gestures as a part of speaker's utterances and analyzed how co-herence between gesture and speech is attained to create meaning. Kendon (1967), Sacks et al. (1974), and Duncan and Niederehe (1974) among other authors unveiled the turn-taking system, by shedding a light on how social signals are used to control the flow of conversation. Goodwin (1981) studied the structure of coordination in conversational interaction, focusing on how participants work with each other to change the engagement type. McNeill (2005) provided a conjecture on the psycho-logical process of language and gesture co-generation, based on the detailed investigation of annotated conversation records created using the recording, transcription, and coding methods he and his colleagues invented.

It is surprising that people are sufficiently proficient in expressing their thoughts and interpreting utterances of other partners in conversation. It is noteworthy that people can initiate and maintain conversations with little difficulty in daily life. People are proficient in planning and shaping thoughts as utterances produced at appropriate moments in conversation to tell stories according to other participants? interest and the discourse of the conversation. People are skillful in interpreting verbal and nonverbal expressions in conversation to incorporate other participants? experience into our memory and use them to plan activities in other occasions. To make the gathering both pleasant and fruitful, people collaborate with each other to read the emotion and intentions of other participants, though not always successful.

We aim to be comprehensive in the study of conversational informatics, trying to encompass as much phenomena intrinsic or extrinsic to conversation as possible. We also analyze and model conversational behaviors of people to unveil underlying principles and attempt to build artifacts for augmenting conversations. It is quite challenging to build a robot or software agent that can participate in daily conversation with people. Last but not least, we place the same amount of emphasis on producing and managing conversation content, as it is an integral part of conversation, just as conversational agents and environments are.

The field draws on a foundation provided by artificial intelligence, natural language processing, speech and image processing, cognitive science, and conversation analysis. It sheds light on meaning creation and interpretation during conversations, in search of better methods of computer-mediated communication, human-computer interaction, and support for knowledge creation.

Let us start our journey with thinking about the role of conversational informatics in the forthcoming society that may be characterized by info-plosion and techno-plosion.

1.2 Conversational Informatics in Info-plosion and Techno-plosion

One of the key features that characterize our society in the 21st century is the exponential growth of the amount of information on the globe. This phenomenon is often called information explosion or info-plosion (Kitsuregawa and Nishida 2010). In the info-plosion age, the user is becoming more interested in information services brought about by artifacts, rather than artifacts themselves. Info-plosion not only visualizes a long tail of internal desire that might serve as a driving-force for technology but also popularizes methods and tools for realizing the internal desire. The collective intelligence empowered by the information infrastructure enables agile fabrication of a solution to desire. The more desire is satisfied, the far more new desire is popping up.

Artificial intelligence is an area of research that has contributed to building intelligent programs and robots. Since its official launch as early as in 1956, artificial intelligence research has marked plenty of great successes and is now regarded as one of the most established areas of research, though its scope and approach are not uniform at all, as it may have been seen. The history of artificial intelligence research is rather winding, having gone through a so-called winter era, as shown in Fig. 1.1.[1]

The early days of artificial intelligence may be characterized as a fight against computational poverty, that is, against very limited computational resources available from computers. It compelled researchers to concentrate on hacking a closed-world, symbolic program that could amaze the computer-naive audiences even with very low-performance computers. Heuristic search, weak method, knowledge representation, and symbolic inferences are common technical issues. Most of these approaches were rather conceptual or even ideological, oriented towards a "philosophy of the artificial". It should be noted that researchers were more enthusiastic in aiming at a stand-alone computational intelligence than open-ended interactive systems due to the desire of demonstrating the power of autonomous problem solving without any online assistance from the users. Apart from those hang around the "AI core", there is a group of "non-mainstreamers" who sought more practical value such as

[1] Descriptions about the history of artificial intelligence are available from various sources, such as http://aitopics.org/misc/brief-history. The description in this section is based on Nishidas (2012).

Year	AI	ICT
1940s and earlier		1936: Turing Machine 1947: von Neumann Computer 1948: Information Theory, by C. Shannon and W. Weaver 1948: Cybernetics by N. Wiener
1950s	1952–62: Checker program by A. Samuel 1956: Dartmouth Conference	1957: FORTRAN by J. Backus
1960s	1961: Symbolic Integration program SAINT by J. Slagle 1962: Perceptron by F. Rosenblatt 1966: The ALPAC report against Machine Translation by R. Pierce 1967: Formula Manipulation System Macsyma by J. Moses 1967: Dendral for Mass Spectrum Analysis by E. Feigenbaum	1961: Mathematical theory of Packet Networks by L. Kleinrock 1963: Interactive Computer Graphics by I. Sutherland 1968: Mouse and Bitmap display for oN Line System (NLS) by D.C. Engelbart 1969: ARPA-net
1970s	1971: Natural Language Dialogue System SHRDLU, by T. Winograd 1973: Combinatorial Explosion problem pointed out in The Lighthill report 1974: MYCIN by T. Shortliffe Mid 1970s: Prial Sketch and Visual Perceptron by D. Marr 1976: Automated Mathematician (AM) by D. Lenat 1979: Autonomous Vehicle Stanford Cart by H. Moravec	1970: ALOHAnet 1970: Relational Database Theory by E.F. Codd 1972: Theory of NP-completeness by S. Cook and R. Karp Mid 1970s: Alto Machine by A. Kay and A. Goldberg 1976: Ethernet 1979: Spreadsheet Program Visicalc by D. Bricklin
1980s	1982: Fifth Generation Computer Project 1984: The CYC Project by D. Lenat Mid 1980s: Back-propagation algorithm was widely used 1985: the Cybernetic Artist Aaron by H. Cohen 1986: Subsumption Architecture by R. Brooks 1989: An Autonomous Vehicle ALVINN by D. Pomerleau	1982: TCP/IP Protocol by B. Kahn and V. Cerf Mid 1980s: First Wireless Tag Products 1987: UUNET started the Commercial UUCP Network Connection Service 1988: Internet worm (Morris Worm) 1989: World Wide Web by T. Berners-Lee 1989: The number of hosts on the Internet has exceeded 100,000
1990s	1990: Genetic Programming by J.R. Koza Early1990s: TD-Gammon by G. Tesauro Mid 1990s: Data Mining Technology 1997: DeepBlue defeated the World Chess Champion G. Kasparov 1997: The First Robocup by H. Kitano 1999: Robot pets became commercially available	1992: The number of hosts on the Internet has exceeded 1,000,000 1994: Shopping malls on the Internet 1994: W3C was founded by T. Berners-Lee 1997: Google Search 1998: XML1.0 (eXtensible Markup Language) by W3C 1998: PayPal
2000s	2000: Honda Asimo 2004: The Mars Exploration Rovers (Spirit & Opportunity)	2001: Wikipedia 2003: Skype / iTunes store 2004: Facebook 2005: YouTube / Google Earth 2006: Twitter 2007: Google Street View
2010s	2010: Google Driverless Car / Kinect 2011: IBM Watson Jeopardy defeated two of the greatest champions 2012: Siri	

Fig. 1.1 Brief history of AI research and development contrasted with that of information and communication technology. Adapted from Nishida (2012)

knowledge-based systems or intelligent systems. In spite of a number of monumental successes, such as Dendral, Macsyma, XCON, and ALVINN, they were deemed rather exceptional, and the entire AI business ran into an AI Winter in the late 1980s.

Since the 1980s, however, AI research began to thaw due to the new trend of data-and-computation-intensive approaches enabled by connected super computers connected to the Internet covering the surface of the globe. Machine learning and data mining have become a main stream of AI, starting to attract both theoreticians and practitioners to form a new synergy. People sought ways of reformulating AI problems as statistical computing, rather than manually coding solutions using AI programming techniques.

Roughly by the end of the 20th century, the first round of AI research had concluded in the sense that almost all initial first paths had been explored, ranging from heuristic search to evolutionary computing, incorporated into a comprehensive textbook written by Russell and Norvig. The success of AI so far may be categorized into six groups: large scale search, knowledge-based systems, intelligent media technology including language, speech and vision, planning, machine learning and data mining, and computational art. Numerous challenges, such as, The DARPA Urban Challenge (2007–), The DARPA Robotics Challenge (October 2012–), and Robocup, provided them with opportunities to demonstrate their inventions and fuel for further progress.

Success of such a new paradigm of computational intelligence includes Deep Blue, IBM Watson, Google Voice Search, and Siri, just to name a few. At the same time,

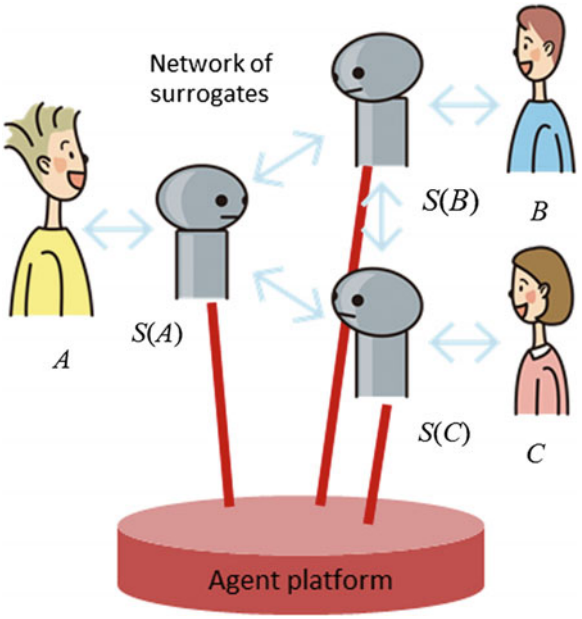


Fig. 1.2 Agent-mediated society. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

progressive efforts in classic paradigm of AI, planning in particular, have brought about a steady success in autonomous/self-driving vehicles such as Mars Rovers, Stanford Cart or Google Self-Driving Car.

Accelerated by artificial intelligence technologies, info-plosion will eventually result in techno-plosion, causing a paradigm shift from the artifacts-as-designed-tools to artifacts-as-autonomous-agents. In this paradigm, we will live with autonomous agents for assistance in the living space. They will be around you and help you both reactively and proactively. They will serve as an integrated interface between the real and cyber worlds. They will be able to collect the situational information around it through sensors and provide a suitable service with the user through various mediators. They will manage your social relations as your surrogate, such as making and maintaining promises, as well. The resulting society cohabited by people and agents may be called the *agent-mediated society* (Fig. 1.2). Inter-human interaction is possible by coupling human-agent communication between the owner and her/his surrogate, and the social communication among mediators. Each person's intention is communicated to her/his surrogate in human-agent communication. Surrogates interact with each other on behalf of the owner. As the technology will implement the details, people will not have to be computer geeks in order to benefit from the advanced information and service environment provided by intelligent autonomous agents. This new framework is deemed as the world supported by "super intelligence" in which each human will not directly interact with each other or play social functions anymore; instead, people will interact with each other through their surrogates.

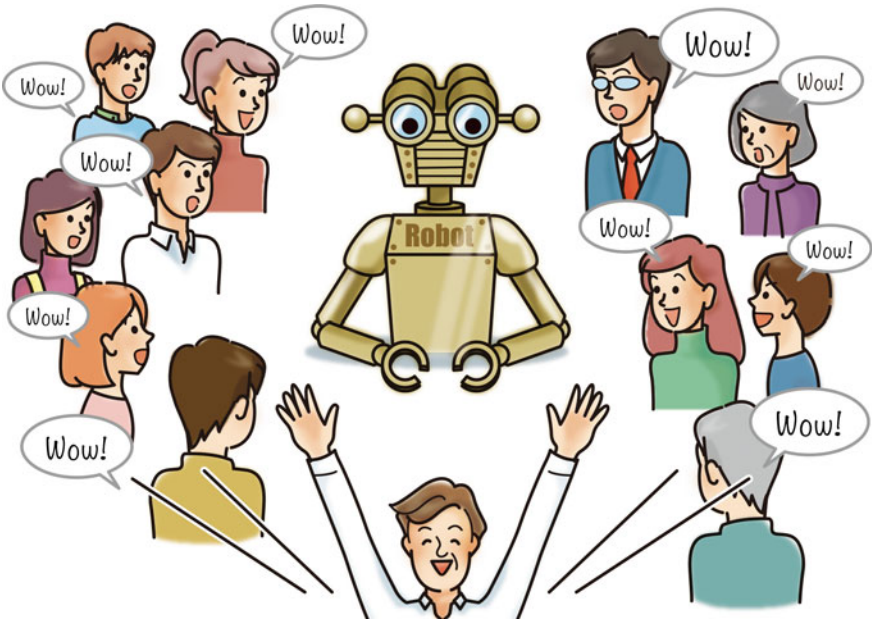


Fig. 1.3 Working for the Wow factor. © 2014, At, Inc. Reproduced with permission

The relationship between humans and artificial intelligence might drastically change as pointed out as a *technological singularity* that occurs when artificial intelligence surpasses human intelligence. However, the influence of techno-plosion has not been discussed seriously, partly because most researchers appear to be naive and optimistic that techno-plosion may not have reached the level of maturity that might deeply impact or even threaten human society. Classic AI has been driven by pioneers with frontier spirits, motivated by a desire to increase a wow-factor by demonstrating pure machine intelligence which can outperform humans (Fig. 1.3). Even though such an attitude may represent researchers' genuine enthusiasm, they are subject to criticism such as, machine intelligence alone will not help much, process of intelligent reasoning is more important than the result and should be made accountable, or researcher's satisfaction may not benefit society.

Nishida (2013) argued that there will be four major problems underlying the naive implementation of artificial intelligence (Fig. 1.4). The first problem is *technology abuse*. New technologies can be applied to illegal or malicious activities. Artifacts that merely act on behalf of the owner extend both the good and evil wills of the owner. This advanced technology might enable an intruder to sneak into your proximity without being noticed, or commit fraud in a novel way. Past experiences and knowledge might not be able to prevent such activities from happening. Serious problems may arise before people become aware of them and take precautions (Nishida and Nishida 2007).

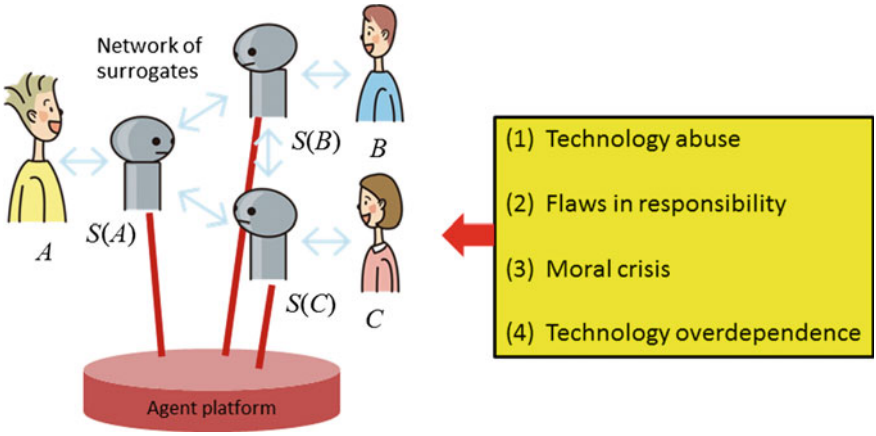


Fig. 1.4 Four major problems in agent-mediated society. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

The second problem is *flaws in responsibility*. Perrow (1984) pointed out the difficulty in sustaining sensibility against accidents as an organization, and accidents might occur even frequently. The more complex artifacts become, the less likely humans can put them under their control. Neither the product maker nor the owner of a complex artifact may take full responsibility of an artifact if it is fairly complex; the owner of the artifact cannot envision all possible outcomes caused by the artifacts s/he possess, for it might be simply beyond her/his intellectual capabilities; a product maker may not be able to ensure that the product may work properly under any usage scenarios.

The third problem is *moral crisis*. In a complex and changing world, ethical conflicts might occur when people have inconsistent beliefs in how ethical principles are instantiated as moral rules, even though people might know that *ethical principles* (Kant 1788; Rawls 1999; Decker 2008) are designed to protect the autonomy and dignity of humans and construe moral rules as entailing the norm of protecting the weak from being used solely as a means to achieve an end by the strong. People might not be fully aware of the consequence of her/his dominance over the weak (Nishida 2009).

The fourth problem is *overdependence on artifacts*. The introduction of an autonomous agent will cause heavy dependence on artifacts. Individuals might use artifacts without considering the consequences. Society might assume the infallibility of artifacts without rationale. There are strong concerns about heavy dependence on artifacts. Among others, Cooley (2007) exhibited a similar concern using "from judgment to calculation", and gave a caveat of being overly dependent on calculation rather than judgment. Heavy dependence will entirely remove motivations of thinking and imagination at the individual level, and might bring about "empty brains" (Maurer 2007). Maurer warns that a serious break down of the computerized social infrastructure might cause a catastrophic disaster and take human society back to the Stone Age.

Some of these problems might not be innate for techno-plosion. Technology abuse and flaws-in-responsibility problems might be resolved by the introduction of a computational framework of mediating social interactions in which *public mediators* are introduced in addition to surrogates and *private mediators* to bring about fair distribution of resources. An important class of mediator is that of public mediators, which may be made available by a (substantially) public organization authorized in the society. Public mediators may attempt to maximize the harmony, if not the merit, of the society as a whole, by maximally arbitrating the requests and offering from participating agents, while private mediators may work for the owner or institution to which they belong, attempting to maximize the satisfaction of the owner's intention, based on a specialized knowledge about how to plug-in to the network of mediators. Public mediators shall be built under a publicly transparent process and operated under public monitoring. Risks caused by public mediators may be underwritten by the public; even if the user may suffer from damage, it is reasonable that the public sector representing the community involving the user takes responsibility to compensate for the damages caused by the mediators at public cost. In contrast, private mediators may be employed for occasions when special conveniences are desired at the user's own cost.

The technology abuse problem will be almost solved if not complete, as people can affect other people only through mediators that comply with the social rules. Their safety is guaranteed by the community so long as they use public mediators, as artifacts are designed to comply with the social rules and hence are transparent at the granularity specified by the social rules. The flaws-in-responsibility problem will be almost, if not completely solved in principle in this framework, in the sense that the infrastructure is underwritten by the community responsible for the user.

It should be noted, however, that even the society of public mediators is inherently incomplete. First, it cannot be free from accidental malfunction. Even worse, the above framework deteriorates the morals in crisis and overdependence on artifacts problems. The morals in crisis problem will escalate, as surrogates will do anything on behalf of the user; as a result, people will lose opportunities to think about morals and practice moral behavior in society. The overdependence on artifacts problem will become desperately serious to the degree of being unrecoverable. It appears that we have to live with being assisted with artifacts (Nishida and Nishida 2007).

Occasionally, mankind might encounter unprecedented disasters and the computational framework might cease to function as an infrastructure. This is the time for mankind to depend on itself. Under such circumstances, a reasonable goal might be to create mutual dependency between empathy and technology: using technology to help people cultivate empathy among people so that empathy in the society may allow people to help each other to restore the infrastructure of civilization, should they suffer from disasters and breakdowns that might be caused by the incompleteness of technology. A substantial portion of technology should be devoted to enhance mankind's self-reliance and resilience to unpredictable breakdown.

Super intelligence will eventually exceed human intelligence, which is referred to as "technical singularity". It appears that people are becoming more concerned with the actual cost and benefit expected from super intelligence. Some people might fight

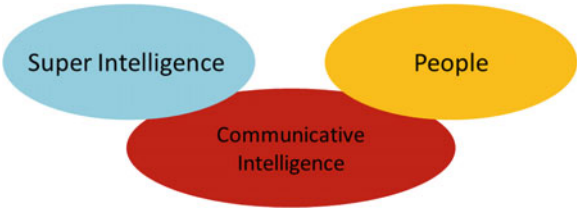


Fig. 1.5 Communicative intelligence for bridging people and super intelligence

against it as a type of neo-luddism. In contrast, a majority of scientists and engineers might simply dismiss the idea as nonsense and keep them involved in inventing a better solution without taking urgent actions.

We take the technological singularity more seriously to consider the issues as rather a lack of common ground in understanding between people and the artificial. We suspect that the singularity may result from the attitude of conventional studies on artificial intelligence that placed too much emphasis on a closed form of computational intelligence rather than communicational aspects of intelligence. We do believe that new lines of research, *communicative intelligence*, are needed which make a sharp contrast to, but complementary to, the traditional lines of research and development in artificial intelligence. The goal of communicative intelligence is to establish a trustable bridge between people and super intelligence (Fig. 1.5).

For such a bridge to be effective, we believe that the key issue is empathy (Nishida 2013). Empathy is defined as "the ability to understand others' emotions and/or perspectives and, often, to resonate with others' emotional states," or as "an affective response that is identical, or very similar, to what the other person is feeling or might be expected to feel given the context: a response stemming from an understanding of another's emotional state or condition." (Eisenberg et al. 2010) Empathy can also be considered to be equivalent to conviviality that allows individuals to identify with each other thereby experiencing each other's feelings, thoughts and attitudes; hence, is deemed a central concept to designing a community (Caire 2009). Empathy is critical for people to understand and help each other to restore the infrastructure should there be a technology breakdown. Intuitively, a condition for an agent to be empathic is to bring about a situation where the user says to the agent "glad to stay with you," as a result of a service (Fig. 1.6).

How can empathy between people and computational intelligence be put into a concrete image? For an agent to be empathic, it should possess: emotional intelligence that allows for behaviors based on sensing the emotion of self and other agents, social intelligence that comprises both the ability of an agent to build a social relationship with others and to use it when solving a variety of problems and the ability of a group to learn from experience when solving problems, and empathic intelligence that permits an agent to act in an empathic manner by sensing/raising awareness of tacit intentions of the user.

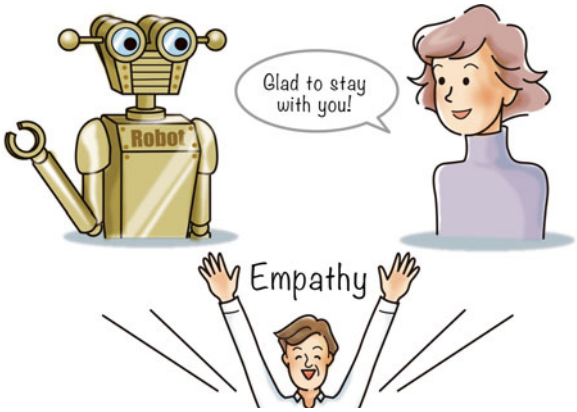


Fig. 1.6 Building empathic agent. © 2014, At, Inc. Reproduced with permission

We base our approach on the *sharing hypothesis*: the more common ground is shared, the more empathy is gained. The sharing hypothesis suggests the direction of our technical development. It should be noted that even today, the Internet and web technologies have brought about a significant impact on helping people share the universe of discourse, first-person view, knowledge and skills, communication style and rituals, and value system.

Since conversation is the most effective and natural means for people to build a common ground, we have a good reason to believe that conversation can also be an effective means for building a common ground between humans and agents that represent super intelligence. We know that effective conversation requires a rich common ground shared by participants, while humans and agents have only a poor common ground. A proposed approach of *conversational informatics* to resolve this dilemma is to introduce a notion of co-evolution, as introduced in the next section.

1.3 Primordial Soup of Conversation

Our approach is to create an artificial *primordial soup of conversation* that allows for co-evolution of common ground and communicative intelligence through the accumulation of conversations. We believe that elements of common ground may be made conceivable in conversations; thereby, super intelligence may be able to sense and learn from them for better communication skills.

We aim at creating a network of artificial environments so participants may benefit from conversation in a cyber-physical space, as shown in Fig. 1.7. It consists of numerous "conversation cells" where people and agents make conversation. Events in each cell are recorded for later reuse at the original cell or transfer to other cells. Cells are interconnected so participants in difference cells can virtually communicate

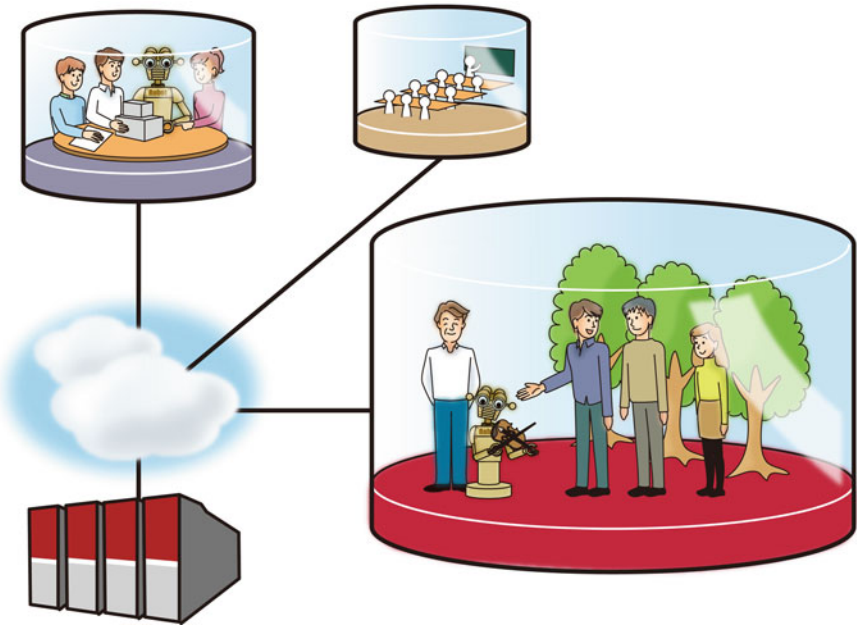


Fig. 1.7 Artificial primordial soup of conversation. © 2014, At, Inc. Reproduced with permission

with each other. How can we make such an environment work as a primordial soup of conversation so that the common ground can emerge as a basis of communication among people and (possibly premature apprentice) agents?

What are the key features for such an environment to serve effectively as a promoter of the evolutionary process of common ground building? The first feature is content. Without rich content, conversations might fade away even though they are somehow initiated. At least, human participants should be able to feel that content is continuously fed into the soup.

General technologies such as augmented reality or augmented virtuality will greatly accelerate the production of attractive content. Projecting supplementary information on the surface of real-world objects will help participants better understand the situation and come up with new thoughts for conversation. Producing a virtual environment from reality, e.g., Google's Street View will provide participants with rather concrete images of the places of mutual interest to increase the opportunity to talk with each other.

It should be noted that records of conversation are valuable content or at least resources from which valued content can be produced. Conversations as content should be well circulated in a community from one cell to another in order to build new thoughts on the basis of subsequent opportunities of conversation.

Content may be more useful if it is interactive, allowing the consumer to interactively enjoy various aspects of the given content in the way he/she likes. Interactive

content may range from menu- or command-based to simulation-based. Interactive content may facilitate creation induced by active engagement of the consumers. For example, students may enjoy an interactive physics textbook in which they may run complex physical phenomena simulations under different parameter settings; therefore, they will gain a deeper understanding from which novel questions may be produced.

The conversation environment may be made smarter by the introduction of intelligent sensors and objects to facilitate conversation. Researchers in ambient intelligence (Weber et al. 2010) believe that embedding the technology in the background to make it invisible so that they can be sensitive, adaptive, and responsive to the presence of people and objects will not only make the environment more friendly and intimate but also enable people to express themselves in an innovative manner.

We believe that the ultimate entities in a conversational environment are *conversational agents*, i.e., possibly physical synthetic agents that can conduct human-like interaction with people as fluently as people do to mutually benefit from other participants in conversations. Unlike, though complementary to, ambient intelligence, conversational agents dispose themselves to talk with people.

What abilities make up such conversational agents that people may admit to regard as being *conversationally intelligent*? First, conversational agents should be able to recognize and produce verbal and nonverbal social signals in conversation. Some social signals are quite subtle and context-dependent, making it difficult for the recognizer to capture them with perfect precision. Second, conversational agents should be able to recognize and produce discourse. Discourse of conversation determines the range of referents in the current utterance. It consists of preceding expressions having been discussed in the discourse and the situation surrounding the participants of the conversation (including the history). Third, conversational agents should be able to carry out joint activities including conversation. Semantic interpretation and production of social signals is necessary, as suggested by Vinciarelli et al. (2012). Common ground needs to be established for joint activities. Participants need to make sure that they share a common ground. If they have doubts, they need to take actions to fix the misunderstanding. Fourth, conversational agents should be able to learn from conversations: build stories and change knowledge accordingly. Conversation is useless unless the participants cannot derive information and insights from conversation to act better and more intelligently in the future. Finally, conversational agents should be able to engage in empathic inter-action and build trusts with each other. Without empathy, participants may not efficiently draw on conversation. The list of desiderata appears quite challenging.

How can we build conversational agents for practical use? In the beginning, the participants of conversations in the primordial soup may only consist of phanerosis or anonymous avatars backed up by real people. In the beginning, it might be quite difficult to develop autonomous conversational agents to which human participants pay much attention, for they may be regarded as a dull partner as their competence in conversation is quite low. The hope is that once we can create a primordial soup of conversation that can provide human participants with a large variety of nutrition, it will produce a large amount of data that can be used to make conversational

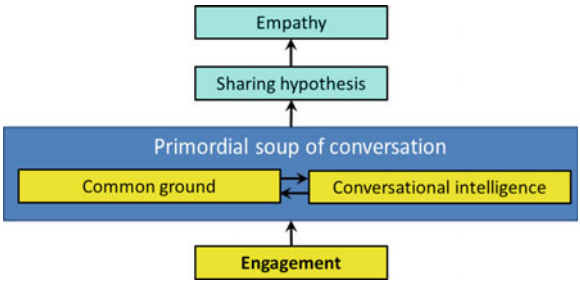


Fig. 1.8 A road to empathic agents

agents intelligent. In particular, a successful regime of structured interaction such as cloud sourcing will create a large number of *mechanized humans* (MHs) whose conversational behaviors will create plenty of high quality, modularized components of knowledge that can be used for pursuing various intelligent tasks. It is quite likely that we can produce *human-like machines* (HMs) by extracting patterns of behaviors demonstrated by MHs. Gradually, the population of HMs may increase, making the primordial soup larger so that more evolution may take place.

In order for an ensemble of MHs and HMs to evolve over time to function as an artificial soup of conversation, strong motivation of participation or *engagement* seems a key to success. With strong engagement, participants may stay long in the primordial soup of conversation to contribute to building common ground, otherwise the soup will shrink and disappear as a result of cultural selection.

What motivates participants to engage? We suspect that the sum of substantial and illusionary sense of value produced by the interactions resulting from engagement provokes active participation and motivates participants to stay therein.

Figure 1.8 summarizes the structure of discussions in the above. We may rely on the sharing hypothesis to realize empathic agents, which may in turn be enabled by building a primordial soup of conversation. Key factors of a primordial soup of conversation are common ground and *conversational intelligence* resulting from conversational agents. The two-factors are reciprocal each other, expected to co-evolve together.

What deserves serious efforts at this stage? It seems that we need not seek for perfection. It is important to actually set out for a technically modest goal target as a feasible next step for a long voyage. It appears that a promising next step might be to build a minimal machinery for launching a primordial soup of conversation. We may not elaborate a technique for building advanced conversational agents. Rather we need a mechanism for circulating conversations in a community. It will contribute to building a firm common ground for the next stage where people and agents are connected together by strong empathic ties.

To make this happen, we need to take a methodologically proper approach. Studies into conversational informatics need to be highly empirical; hence, need to cover a rather comprehensive range of aspects encompassing: theory, platform, measurement

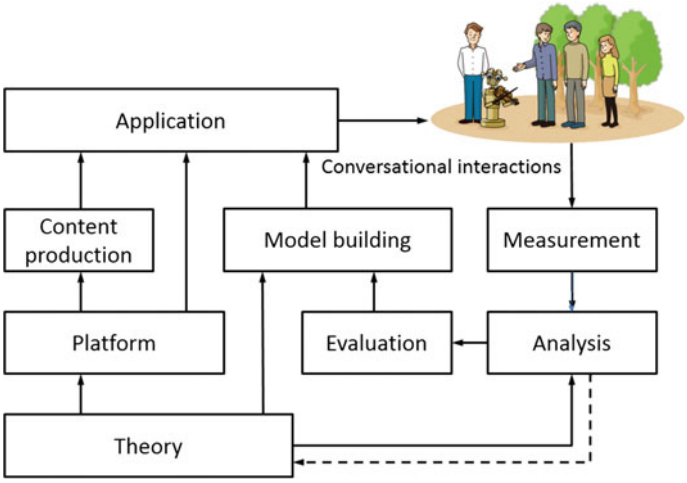


Fig. 1.9 Methodology of conversational informatics. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

and analysis, model building, content production, application, and evaluation. Platforms need to be built to support a broad range of conversations conducted by people, their avatars, and artificial conversational agents across cyber and physical spaces. The challenge is to build a conversational agent that can participate in conversation with humans to contribute relevant stories and update its memory according to what has been learned from the conversation. A method for building a conversational model is needed not only to understand how people conduct conversation in a structured fashion but also to develop the communicative behaviors of conversational agents. A suite of tools needs to be built to help researchers quantitatively specify the ideal behaviors of conversational agents in a wide variety of conversational situations. Content production needs to be effectively assisted so that content producers can easily create content for augmented conversation systems without much technical knowledge about the conversation augmentation system. Applications are not only beneficial for the society but also valuable to researchers as opportunities to learn real-world problems. Evaluation is needed to understand the achievement and limitation of individual projects.

There are dependencies among these issues, as shown in Fig. 1.9. Platform and analysis depend on theory, which may be revised as a result of analysis. Applications draw on the platform, model building, and content production. Application will bring about conversational interactions for analysis from which measurement will be made, and insights are obtained for evaluation which will be in turn reflected on model building. Compared with an approach aiming at pioneering a new theory of communication, such as proposed by Wachsmuth et al. (2013) in whichalignment in communication is focused, our approach places more emphasis on content and data-oriented evolutionary process than on interaction. In spite of the differences, the

two approaches are complementary to each other and integrated in a fruitful fashion. For example, the alignment mechanism may contribute to enhance engagement and empathy.

1.4 Organization of This Book

In what follows, we start by surveying previous studies on conversations in Chap. 2. It ranges from story-telling and narratives to cognitive processes. On the surface, we look at major work in communication science and anthropology where thoughtful observation and analysis plays a critical role. Beneath the surface, in contrast, we see how observable processes are supported by cognitive neuro processes in the brain. In Chap. 3, we overview historical development of conversational systems. In Chap. 4, we describe a collection of standard techniques developed so far to develop conversational agents. In Chap. 5, we introduce a theory of conversation quantization that underlies the research and development of conversation agents. We then turn to the technological aspects of our work. In Chap. 6, we introduce an immersive conversation environment called ICIE that serves as a platform of our research on conversational informatics. In Chap. 7, we specialize our view on computer vision techniques that play a critical role in both recognizing human behavior and producing content. Chapter 8 is about measurement, analysis, and modeling of conversations. It is more scientifically motivated, aiming at unveiling the mechanisms governing human's behavior on/beneath the surface. Chapter 9 addresses learning by imitation for producing conversational agents' interactions from observation or by demonstration. Chapter 10 presents an application and extension of the framework presented in Chap. 9. Chapter 11 focuses on sensing and engineering the cognitive process people rely on making conversations creative. Chapter 12 discusses high-level issues such as social intelligence design, ethics, and empathy. Chapter 13 concludes the entire discussions.

1.5 Summary

Conversational informatics focuses on communicative aspects of intelligence, aiming at understanding and augmenting conversations—a fundamental human activity. In contrast to conventional research on artificial intelligence oriented towards autonomous intelligence, conversational informatics attempts to bridge human society and computational intelligence. The engineering goal of conversational informatics is building empathic agents that can dynamically establish empathic relationships with humans and other agents by accumulating conversations. Instead of directly diving into building full-fledged empathic agents, we consider that building an ever-evolving primordial soup of conversations—an ensemble of mechanized humans and human-like machines—should be promising. In order to function in a

maximally effective manner, a primordial soup of conversations should be designed so that dense and meaningful interactions among MHs and HMs may happen, as suggested by the sharing hypothesis—the more common ground is shared, the more empathy is gained. Evolution of a primordial soup of conversation will be supported by synergy of the common ground and conversational intelligence. The key issue should be a strong engagement by participants. Conversational informatics spans a broad range of interrelated topics from science to engineering, and from the theoretical to empirical approaches. A data-intensive approach presented in this book exploits abundance of data rapidly growing to cover a wide spectrum of our conversational behaviors.

Chapter 2
Conversation: Above and Beneath the Surface

Abstract Conversation is complex. It is amazing how easily participants coordinate their actions to establish a discourse and make points often with little conscious effort in daily conversation save for a few special cases. To date, numerous authors have investigated conversations from a wide variety of angles. In this chapter, we overview major theories that enable us to understand conversation in a structured manner.

Keywords Narrative · Social discourse · Focused gathering · Joint activity theory · Turn taking system · Affective computing · Theory of mind

2.1 The Horizon of Conversational Communication

Throughout the history of scientific and engineering research, simplification and abstraction have often been the means to making a project feasible. Without simplification and abstraction, we are often unable to draw any useful and reliable conclusions from such scientific and engineering projects; however, an over-dependence on simplification and abstraction may cause researchers to become blind to the very phenomena they are studying, hindering them from recognizing important phenomena, even if such phenomena materialize in front of them by chance. This is quite the opposite of serendipity.

Therefore, our approach to resolving this problem is to introduce a number of prescribed viewpoints such that we do not miss any insights from this set of standard angles. More specifically, as shown in Fig. 2.1, we look at conversations from the following five viewpoints: verbal communication, nonverbal communication, social discourse, narratives and content flow, and cognitive processes. Although verbal communication might seem the most central viewpoint, as one might choose verbal communication if asked to select just one abstraction of conversation, verbal communication is usually focused on describing the content to be communicated rather than the structure of the conversation or the speaker's underlying attitude and emotions.

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_2

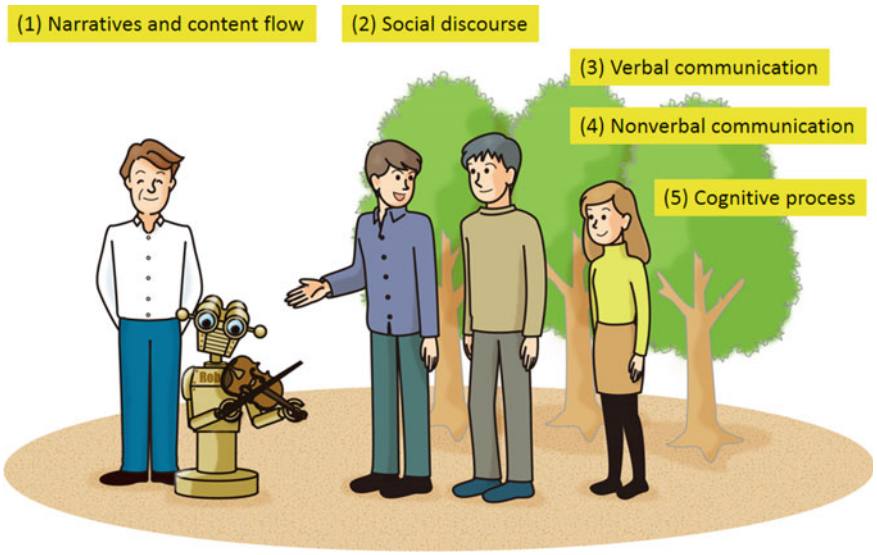


Fig. 2.1 Five viewpoints employed in this book to look at conversation. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Nonverbal communication may disclose a speaker's attitude and emotions; furthermore, it may supply information that complements the verbal utterances or provide social signals that control the flow of conversation.

Social discourse sheds light on the social implications of a conversation; presumably, people converse to together achieve social goals. The social discourse viewpoint highlights the function that a given conversation may play in accomplishing social goals at higher levels of abstraction. The narratives and content viewpoint focus on the content-oriented aspects of conversation, particularly how pieces of conversation contribute to the structure and organization of the content to be shared among participants or an individual participant's personal memory organization. Finally, the cognitive process viewpoint enables us to look inside the mental space of each participant by studying what processes allow each participant to interpret or produce verbal and nonverbal behaviors in conversation.

2.2 Stories and Narratives

At a coarse approximation level, a conversation is characterized as a process of exchanging small talks or chitchats that are regarded as components of larger stories. Occasionally, small pieces of conversation might be brought together by participants to jointly compose a new piece of conversation. Participants may take home a collection of new pieces of conversation for further extension in the course of individual life, where pieces of conversation will be recalled to interpret new experiences or

contrast with other pieces of stories or be synthesized with an individual's own stories.

Dawkins (1976) uses memes, or cultural genes, to explain the evolutionary process of cultural development. Incomplete imitation may play an important role in the evolutionary process that effectively functions in search of new ideas. Conversations can be regarded as an important class of memes that can be circulated in society via inheritance, mutation, selection, or crossovers.

Researchers in narratology, discursive psychology, and social constructionism consider storytelling a source of our intelligence. They believe that stories and narratives are not only a means for communicating messages with each other, but also for recognizing and understanding the world (Edwards 1997). Our cognition depends on the language used to describe cognition. A group led by Schank formulated a theory of dynamic memory to account for how people learn from stories to build dynamically evolving memory (Schank 1982).

People can indeed manage complex information by structuring it based on a collection of organization principles. According to discourse theory by Brown (1983), topics play a critical role in recognizing a collection of information as a cluster. A story is a staged presentation of complex information. Coherency of information presentation is critical for the audience to understand the presentation without difficulty.

2.3 Conversation in a Social Discourse

Goffman established a series of seminal works on observing and analyzing people's behaviors from the viewpoint of sociology. Not surprisingly, conversation was within the scope of his analysis, but from much wider perspectives, such as gatherings. Goffman (1963) highlighted behaviors in public places, as opposed to those in private. The target of the analysis was the situation, i.e., the relationships people may exhibit in relation to other people in public, where members of a given society can act freely. Goffman attempted to unveil situational properties underlying social occasions that exist when a group of people in a shared public place influence one another. Interactions were classified as follows: (1) *unfocused interactions* are provoked to manage contingent encounters, such as people without any social relationship happening upon one another; and (2) *focused interactions* occur when people actively engage in openly collaborating with each other to achieve some joint goal. His findings involved imaginary objects, such as an involvement shield, which only exist in peoples' behavior in the sense that people act as if there were an invisible wall that separates them from others; another example is civil inattention, which people often exhibit to avoid social overhead when they are within a distance that compels them to initiate some actions to be deemed polite. These phenomena clearly do not result from any physical laws of nature, but rather from principles for managing social relationships.

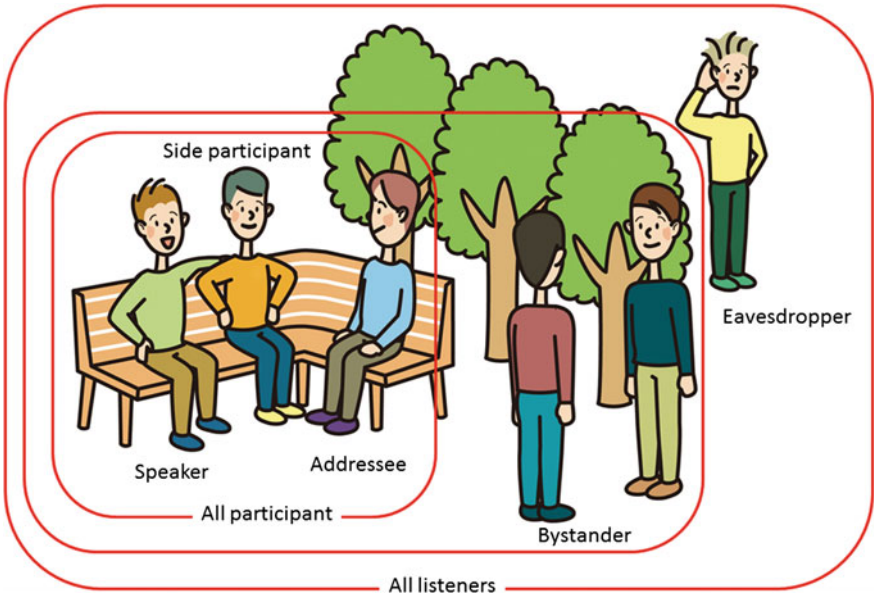


Fig. 2.2 The type of participation. Drawing inspired by Goffman (1981). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Goffman (1981) analyzed conversations from the viewpoint of ritualization, trying to uncover social mechanisms for participation, in particular conversations. As illustrated in Fig. 2.2, he classified participants into the following five categories: *speaker*, *addressee*, *side participants*, *bystanders*, and *eavesdroppers*. The speaker and the addressee are those who have the deepest engagement in the conversation. Side participants are those who are close to the speaker and the addressee, and may have previously been a speaker or addressee. Furthermore, they have good reason to become the speaker or addressee in the near future.

Kendon (1990) noted that there is regularity in the spatial arrangement of participants in conversation. As shown in Fig. 2.3, Kendon identified one such pattern sustained during conversation called the *F-formation*, in which each participant occupies an equal, direct, and exclusive position around an O-shaped area (i.e., *O-space*). Kendon pointed out that the spatial arrangement generates social influences on both the participants and non-participants, because the F-formation introduces three functional spaces near the participants that are introduced by a narrow zone called the *P-space*, which delineates an area surrounded by the participants.

The O-space, a space encapsulated within the P-space, is a zone agreed to be reserved for interactions among the participants. Even though the F-formation is often observed in a free-standing area, such as a party, only the participants comprising the F-formation are allowed to control the interactions within the O-space. The third space, which is called the *R-space*, extends backward between the participants and serves as a gateway for newcomers to approach and obtain approval to join

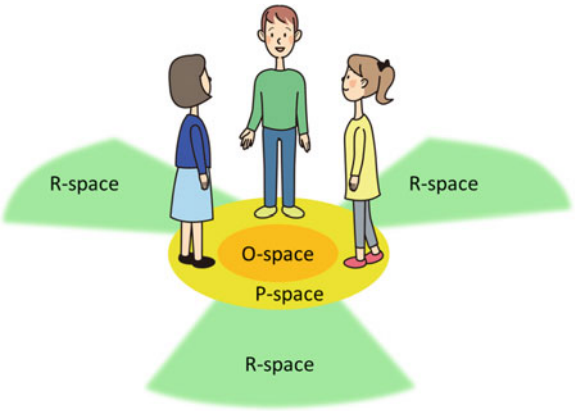


Fig. 2.3 F-Formation. Drawing inspired by Kendon (1990 p. 209). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

the F-formation. As people usually move around in an open space, F-formation is dynamic in nature. It is created when two or more people start a conversation and disappears when participants finish their interactions and leave the focused gathering. During interactions in an F-formation, peoples' behaviors are coordinated in a frame attunement as noted by Kendon (1990 p. 253).

Kendon also observed how people exchange greetings with one another, where a greeting was defined as a unit of social interaction when people enter a focused gathering. Kendon videotaped a party and investigated how people greet one another. Figure 2.4 shows an illustrative example in which guest *B* approaches the host *A*. Distinctive gestures in greeting interactions are called salutations. Greetings normally begin with sighting, orientation, and approach. Greetings may be initiated even when potential participants are far away from each other; such occasions are called distance salutations and include such behaviors as a head toss, a head lowering, a nod, or a wave. Distance salutations may often be followed by a head dip or a lowering of the head via a forward bend of the neck.

While a party approaches, one observes characteristic behaviors, such as glancing, looking, or gazing, and facial orientation varies depending on the looking behavior. Also, body cross and grooming is occasionally observed. In the former, a participant crosses one or both of the arms in front of the upper body; in the latter, he or she adjusts his or her clothing, for example straightens a tie or strokes the hair. When the two participants make their final approach, they may smile, set a head position, or make a presentation with a palm. Finally, the greeting sequence comes to an end, i.e., the close salutation, which may include a handshake or an embrace.

A newcomer need not always wait until the current conversation is completed to participate. Instead, he or she can wait at a distance and send weak social signals, such as synchronization of behaviors while avoiding any strong social signals, which include direct eye contact, until he or she receives a salutation display. Thus, a greeting

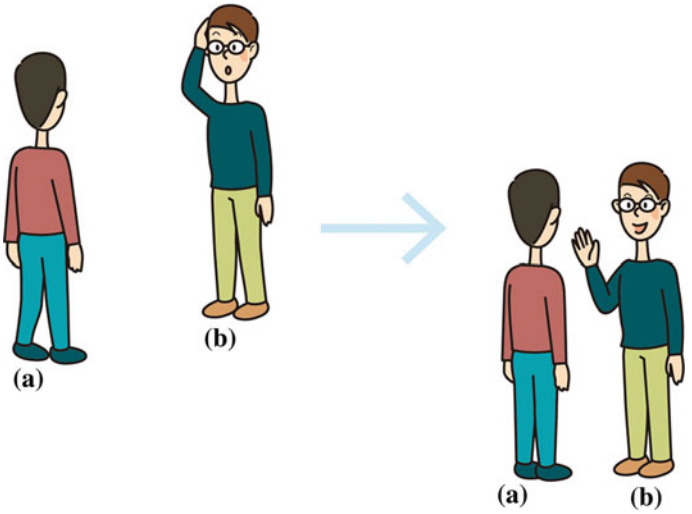


Fig. 2.4 Greetings (2 party). Drawing inspired by Kendon (1990). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

might comprise the sequence illustrated in Fig. 2.5, in which the person on the right has arrived during the dialogue between the other two. The person on the right sends a weak social signal, such as a synchronization of movement or a short glance, to the person on the left. If he recognizes the signal and is ready, he will send a salutation, such as a head toss, to which the person on the right may respond with a head lowering and hand raise, followed by a handshake and an extended conversation.

2.4 Interactions in Focused Gatherings

During focused gatherings or a core of conversation, participants exchange various social signals to coordinate their behaviors. Signs of focused participation include such phenomena as mutual attention, in which both participants look at each other (Fig. 2.6a), or joint attention, in which participants gaze at an object in focus of the talk (Fig. 2.6b). Sometimes, the signal is not as explicit, as is seen in synchrony in which participants behave in the same patterns, as illustrated in Fig. 2.7; however, the boundary between participation and non-participation is blurry. As illustrated in Fig. 2.8, the status of participation, or *engagement*, varies in terms of facial expressions, head directions, poses, and so on.

The social signals may adhere to a ritual by their very nature, expressing rather subtle care for other participants that can be clearly identified by people who share the same cultural background. For example, in Fig. 2.9, both fragments of conversations C1 and C2 are simple pairs of questions and answers concerning time. Although

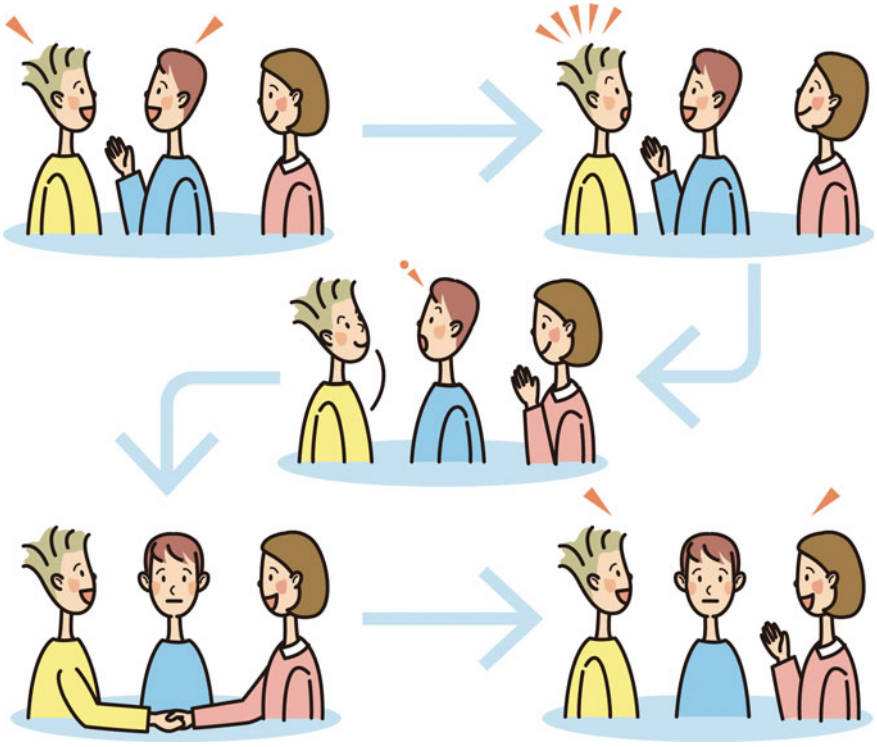


Fig. 2.5 Greetings (3 party). Drawing inspired by Kendon (1990). © 2014, At, Inc. Reproduced with permission

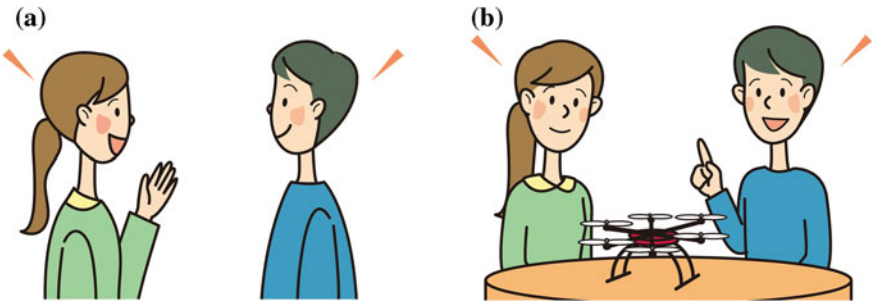


Fig. 2.6 Signs of focused participation. © 2014, At, Inc. Reproduced with permission. **a** Mutual attention **b** Joint attention

the transaction in C1 is quite simple and straightforward, C2 involves an exchange of ritual signals, such as remedy to neutralize the potentially offensive consequence of encroaching on another with a demand, relief to demonstrate that the potential offender's effort to nullify offense is acceptable, appreciation or a display of gratitude

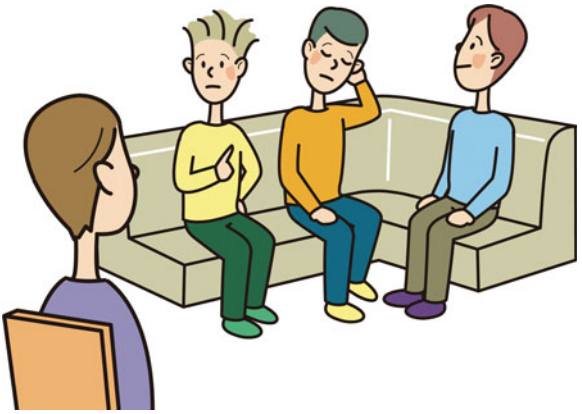


Fig. 2.7 Synchrony. Drawing inspired by Kendon (1990). © 2014, At, Inc. Reproduced with permission

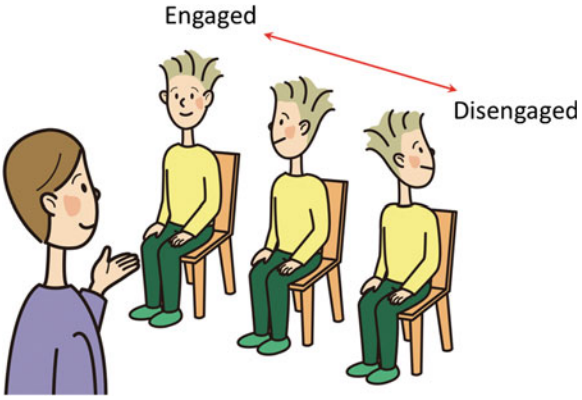


Fig. 2.8 Different degrees of engagement. Drawing inspired by Goodwin (1981). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

for the service rendered and for not taking the claim the wrong way, and minimization, which demonstrates that enough gratitude has been displayed.

Goffman (1967) introduced the concept of face as the positive social image a person may claim for himself in a line of interactions with other participants in a social encounter. A participant may be satisfied if he or she can maintain a face; otherwise he or she is out of face. Actions a person takes to ensure whatever he or she is doing is consistent with face is called facework. Although an offending or threatening face is normally avoided, a corrective process, such as providing compensations to the injured, may be conducted even when an incidental offensive action is recognized. Goffman called these processes ritual, as the actors resort to socially approved conventions to manipulate social values.

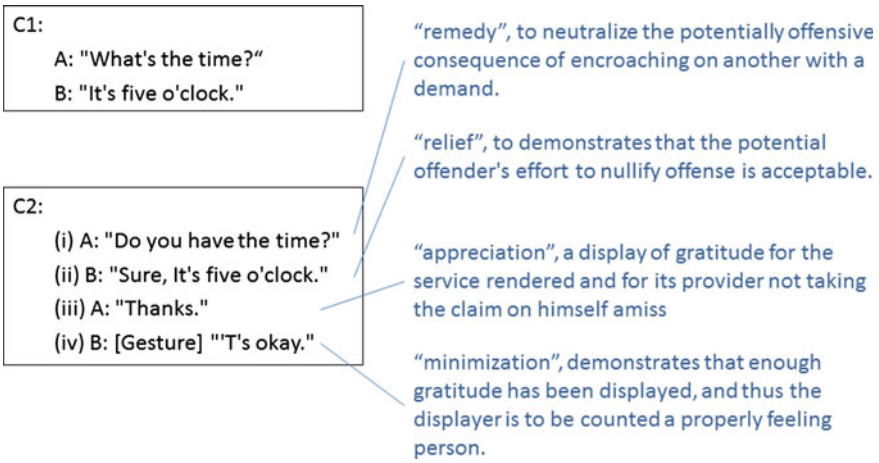


Fig. 2.9 Exchange of ritual signals in conversation (Goffman 1967)

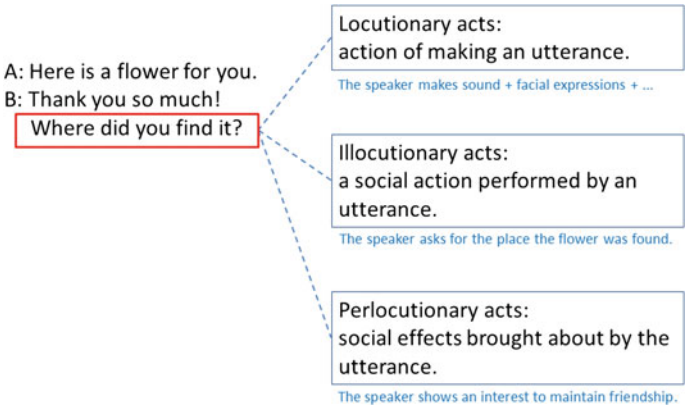


Fig. 2.10 Locutionary acts, illocutionary acts, and perlocutionary acts (Austin 1962)

Speech act theory analyzes utterances of conversation with respect to social implications. Austin (1962) classified social implications into the following three categories: *locutionary acts* that refer to the very action of making an utterance, *illocutionary acts* that given utterances make as social acts, and *perlocutionary acts*, which refer to the social effect caused by the given utterances. Figure 2.10 shows an illustrative example of the classification introduced by Austin. In the figure, after person A gives person B a flower, person B asks, "Where did you find it?" The locutionary act is the sound and associated nonverbal expressions the speaker makes; the illocutionary act indicates that the speaker has asked where the flower was found; and the perlocutionary act reveals that the speaker shows an interest in maintaining the relationship.

Unfortunately, speech act theory is limited, even after improvements by Searle, Grice, and other authors, because it identifies individual actions of the social actor in an isolated context. Joint activity theory was therefore proposed to remedy this fundamental weakness, and is discussed in the next section.

2.5 Joint Activity Theory

In joint activity theory, Clark (1996) criticized speech act theory, claiming that it had not paid much attention to the interactive aspects, collaborations, or coordination among participants, which should be of primary importance in conversation. The basic hypotheses employed in joint activity theory are comprised of the following six propositions: (1) language is fundamentally used for social purposes; (2) language use is a species of joint action; (3) language use always involves speaker's meaning and addressee's understanding; (4) the basic setting for language use is face-to-face conversation; (5) language use often has more than one layer of activity; and (6) the study of language use is both a cognitive and a social science.

According to joint activity theory, participants of conversation build a layered space of new information on the common ground using shared coordination devices. Actions of participants consist of multiple levels of abstraction, or *action ladders*, in which the actions on the upper levels are realized by those on the lower levels. Clark called these functions "downward evidence" and "upward completion."

For example, as shown in Fig. 2.11, Adam said to Bart, "Sit down here would you" and used a pointing gesture. As shown in the figure, the interaction between Adam and Bart may be analyzed at four levels. At Level 1, the execution-and-attention level, language use is characterized in terms of physical interactions among participants. At Level 2, the presentation-and-identification level, the interaction is specified in terms of information. At Level 3, cognitive terms are used to describe interaction. Finally, at Level 4, interaction is modeled as social interaction. Figure 2.12 shows via an action ladder how speech act theory is reformulated and extended into interactions at multiple levels.

As summarized in Table 2.1, Clark argues that signs in language use are classified into *icons*, *indexes*, and *symbols*. Icons perceptually resemble the referenced object and are used to demonstrate things. Indexes produce a physical connection with the referenced object and are used to indicate things in the environment. Symbols are associated with the target objects by rule and are used to describe the types of things.

People use various parts of their body for signaling, as summarized in Table 2.2. These phenomena include both polymorphism (for example, a given social action such as assent may be realized by more than one medium) and polysemy (for example, a nodding may mean assenting or just an acknowledgment of a received mess age).

Rather than a one-way open loop, information flow is normally a reciprocal closed loop with synchronous or asynchronous feedback. Clark introduced the notion of *tracks* to account for the channels of information flow. Track 1 is dedicated to official business, conveying propositions in the higher level social interactions for which

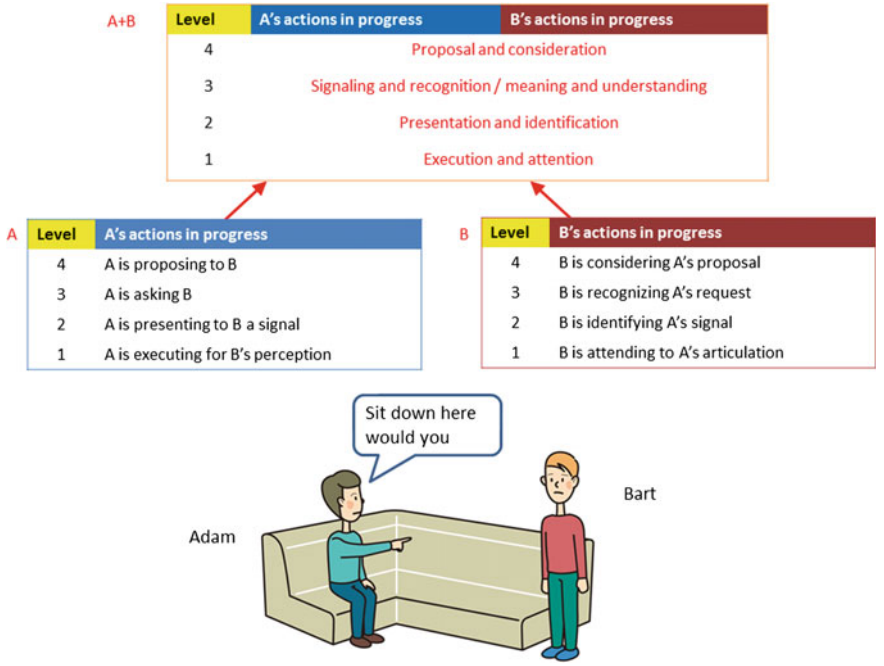


Fig. 2.11 Participants' abstracted actions on multiple levels. Drawing inspired by Clark (1996). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

A -> B: "please sit down"

Perlocutionary act	*A* is trying to get *B* sit down.
Illocutionary act	*A* is asking *B* to sit down.
Locutionary act	*A* is saying to *B* "Please sit down."
Rhetic act	*A* is using the words *please*, *sit*, and *down* with a certain sense and reference.
Phatic act	*A* is uttering the words *please*, *sit*, and *down*.
Phonetic act	*A* is producing the noises that constitute "Please sit down."

downward evidence ↓

↑ upward completion

Fig. 2.12 Speech acts as action ladder (Clark 1996)

the conversation is used. In contrast, as summarized in Table 2.3, Track 2 contains control signals needed for bidirectional transmission of signals, such as acknowledgment of receipt of a signal, or correction of signals. Tracks may be recursively embedded; for example, Track 3 concerns information transmission in Track 2. Clark also emphasized the importance of common ground, which is a self-awareness that can serve as a basis for communication. Table 2.4 enumerates the types of common ground used in conversation.

Table 2.1 Signs (Clark 1996)

Type of sign	Relation of sign *S* to its object *O*	Method of signaling
Icon	*S* resembles *O* perceptually	Demonstrating a thing
Index	*S* is physically connected with *O*	Indicating a thing
Symbol	*S* is associated with *O* by rule	Describing as a type of thing

Table 2.2 Media for signaling (Clark 1996)

Instrument	Describing-as	Indicating	Demonstrating
Voice	Words, sentences, vocal emblems	Vocal locating of "I", "here", "now"	Intonation, tone of voice, onomatopoeia
Hands, arms	Emblems, junctions	Pointing, beats	Iconic hand gestures
Face	Facial emblems	Directing face	Facial gestures, smiles
Eyes	Winks, rolling eyes	Eye contact, eye gaze	Widened eyes
Body	Junctions	Directing body	Iconic body gestures

Table 2.3 Grounding using Track 2 (Clark 1996)

Track 2 signal	Example	Interpretation
Trial constituent	A: a man called Annegra? B: yea, Allegra	"Confirm that you know who I mean by Annegra"
Installment	A: so Mr. D. Challam, B: yes	"Confirm that you understand this installment"
Fade-outs	A: you know, she's just gonna B: yeah	"I am sure you understand without my completing this"

Table 2.4 Types of common ground (Clark 1996)

1. Communal
a. Human nature
b. Communal lexicons
c. Cultural facts, norms, procedures
2. Personal
a. Perceptual bases gestural indications, partner's activities, salient perceptual events
b. Actional bases
c. Personal diagies

In conversation, participants often hypothesize an imaginary situation and discuss various pertinent issues. Clark introduced the notion of layer-to-model possible worlds dynamically created in conversations. This kind of interaction is quite useful in education and training such that participants can establish a deep understanding

Table 2.5 The structure of ostensible invitation (Clark 1996)

1. Joint pretense. A engages B in a joint pretense
2. Communicative act. The joint pretense is that A_i is performing a sincere communicative act toward B_i
3. Correspondence. A is to be taken as A_i, and B as B_i
4. Contrast. A intends A and B to mutually appreciate the salient contrasts between the demonstrated and actual situations
5. Deniability. If asked, A would deny meaning for B what A_i means for B_i

of the given topic by taking into account what could have happened in a given situation (but did not actually happen). Furthermore, hypothetical situations may be used to communicate sophisticated feelings or thoughts. Participants may be readily participating in a joint play to actively communicate and share subtle emotions. Ostensible communicative acts are used to communicate politeness by introducing a hypothetical world in which participants of conversation pretend to play a given role to communicate their delicate feelings. Table 2.5 shows an elegant case described by Clark in which layers are used for analysis.

High-level issues in communication such as politeness may be analyzed using a notion of social objects. Based on the arguments by Goffman (1967), Clark discusses how people try to achieve social equity, or *face*, to maximize the outcome of social interactions and minimize the distress assumed to be resulting from inequitable situation. Brown and Levinson (1978) discusses in detail the generic principles for politeness in language use, based on the assumption that politeness is rationally derived from mutual knowledge assumption about face.

2.6 Integrating Multiple Modalitis to Make Sense

In conversation, the speaker usually coordinates multiple parts of his or her body, including his or her eyes, face, hands, head, and torso; the speaker does this to express what he or she wants to communicate (Knapp and Hall 2010; Richmond et al. 2004). As shown in Fig. 2.13a, a *pointing gesture* is used to associate utterances with the objects or events in the environment surrounding the speaker. *Illustrative gestures* (e.g., Fig. 2.13b) are used to demonstrate a certain visual feature of the referent. Furthermore, a *precision grip* or a gesture with a narrow symbolic interpretation (e.g., Fig. 2.13c) may be used to supplement information with a speaker's utterances.

Kendon (2004) defined gesture as a visible part of the human body that can function as a part of utterance and analyzed how people use gestures as an integral part of utterances, or gesticulation, by investigating the records of monologues in detail. Kendon investigated the time domain and observed in detail how speech and gesture components are coordinated, or orchestrated, to produce meaning. He then studied how gesticulation adds meaning to utterances; for example, by visually

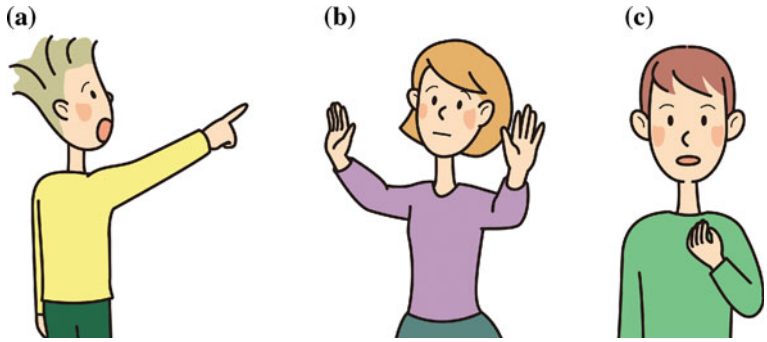


Fig. 2.13 Example of gestures as a part of utterances. **a** Pointing (Kendon 2004) **b** Illustrative (Kopp et al. 2007) **c** A Precision grip (Kendon 2004). © 2014, At, Inc. 2014. Reproduced with permission

illustrating size, shape, structure, and spatial arrangement. Temporal arrangement of speech and gesture is used to specify how the meanings conveyed in different tracks by speech and gestures are combined.

Furthermore, Kendon detailed *narrow glosses*, or gestures with an almost uniquely identifiable interpretation, in studying how gestures co-occur with speech to contribute to the referential meaning of what is spoken. Kendon pointed out that narrow glosses co-occurring with equivalent verbal expressions may add pragmatic meaning, such as emphasis, such that the utterance may not be ignored by the listener. In contrast, he further noted that narrow glosses co-occurring with non-matching verbal expressions may bear additional propositions beyond what is spoken.

Kendon suggested that gesture families with similarity in shape or movement patterns share their own semantic theme in terms of their pragmatic function. For example, as illustrated in Fig. 2.13c, a gesture family called grappolo ("bunch") or a G-family in the gesture family called precision grip marks the topic of a speaker discourse or shares such meaning as essence, substance, core, or heart. Gestures in the R-family, in which the thumb and the index finger are in contact in the form of a ring, are used when the speaker appears to demand something precise. Kendon also suggested that gestures using an open hand that is prone share the theme of stopping or interrupting a line of action in progress, whereas those using an open hand (as shown in Fig. 2.14) are accountable in terms of the theme of offering and receiving.

Communication behaviors are influenced by personal and cultural factors. The five-factor model (FFM) of personality (McCrae and Costa 1987; Costa and McCrae 1992; McCrae and Costa 1997) characterizes personal traits in terms of five basic dimensions: *openness*(O) to experience or culture, *conscientiousness* (C) or will to achieve, *extraversion* (E) or surgency, *agreeableness* (A) versus antagonism, and *neuroticism* (N) versus emotional stability. The model is based on a wide range of investigations ranging from longitudinal and cross-observer studies to linguistic studies of adjectives. Differences in personal traits may arise everywhere depending

Fig. 2.14 Pointing with open hand supine (palm up). Drawing inspired by Kendon (2004). © 2014, At, Inc. Reproduced with permission.

on the degree of the individual differences of mental software that may be caused not only by national culture but also by educational background, business practice, and even by age. Those difference may be grouped together depending on a culture, or the social background shared by people. Hofstede(2001), Hofstede et al. (2002), Hofstede and Hofstede (2005) elaborate on the intuition that a culture is a mental programming or software of the mind. Their uniqueness lies in the use of five dimensions to parametrically specify national culture: *identity* for collectivism-individualism, *hierarchy* for large *versus* small power distance, *gender* for femininity *versus* masculinity, *truth* or anxiety for strong *versus* weak uncertainty avoidance, *virtue* for long- *versus* short- term orientation. CUBE-G project presented in Sect. 4.4 in this book is based on Hofstede's dimensional model.

2.7 Turn-Taking System

Most societies have developed a system of interaction for regulating and organizing message flow to promote efficient communications in spoken language (Goffman 1955). This conjecture was substantiated by authors in conversation analysis.

Kendon (1967) reported an early attempt to analyze the *system of turn-taking* by focusing on gaze direction during conversation. As shown in Fig. 2.15, Kendon devised a conversation recording system to analyze how eye gaze behaviors correlated with turn-taking in dialogue. Thirteen undergraduate students from Oxford University were asked to introduce themselves to the auditor. Kendon and his colleagues observed the following two types of gaze behaviors: (1) *q-gaze* in which the gaze of the speaker (p) is directed to the conversation partner (q); and (2) *a-gaze* in which p is not looking at q. As shown in the figure, a small mirror was placed on the table such that a single video-tape recorder could record the two participants' faces in the dialogue in a single frame and the timing of the utterance and corresponding gaze behaviors of the two participants could be quantitatively compared.

Kendon obtained a number of interesting results from this rather simple experiment. First, as illustrated in Fig. 2.16a, q-gaze is mostly associated with the role of the speaker, whereas a-gaze is associated with the listener (i.e., the auditor, according to

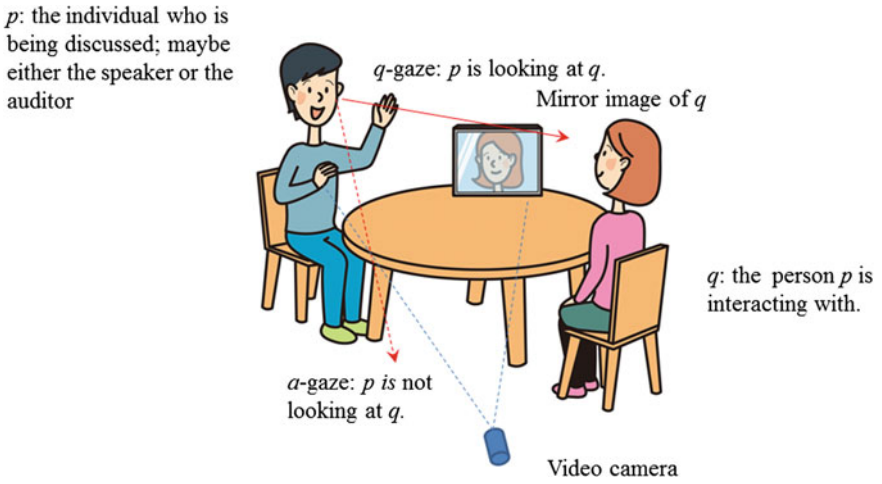


Fig. 2.15 A dialogue setting used in Kendon (1967) to analyze gaze behaviors in dialogue. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

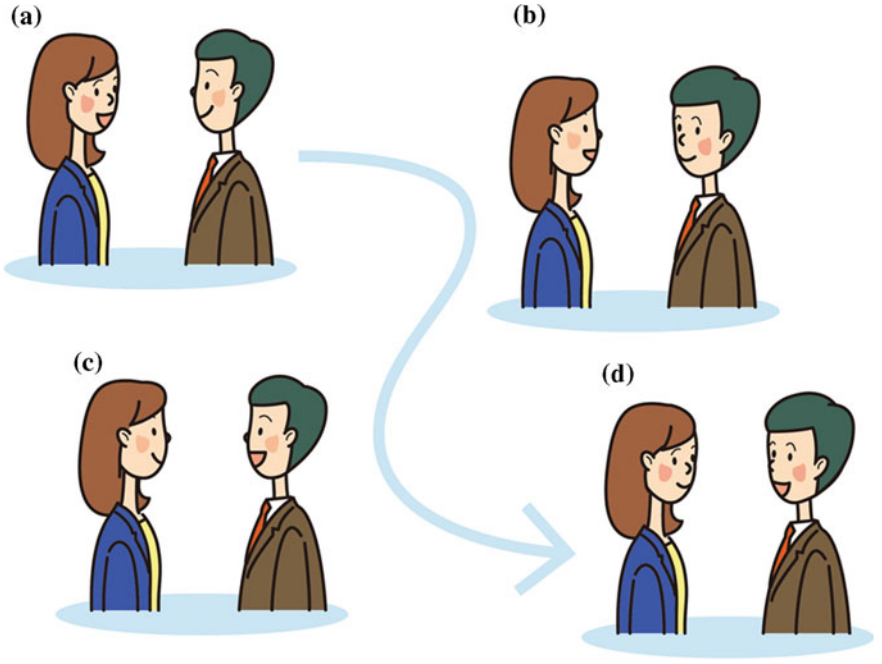


Fig. 2.16 Turn Taking. Drawing inspired by Kendon (1967). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Fig. 2.17 Back channel. Drawing inspired by Yngve (1970). © 2014 At, Inc. Reproduced with permission

Kendon's terminology). Second, the average length of duration of *q*-gazes displayed by auditors depends on who the speaker is.

Third, a sequence of normal gaze behaviors during turn-taking was identified; as shown in Fig. 2.16b, the current speaker starts to make a *q*-gaze as he or she comes to the end of his or her turn, while the current auditor starts to make an *a*-gaze during that same time period. After the next speaker starts speaking, he or she will start to gaze at the next anticipated auditor. In Kendon's interpretation of the *q*-gaze of the previous speaker, as illustrated in Fig. 2.16c, d Kendon noted that the role of the previous speaker is to monitor the turn shifting to the next speaker. The interpretation of the *a*-gaze of the next speaker is to reject interruption and prepare the next utterance.

Fourth, gaze behaviors differ depending on whether the speaker is finishing and preparing to yield his or her turn to the next speaker; i.e., although the speaker stops speaking, he or she might simply want to pause and retain the turn after a short break. Fifth, gaze behaviors of the auditor may differ based on whether he or she intends to send a signal to follow the speaker or assent what the speaker has said. The auditor tends to make strong *q*-gazes if the former and *a*-gazes if the latter. Sixth, a participant may avoid too much mutual gaze, as mutual gaze may often induce a strong sense of affective involvement.

Yngve (1970) reported an early study on observing back channels in conversation using a recording setting similar to that used by Kendon and his colleagues (depicted in Fig. 2.17). Yngve found that *a back channel* was often made simultaneously with the speaker's utterance, less regularly than turn-taking signals. He pointed out that the notion of turn is sometimes unclear and hence better classified turn in a narrower sense, noting that floor may be intervened by short breaks.

Duncan (1972), Duncan and Niederehe (1974), Duncan (1974) identified six signals relevant to turn-taking and presented a model of a turn-taking system, as illustrated in Fig. 2.18. A *speaker continuation signal* is produced by the speaker directly after he or she has obtained the turn and when he or she starts speaking. A *speaker within-turn signal* is produced by the speaker when the utterance comes to the boundary of grammatical clauses; this signal is used to indicate the speaker's intention to continue his or her current utterance, possibly to be followed by auditor back channel signals or speaker continuation signals. A *speaker gesticulation signal* is emitted via

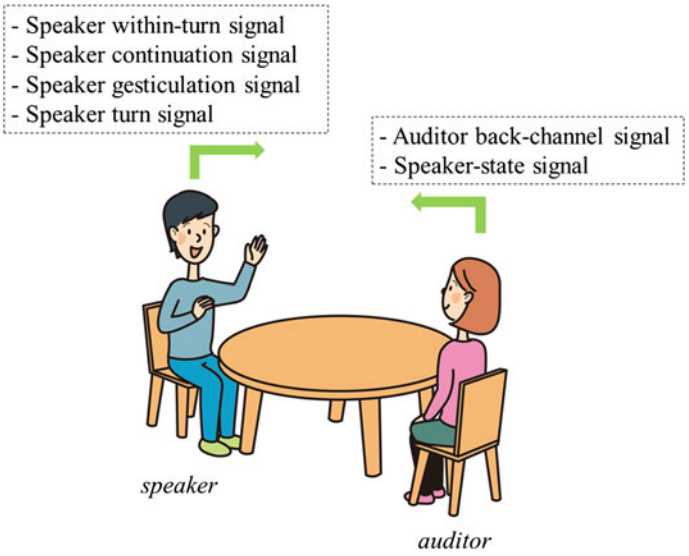


Fig. 2.18 A model of turn taking. Drawing inspired by Duncan (1974). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

a speaker's gestures, typically to suppress turn yielding requests from the auditor. A *speaker turn signal* is used to indicate that the speaker is granting permission for the auditor to take his or her turn. If the auditor wants to actually take the turn as the next speaker, he or she generates the *speaker-state signal*. Throughout all of this, the *auditor back-channel signal* may be employed to indicate that the auditor is attending to the speaker's utterances. Although the model might appear to be quite idealized (as criticized by numerous researchers in conversation analysis, including Goodwin and Sacks), it provides a sense of nominal turn-taking behaviors.

In conversation analysis, researchers try to analyze conversations as they are without idealization, and the system of conversation has been elaborated from the viewpoint of social interactions (Sacks et al. 1974; Schegloff 1968, 1999; Schegloff and Sacks 1973; Schegloff et al. 1977).

2.8 Cognitive Process

Scientists and engineers know how important looking for complex phenomena from the inside is, as hypothesizing an internal mechanism may allow us to understand complexity in a structured fashion. It also helps us artificially reproduce the process. In this subsection, we overview major approaches in psychology and cognitive neuroscience that may help us better comprehend the complex phenomena regarding conversation.

Early attempts of modeling the cognitive process underlying conversations were made in cognitive linguistics. Lakoff (1987) discussed how the cognitive process

works when people are involved in communication, using such cognitive apparatus as a *proposition model* that represents elements, their features, and their interrelationship, a collection of *image schemata* that include generic, geometric, and *spatial prototypes* (such as containers, conduits, connections, part-and-whole, center-and-peripherals, and metaphors projecting among different discourses of universe), and *metonymies* that allow parts to represent the whole.

McNeill (2005) defined *growth point* as a cognitive entity from which coherent verbal and nonverbal expressions are generated. As illustrated in Fig. 2.19, each growth point represents an idea that a cognitive agent has chosen to express from the discourse of internal thoughts (i.e., *catchment*). Once a growth point is generated, two lines of cognitive processes—one linguistic, the other gestural—are instantiated to substantiate the idea in the growth point. Sub-processes invoking the two lines interact with each other in making dialectics, resulting in coordinated communicative behaviors in linguistic and nonlinguistic media.

It is widely believed that we can read other people's minds. Such an ability is called *theory of mind* and is deemed quite essential to live a better social and emotional life. More specifically, we believe that theory of mind is essential to empathic agents. Attempts have been made to better understand theory of mind, addressing what the computational mechanism underlying theory of mind is, how people formulate theory of mind at a young age, and whether primates and species other than homo sapiens have theory of mind. Developmental psychologists devised experiments to investigate at which stage of development infants acquire theory of mind. In the *false belief task* experiment illustrated in Fig. 2.20 (Wimmer and Perner 1983),

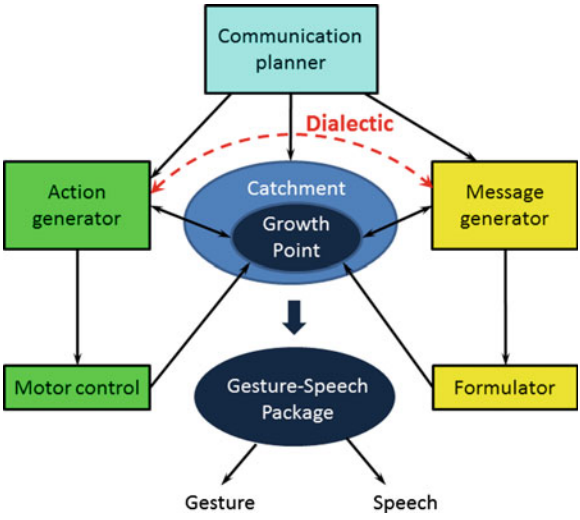


Fig. 2.19 A hypothetical process of generating verbal and nonverbal expressions from growth points. Drawing inspired by McNeill (2005)

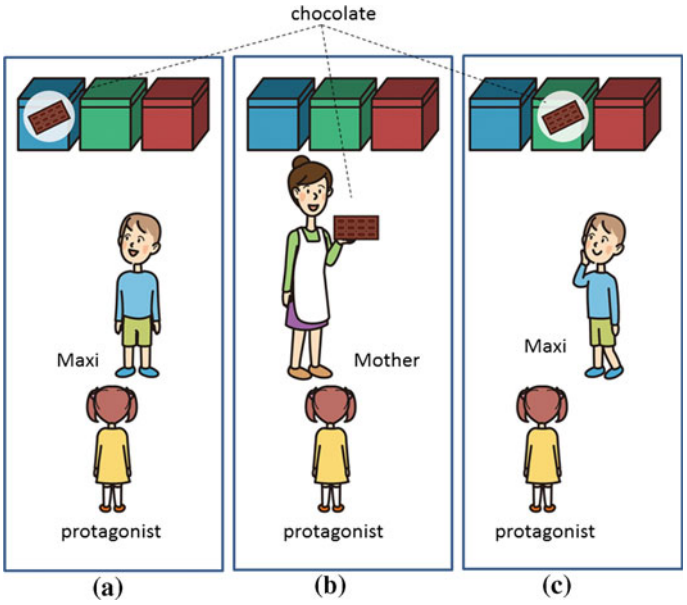


Fig. 2.20 A false-belief task. Drawing inspired by Wimmer and Perner (1983). **a** Scene 1 **b** Scene 2 **c** Scene 3. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

each participant (child) is first shown a picture-based story in which the protagonist witnesses a small boy named Maxi identify that his favorite chocolate is in a blue box (Fig. 2.20a); next, the mother moves the chocolate to a green box while Maxi is absent (Fig. 2.20b); finally, the participant was asked which box Maxi would search in upon his return and whether the protagonist knew where the chocolate was (and even what the protagonist would do if she wanted to eat the chocolate herself). Theory of mind was considered necessary to provide correct answers for the above questions. Wimmer and Perner found that normal infants were not able to answer correctly until they were 4–6 years old.

As illustrated in Fig. 2.21, Leslie (1987) described that an ability to pretend, such as playing with a banana by pretending it is a telephone, an ability infants aged 18–24 months acquire, is deeply related to the development of intelligence; because to pretend, one must not only create a referentially transparent primary representation that can be used to refer to existing objects, but also create a referentially opaque meta-representation for objects not existing in the universe of discourse. The *decoupling model* proposed by Leslie is used to account for pretending.

Premack and Woodruff (1978) conducted an interesting experiment to investigate whether primates other than homo sapiens have theory of mind. Their findings were positive; a chimpanzee named Sarah appeared to understand such scenes as "a human actor struggling to escape from a locked cage," "a heater malfunctioning because the human actor glanced wryly at it, and even kicked at it a little with shivered clasped arms to the chest," "an actor unable to wash a dirty floor because the hose

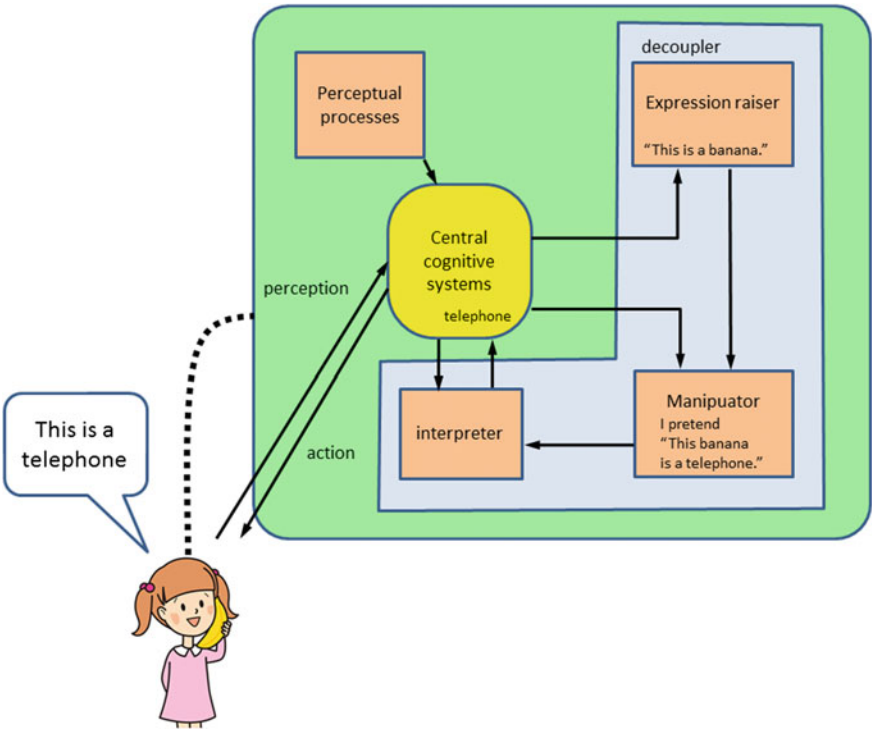


Fig. 2.21 A decoupling model of pretending. Diagram inspired by Leslie (1987). © Toyoaki Nishida and At, Inc. Reproduced with permission

is not properly attached to the faucet," and "an actor seeking to play an unplugged phonograph." Furthermore, Sarah seemed to follow lines of reasoning, such as "the human actor wants the banana and is struggling to reach it."

Detailed in Fig. 2.22, Baron-Cohen (1995) proposed a computational architecture of theory of mind consisting of the following four components: intention detector, eye direction detector, shared attention mechanism, and a theory of mind mechanism that infers the intention of other actors by integrating visual, auditory, and tactile cues.

Apart from rational reasoning, emotion plays an important role in human intelligence and hence in communications (Nishida 2010a). Early research on emotions identified and classified emotions as represented by Ekman (1992). The major concern here was to identify basic emotions that may be combined to comprise non-basic emotions. For example, Ekman proposed happiness, disgust, surprise, sadness, anger, and fear as six basic emotions. Classifications proposed by Plutchik (1980) were an attempt to define non-basic emotions as one may compose a new color by mixing primitive colors.

Mehrabian (1996) proposed a parametric model of emotions called the *PAD* model in which the space of emotion was spanned by the following three axes: *pleasure* (P), *arousal* (A), and *dominance* (D). Depicted in Fig. 2.23, Mehrabian claimed that

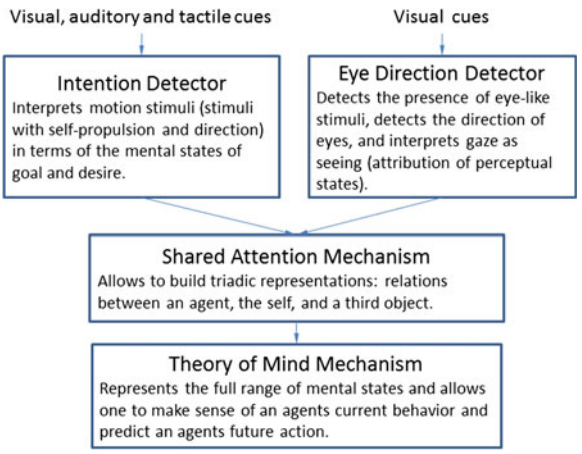


Fig. 2.22 Computational architecture of theory of mind. Diagram inspired by Baron-Cohen (1995)

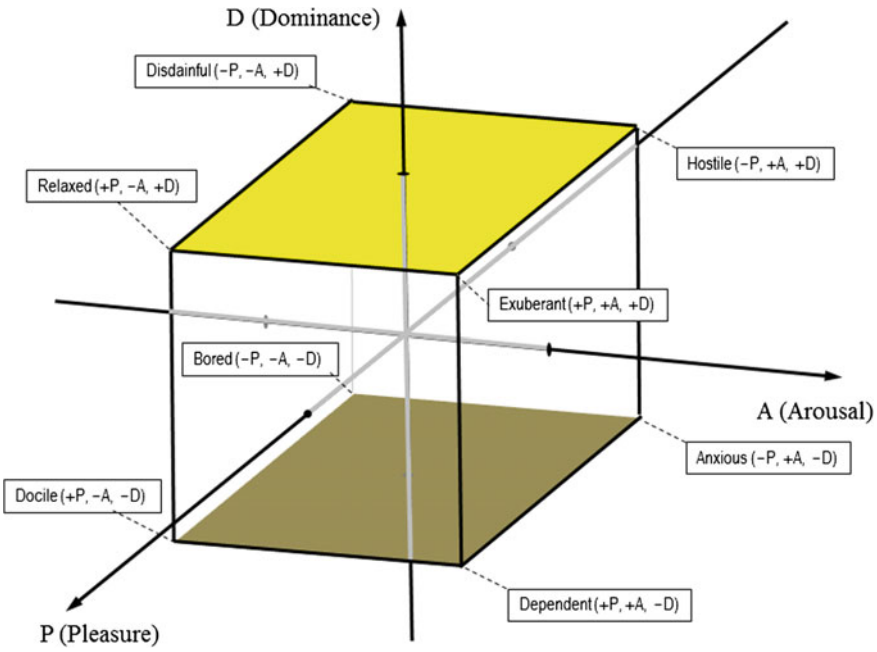


Fig. 2.23 The PAD model. Diagram inspired by Mehrabian (1996). © 2010, Springer. Reproduced with permission

other emotions can be consistently accommodated in this three-dimensional space. Such emotions include, for example, exuberant (P: +, A: +, D: +), hostile (P: −, A: +, D: +), relaxed (P: +, A: −, D: +), dependent (P: +, A: +, D: −), docile (P: +,

Consequences of events	[+]				"pleased"
		Consequences for others	Desirable for other		"happy-for"
			Undesirable for other		"gloating"
		Consequences for self	Prospects relevant		"hope"
				Confirmed	"satisfaction"
				Disconfirmed	"relief"
			Prospects irrelevant		"joy"
				Attributed to self (agent)	"gratification"
				Attributed to others	"gratitude"
	[−]				"displeased"
		Consequences for others	Desirable for other		"resentment"
			Undesirable for other		"pity"
		Consequences for self	Prospects relevant		"fear"
				Confirmed	"fears-confirmed"
				Disconfirmed	"disappointment"
			Prospects irrelevant		"distress"
				Attributed to Self (agent)	"remorse"
				Attributed to others	"anger"
Actions of agents	[+]				"approving"
		Focusing on self agent			"pride"
		Focusing on other agent			"admiration"
	[−]				"disapproving"
		Focusing on self agent			"shame"
		Focusing on other agent			"reproach"
Aspects of Objects	[+]				"liking"
		Attraction			"love"
	[−]				"disliking"
		Attraction			"hate"

Fig. 2.24 Ortony-Clare-Collins (OCC) model (Ortony et al. 1988; Nishida 2010a). © 2010, Springer. Reproduced with permission

A: −, D: −), anxious (P: −, A: +, D: −), disdainful (P: −, A: −, D: +), and bored (P: −, A: −, D: −). The actual model is quantitative, and values are taken from real numbers.

How are these emotions associated with non-emotional mental activities, such as goal-oriented behaviors? Cognitive appraisal theory suggests that emotion arises as a result of evaluating incoming events based on one's mental state. Ortony et al. (1988) presented a comprehensive model (i.e., the *OCC model*) based on this idea, as shown in Fig. 2.24. The OCC model accounts for how emotions are associated with valenced responses to goal-oriented events. For example, if the consequences of an event are positive, the resulting emotion might be designated as "pleased"; otherwise, it is "displeased." A more detailed appraisal may arise by taking into account other aspects, such as occurrences. Whether the focus is the consequences for self, prospects are irrelevant, and the consequences are attributed to other agents. The resulting emotion might be "gratitude" or "anger," depending on whether the consequences are positive or negative.

A deeper model of emotion was proposed in cognitive neuroscience. A model proposed by Damasio (1994) is remarkable in the sense that it distinguished *primary* and *secondary* emotions, as depicted in Fig. 2.25. The former is rather reflective as the signal from the real world is sent directly to the amygdala such that bodily reactions occur quickly to cope with any immediate problems. Meanwhile, the latter

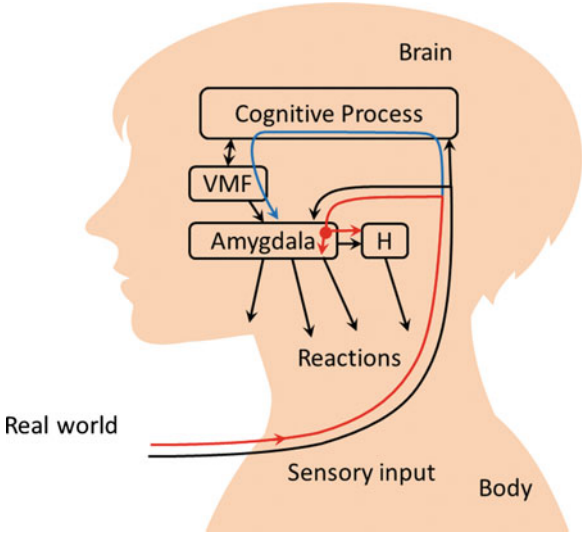


Fig. 2.25 Model of emotion. Diagram inspired by Damasio (1994). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

is deliberative in the sense that the sensory input is sent to cognitive processing for detailed analysis based on past experiences, and the results are sent to the amygdala through the ventromedial prefrontal cortex (VMF) to cause emotional reactions. Although primary and secondary emotions have the same type, the latency is different due to the difference in the signal route. Damasio also proposed a somatic marker hypothesis suggesting how past experiences will be recalled to avoid the same or similar mistakes from the past.

Recent progress in cognitive neuroscience (Rizzolatti and Sinigaglia 2008; Iacoboni 2008) involves the role of imitation with a *mirror neuron system* that, upon seeing actor B perform action X, generates the same signal in actor A as would be produced if actor A performed that same action X. Thus, the mirror system may allow an actor to guess the intention of other people and hence contribute to empathic communication.

2.9 Summary

In this chapter, we have overviewed existing observations and theories for understanding conversation. We employed five viewpoints to overview a vast collection of previous work, namely, verbal communication, nonverbal communication, social discourse, narratives and content flow, and cognitive processes. We started our journey with narratives and content flow. Although this subject includes topics such as

memes, narratology, discursive psychology, and social constructionism, and deemed to be less relevant to conversation than other topics, we believe that storytelling aspects of conversation is quite important to characterize the role of conversation in our social life. Then, we introduced Goffman's seminal work on unfocused and focused interactions in gathering to delineate the outer appearance of conversation. Goffman's characterization of types of participants in conversation and Kendon's F-formation for analyzing spatial arrangement of participants helped us grasp conversation scenes in a structured fashion. Focused gathering is a core of conversation which can be categorized into verbal interactions using symbols and nonverbal interactions using social signals. The former belongs to the domain of language use. Our approach draws on Clark's joint activity theory comprising a rich repertoire of concepts, such as levels, tracks and layers, for formalizing conversation as joint activities. The latter has attracted an attention from scholars with different background from anthropology to social sciences and even to cultural studies. The focus was the identification of social signals and their interpretation. We illustrated the analysis of the turn-taking system in this framework. Finally, we shed light on the mental processes that happen beneath the surface of conversations, to understand conversations more deeply. We looked at Lakoff's schema, theory of mind, various models of emotion, and Damasio's model of primary and secondary emotions in particular.

Chapter 3
History of Conversational System Development

Abstract Conversational system development dates back to the early days of computer science when pioneering researchers started to take up serious projects aimed at having computers interact with people using natural language. Their endeavors have produced a broad range of theories, techniques, and systems, ranging from basic research to applications, from text to multimodal signals, from dialogue to story, and from computational to cognitive. In this chapter, we will present a bird's eye view of these activities and highlight epochs relevant to conversational informatics.

Keywords Natural language dialogue system · Speech dialogue system · Multimodal interface · Embodied conversational agent · Story understanding system · Cognitive computing

3.1 A Bird's Eye View

Before departing for a historic tour of the past 50 years of research and development, it is important to be clear about the questions we ask about it. The questions, as usual, are twofold: (a) on what lines of research can we base our research of conversational systems, and (b) how much has been achieved and what are the limitations of previous research?

Figure 3.1 presents our findings as a panoramic view of the terrain of research activities relevant to the scope of this book. Although many researchers may believe that the mainstream of conversational system development is a path toward embodied conversational agents or intelligent virtual humans, which are defined as synthetic characters that can engage in face-to-face conversations using verbal and nonverbal communication means, we would like to emphasize two other lines of research as closely relevant to conversational system development: one towards storytelling and understanding systems and another motivated by simulating human cognition.

The history of conversational system development starts in the 1960s when attempts to develop text-based natural language dialogue systems began. In that period, pioneering research projects in artificial intelligence succeeded in building the first computer programs that could answer questions typed in English. These groundbreaking projects were undertaken not long after the first commercialization

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_3

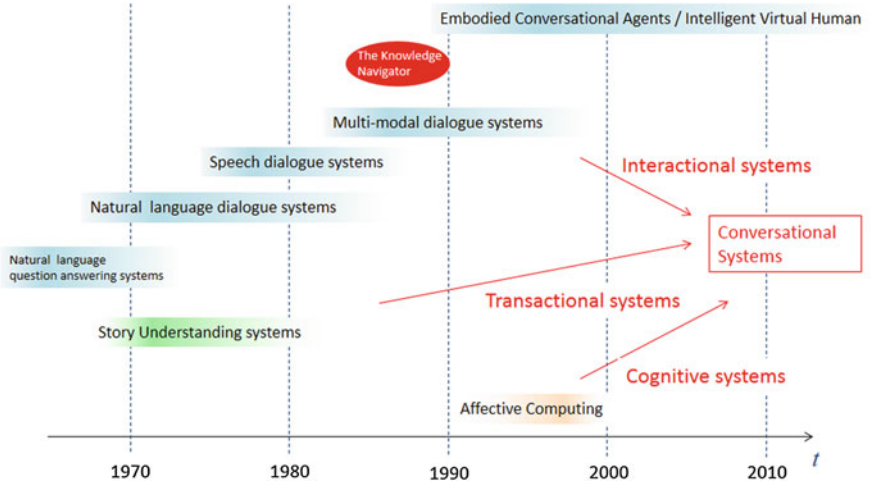


Fig. 3.1 Three threads of research toward conversational system development

of computers and the very beginning of intensive artificial intelligence research, which was announced at the 1956 Dartmouth summer research project on artificial intelligence. Although the first natural language dialogue systems could only handle simple sentences, they were really something for their time. The task required much more than simple statistics: a certain degree of human intelligence was necessary to accomplish such tasks.

After the initial success of natural language question answering systems, a bunch of AI researchers became interested in extending them as *interactional systems* that could pursue *goal-oriented dialogues*; they were looking for better *human-computer interfaces*. Around 1980, speech recognition systems that contributed to a more natural user interface were developed. In the 1990s, these were followed by multimodal dialogue systems that allowed the user to combine speech with nonverbal input, such as gestures, to interact with the system. In 1987, a concept video titled "The Knowledge Navigator" was released by Apple, Inc. This video eloquently illustrated how an artificial intelligence system employed as an embodied conversational agent could help people. Inspired researchers started to build such agents that bore key features of the agent illustrated in The Knowledge Navigator, such as anthropomorphism and verbal-nonverbal interactions with the user in the field of research on embodied conversational agents and intelligent virtual agents (Cassell et al. 2000; Prendinger and Ishizuka 2004; Nishida 2007b).

Some other AI researchers addressed the extension of the early natural language question answering systems in a different direction, i.e., towards generating and understanding narratives and stories. This line of research, which could be called an approach towards *transactional systems*, is critical to address content management of conversational systems, not just telling stories but also learning from conversations, and retaining accumulated conversations in memory.

Since the 1990s, researchers in cognitive science and artificial intelligence became more interested in building a *cognitive system* that may exhibit more human-like intellectual behaviors resulting from autobiographical dynamic memory, emotion and theory of mind. Among others, affective computing has become a visible line of research, since affective computing was proposed by Picard (1997).

After a long period of explorative work in a different vein, we come to witness the confluence, manifesting as a significant overlap of interest and sharing of in-sights and approaches, in addition to maturity of techniques and tools. As a matter of fact, integration is considered necessary for conversational systems to be successful; without proficient interaction, stories may be told poorly, without stories, conversation is boring, and without a cognitive model, we cannot build an attractive conversational agent.

In the subsequent sections of this chapter, we will look at the history of research and development of conversationally intelligent systems in more detail and try to identify major landmarks.

3.2 Early Natural Language Dialogue Systems

The earliest natural language diaglogue systems were text-based natural language question answering systems, such as Baseball (Green et al. 1961), LUNAR (Woods 1973), ELIZA (Weizenbaum 1966), and SHRDLU (Winograd 1972). These systems translated user input into database queries to answer questions. Later, more generic natural language dialogue systems used a dialogue engine. Basic techniques were developed for handling fundamental linguistic constructs, such as syntax, semantics, and discourse.

3.2.1 Baseball

As the name suggests, Baseball is a natural language question answering system built to answer questions about baseball games (Green et al. 1961). Based on a small dictionary and programmed grammar, it produces database retrieval commands to answer information requests in natural language. For example, given a question "Where did the Red Sox play on July 7?" Baseball will generate an internal structure, such as

```
Place = ?
Team  = Red Sox
Month = July
Day   = 7
```

in order to issue a database search. As a result, an output like this:

```
Month = July
Place= Boston
Day = 7
Game Serial No. = 96
(Team = Red Sox, Score= 5)
(Team = Yankees, Score = 3)
```

will be produced, which when roughly translated reads as "It was Boston where Red Sox and Yankees played game #96, resulting in Red Sox defeating the Yan-kees 5-3." It would not have been difficult to produce an English output; however, it was not implemented. No explicit representation of English grammar was used to analyze the syntactic structure of the input. Instead, specialized programs scanned the input to detect and mark syntactic constructs, e.g., brackets were inserted to mark noun phrases for use in the succeeding content analysis phase to produce a database search command. This style of architecture was quite limited and difficult to extend and, consequently, was only applicable to small-scale systems.

3.2.2 LUNAR

The Lunar Natural Sciences Natural Language Information System (LUNAR) is a natural language question answering system that can answer queries about the rock samples brought from the moon (Woods 1973). LUNAR features (a) syntactic analysis using heuristic and semantic information to select the most likely syntactic analysis from candidates, and (b) semantic representation used as an intermediate representation to produce database queries. LUNAR was able to answer complex queries such as

```
Give me all lunar samples with Magnetite
In which samples has Apatite been identified
How many samples contain Titanium
Which rocks do not contain Chromite and Ulvospinel.
```

Augmented Transition Network Grammar (ATNG) was developed to analyze inputs by combining syntactic and semantic analysis (Woods 1970). ATNG is a state transition machine with some extensions. An ATNG transition graph segment for parsing simple English sentences is shown in Fig. 3.2.

For example, the subgraph on the top, specifies that a sentence (S) consists of a noun phrase (NP) followed by either a verb (V) or an auxiliary verb (AUX) and a V, which are further followed by a NP and possibly one or more prepositional phrases (PP). The programmer can associate any program with each edge to make additional semantic checks or to build part of a formal expression. Although ATNG has become a popular tool for natural language analysis and has some advantages in combining syntactic and semantic processing, it is evident that it would be quite difficult to write a large-scale ATNG.

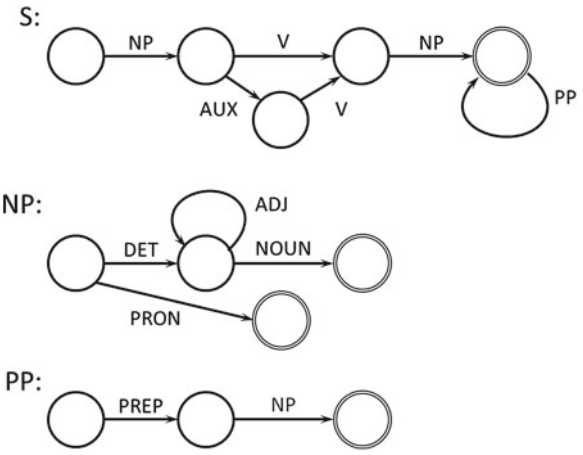


Fig. 3.2 A piece of ATNG for parsing simple English sentences. Diagram inspired by Woods (1970)

3.2.3 SHRDLU

SHRDLU (Winograd 1972) is a program that can interact with the user by understanding the meaning of natural language expressions with respect to the "block world," which is a simple simulated world consisting of simple objects like blocks, pyramids, and boxes (Fig. 3.3).

```
Person:   PICK UP A BIG RED BLOCK.
Computer: OK.
Person:   GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person:   FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE
          HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS
          TALLER THAN THE ONE I AM HOLDING.
Computer: OK.
Person:   WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person:   WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person:   HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
...
```

SHRDLU is remarkable in that it was the first program to prototype the entire process of natural language understanding in computational terms. It successfully demonstrated how various types of components, ranging from natural language processing to a planner for the block world, can be combined to make up a rather simplified virtual agent that behaves as if it understands the meaning of natural language. For example, given the utterance "pick up a big red block," SHRDLU analyzes

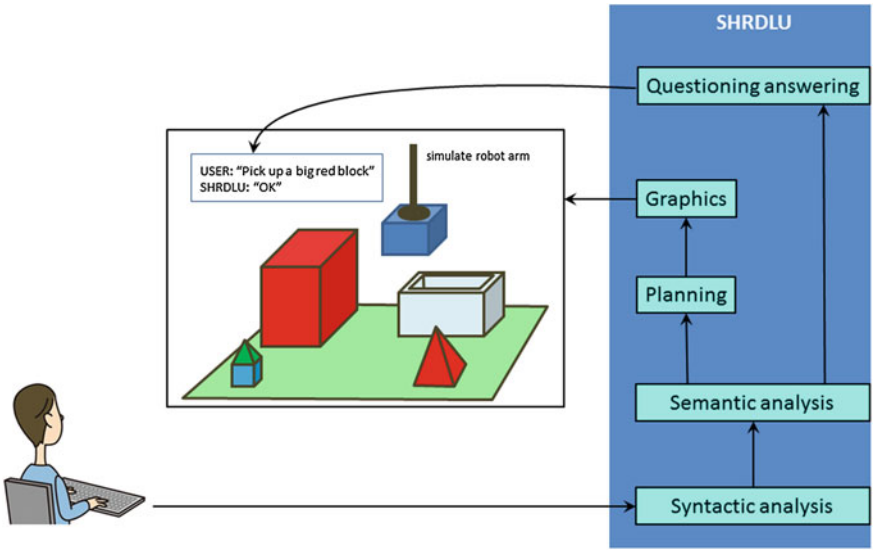


Fig. 3.3 The block world and SHRDLU. Drawing inspired by Winograd (1972). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

its syntactic and semantic structure to determine which physical object is referred to and what goal the user wants it to achieve. It can execute what is said in natural language and can also answer or ask questions when it identifies ambiguity in a given utterance.

3.2.4 ELIZA

ELIZA (Weizenbaum 1966) is very unique in the history of natural language dialogue systems as it does not use any extensive grammar or semantic knowledge to produce surprisingly human-like responses. A typical interaction with ELIZA is as follows:

```
User:   He says I'm depressed much of the time.
ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED
...
User:   It's true. I am unhappy.
ELIZA: DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
```

The techniques employed in ELIZA are (a) identification of keywords, (b) detection of a minimal context, (c) selection of proper transformations, and (d) response generation when no keywords are found in the input. The format of the transformation rules employed in ELIZA is as follows:

$$
\begin{array}{lllll}
(K & ((D_1)) & (R_{1,1})) & (R_{1,2})) & \ldots\ (R_{1,m_1}))\\
 & ((D_1)) & (R_{2,1})) & (R_{2,2})) & \ldots\ (R_{2,m_2}))\\
 & \ldots & & & \\
 & ((D_n)) & (R_{n,1})) & (R_{n,2})) & \ldots\ (R_{n,m_n})))
\end{array}
$$

where K, D_i, and $R_{j,k}$ denote a key word, a decomposition template, and a reassembly rule, respectively. The keyword indexes the transformation rule so that the system can quickly judge whether the transformation rule is applied to the input. The decomposition template is used to specify the detailed conditions for the transformation rule to be applied. For example, the decomposition template D:

```
(0 YOU 0 ME)
```

specifies that there is an indefinite number of words before or after "YOU" followed by "ME," where "0" indicates a segment of an indefinite length. For example, D will match the following sequence:

```
(IT SEEMS THAT YOU HATE ME)
```
"It seems that you hate me",

and decompose into the following collection of components:

the first segment: `(IT SEEMS THAT)`
the second segment: `(YOU)`
the third segment: `(HATE)`
the last segment: `(ME)`.

The reassembly rule specifies how the output will be assembled from the detected components of the input. To avoid repeating the same response for the same input, which would considerably damage the naturalness of the output, the transformation rule allows for more than one reassembly rule to be specified by the programmer. If reassembly rule $R_{i,j}$ is applied in an invocation of the translation rule, the reassembly rule $R_{i,j+1}$ will be used in the next invocation. For example, when the third segment is bound to "HATE," the reassembly rule R:

```
(WHAT MAKES YOU THINK I 3 YOU)
```

will produce the following response:

```
(WHAT MAKES YOU THINK I HATE YOU)
```
"What makes you think I hate you."

ELIZA produces artful back-channel communication realized through string pattern matching that produces not only acknowledgement phrases but also questions that are relevant to the given utterances and that provoke new thoughts on the human side. Although the techniques are only applicable to limited tasks, the work strongly suggests that the user judges the quality of a dialogue system by its responses. Sometimes, a canned joke may make sense to the user; however, it is quite challenging to produce a joke as a result of normal techniques for syntactic, semantic, and discourse processing.

3.3 Speech Dialogue Systems and Multimodal Interfaces

Speech dialogue systems, such as HEARSAY-II (Erman et al. 1980), that started to appear in the 1970s allowed the user to speak to the system. Techniques, such as *blackboard systems*, were developed to overcome difficulties in interpreting continuous noisy signals to elicit maximally plausible interpretation in real time.

The architecture blackboard system architecture appears to be generic; it is used to manage a suite of processes for interpreting low level signals to elicit highly symbolic conceptual representations to create a response. Figure 3.4 shows the blackboard system architecture employed in HEARSAY-II. A blackboard system consists of a shared database called a *blackboard* (the left half) and a collection of cooperating processes (the right half). Information representation of the blackboard may consist of multiple levels of abstraction. HEARSAY-II has seven levels of abstraction, ranging from the parameter level to the phrase level. Each small labeled box on the right represents a programming module, referred to as a knowledge source, which executes a tiny specialized task that is necessary to accomplish interpretation. For example, a knowledge source called SEG takes continuous signal data at the parameter level as input to produce a hypothesis of segmentation at the segments level. Generally, the task of a knowledge source is either hypothesis generation at a higher level of abstraction or its verification at a lower level.

Knowledge sources need to be executed concurrently so that they can contribute to "co-authoring" the most plausible interpretation on the blackboard. At the same time, their execution should be focused at producing responses in real time. Any update

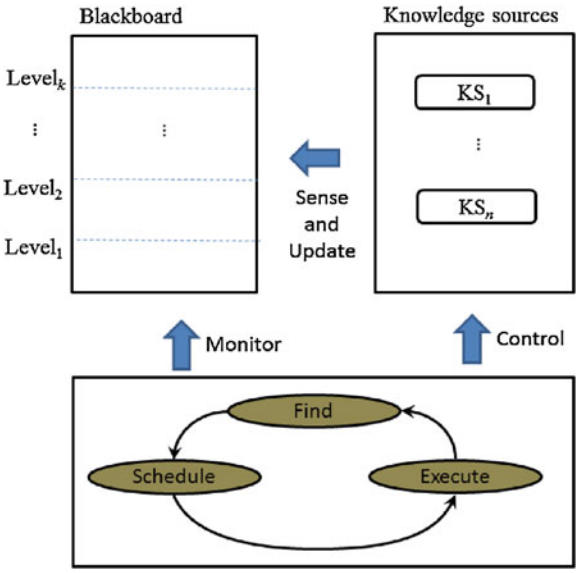


Fig. 3.4 The blackboard system architecture. Drawing inspired by Erman et al. (1980)

on the blackboard is notified to a module called Blackboard Monitor, which will notify the knowledge source relevant to the event. Each notified knowledge source will examine the blackboard and propose an update to the blackboard as a resolution. Then, an agenda is created in which such proposals are prioritized. The scheduler will determine which update is to be made by consulting a focus of attention mechanism.

A series of improvements of the blackboard architecture have been made after HEARSAY-II. In the goal-oriented blackboard architecture, the blackboard monitor is replaced by the goal blackboard and goal processor, which are dedicated to goal-oriented processing. In BB1, the agenda-based control structure is replaced by the control blackboard and control knowledge sources so that a knowledge source in Control Knowledge Sources can control the execution of knowledge sources explicitly (Hayes-Roth 1985) . This allows for distinguishing running knowledge sources and those that can run. The scheduler will assign a priority based on the heuristics associated with active focuses in the control blackboard.

In the 1980s, speech dialogue systems were extended to multimodal interfaces. Thus, it became possible to use more than one modality of communication in human–computer interaction. Put-That-There (Bolt 1980) is pioneering work in this direction. It allowed the user to arrange simple shapes on a large graphics display surface by voice and a simultaneous pointing gesture (Fig. 3.5). It manipulates graphical objects according to typical input from the user as follows:

```
Create a blue square there.
Move the blue triangle to the right of the green square
Move that to the right of the green square.    (with pointing)
Put that there    (indicated by gesture)
Make that smaller    (with pointing gesture)
Make that    (indicating some item)    like that    (indicating some other item)
Delete that    (pointing to some item).
```

Fig. 3.5 The concept of Put-That-There. Drawing inspired by Bolt (1980). © 2014, At, Inc. Reproduced with permission

The user can use pronouns, and the pronunciations need not be correctly recognized as long as the accompanying pointing gesture provides sufficient information. On the other hand, the user's pointing gestures did not to be precise because simultaneous voice was also used to gain precision.

The key issue in multimodal interaction is fusion. Robust and efficient algorithms are needed to interpret multimodal input from heterogeneous sensors by handling errors and missing data. To produce multimodal output, the components of the output from different knowledge sources are tailored and synchronized. It involves content selection and organization, coordinated distribution of information on available modalities, modality-specific content realization, and laying out generation results (André and Rist 1995). The presentation task may be characterized as a knowledge-based, goal-directed activity under constraints. Planning algorithm is needed to assemble the components into a coherent structure of presentation. Complexity may arise because selections made in each component need to be aligned so that the resulting presentation as a whole may serve as a goal. A high level topic structure, such as dialogue acts, may also be required to integrate components (Stein and Thiel 1993).

3.4 Embodied Conversational Agents and Intelligent Virtual Humans

Eventually, we witnessed anthropomorphic agents that have a visual human-like embodiment. Interestingly, a concept video drove this research and development. The Knowledge Navigator video based on Sculley and Byrne (1987) released by Apple, Inc. in 1987 gave a clear image of multimodal conversational interaction mediated by an embodied conversational agent.

The agent, Phil, had a life-like appearance. Phil was modeled as an animated talking head embedded in a multimodal interaction environment (Fig. 3.6). Phil had a unique appearance as a male butler; therefore, the user could identify him as an individual character. Phil knew with whom "he" was talking; consequently, "he" could provide the user with a customized service based on the user's personal profile and discourse of interactions. Phil was socialized in the sense that "he" had knowledge of social events and relationships. For example, Phil was able to sense interactions among other participants and not interrupt their conversations, and "he" could tailor information for a social context. Phil symbolically embodies a personalized social agent that can interact with the user using multimodal communication.

It should be noted that there has been a debate on the delegation and anthropomorphism illustrated in The Knowledge Navigator. Is it really a good idea to employ an indirect manipulation metaphor, i.e., to delegate tasks to artificial agents? Should artificial agents look human? On one hand, these two features are considered useful as a metaphor in building user interfaces in the long run, as delegation hides complexity from the user to benefit from the service without having to be licensed to

Fig. 3.6 The concept of the Knowledge Navigator. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

manipulate the tool. Additionally, anthropomorphism can allow the users to apply proficient inter-human communication skills, such as those employed in daily conversation. On the other hand, incomplete or improper implementation of delegation and anthropomorphism may make the user interface less useful or even introduce confusion. The appearance of an agent should be balanced with underlying intelligence, and we can delegate tasks to the agents so long as the benefits and risks are clear and acceptable to the user. Further research is required to clarify exactly how much technical maturity is required for delegation and anthropomorphism to be employed successfully in a user interface.

Peedy, the Conversational Personal Assistant, (Ball et al. 1997) is one of the earliest realizations of the key concepts introduced in The Knowledge Navigator video. Peedy is a conversational system in which a parrot-agent named Peedy helps the user select songs from a collection of audio CDs (Fig. 3.7). Peedy integrates speech I/O, a natural language dialogue engine, and multimodal output to realize interactive give-and-take, recognition and management of the costs of interaction and delay, proper handling of interruptions, and means to deal with emotional and social aspects of interaction. A typical dialogue between the user and Peedy is as follows:

```
…
User: Play some rock after that.
(Peedy scans the notes again, selects one)
Peedy: How about "Fools in Love"?
User: Who wrote that?
(Peedy cups one wing to his 'ear')
Peedy: Huh?
User: Who wrote that?
(Peedy looks up, scrunches his brow)
Peedy: Joe Jackson
…
```

Peedy was designed to realize an assistive interface that is aware of the social natures of interaction. First, it supports interactive give-and-take, in the sense that it not only responds to questions but also asks questions. Second, it recognizes the cost

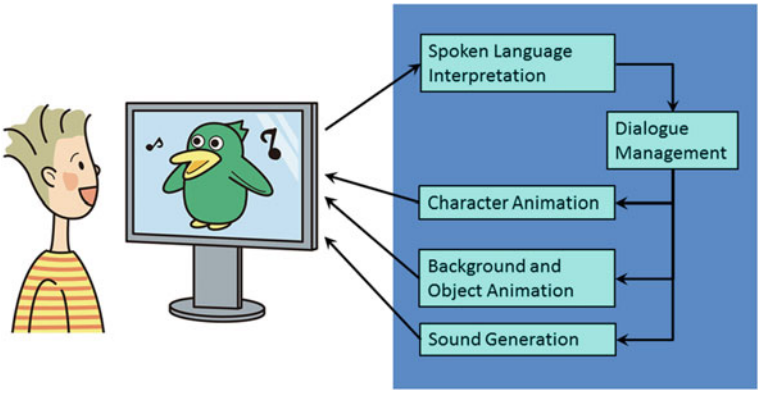


Fig. 3.7 Peedy. Drawing inspired by Ball et al. (1997). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

of interaction and delay. It knows, for example, that requiring the user's confirmation may take time and introduce delay while carrying out a task. Third, it attempts to manage interruptions effectively by delaying the initiation of interaction with the user while they are occupied by a task, e.g., a phone call. Finally, it acknowledges the social and emotional aspects of interaction by sensing the social situation and emotional status of the user.

To achieve high quality and high performance interactions with limited technology, integration is emphasized and task-oriented techniques are preferred over generic techniques. A name database is used to handle proper name substitution. Task-oriented template matching is used for semantic analysis, and spoken commands are limited to approximately 150 "typical" utterances, which might be encountered in the CD audio application, that are paraphrases of one of 17 canonical requests. The state transition model is employed, consisting of five conversational states and 17 input events, which comprise approximately 100 distinct transitions. The Peedy modules are grouped into three subsystems that are dedicated to spoken language processing, dialogue management, and video and audio output. The modules are connected in a rather straightforward fashion.

Hayes-Roth launched the Virtual Theater project, which employs *improvisational interaction* of animated smart puppets (Hayes-Roth and Gent 1997). In the improvisational story-making proposed as collaboration among children and smart puppets, the smart puppets ask the children to choose a high-level direction for a given situation. When the children give a choice, the smart puppets improvise a joint course of behavior. Each smart puppet is controlled by an agent consisting of a "body," "mind," and a "mind-body interface." The "body" is a computer program that performs graphical manifestation in a virtual world. The "mind" integrates perceptual inputs with knowledge and inferences to judge the agent's situation in the virtual world. It instantiates and decides when to execute individual behaviors, and it also performs processing to intervene between situation assessment and behavior. The

"mind-body interface" coordinates interactions between its mind and body in a control loop.

Jennifer James is a conversational agent that simulates a virtual auto salesperson for a fictional auto company. As a virtual character, Jennifer James can engage in a free dialogue with the user by displaying vehicles, opening hoods, etc (Fig. 3.8). The personality of Jennifer James can be shaped by a back story given to the user and adapt to the user based on the information acquired during the conversation session (Hayes-Roth 1998).

Rationale for agents can be found in social psychology. Nass et al. (1994) proposed the notion of "Computers As Social Actors (CASA)," claiming that computers, or those interacting with people through an interface, retain social norms, such as politeness and even gender, when they are used in a social context. Even though people know that computers are media and not real human agents, people tend to presume or pretend that they are interacting with real humans. This theory is now known as the *media equation*.

Rea is an embodied conversational agent that can effectively use a human-like body for verbal and nonverbal communication, as shown in Fig. 3.9. Rea is significant in that the system actually implements conversational interactions, in contrast to employing conversation as a metaphor of interface. Cassell et al. (1999) focused on effective use of nonverbal communication media as social signals, such as gaze or hand gestures, to achieve communication functions, such as acknowledgment of the communication partner or taking turns. Cassell implemented the ideas in the real-estate domain, where Rea was characterized as a real-estate agent who shows users the features of various houses.

Rea was able to follow interactions where proper management of turn-taking was required. For example, when Rea recognizes the arrival of a new user, "she" will stop idling behaviors and turn to face the user. During the conversation session, she can follow the interaction sequence as follows:

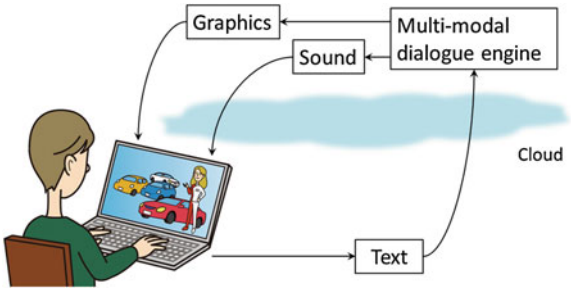


Fig. 3.8 Jennifer James. Drawing inspired by Hayes-Roth (1998). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

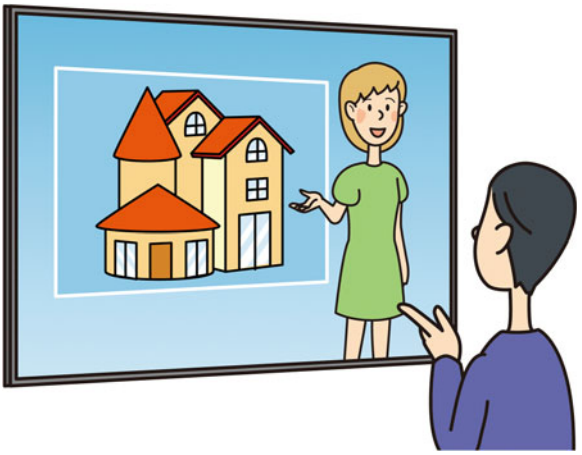


Fig. 3.9 Rea. Drawing inspired by Cassell et al. (1999). © 2014, At, Inc. Reproduced with permission

```
Rea:  It is a large Victorian.
      It has four bedrooms and two baths.
Tim:  (hand gesture)
Rea:  (interpret it as the user's wanting the turn)
      (concludes the utterance and yields the turn)
Tim:  Tell me more about the house.
...
```

Full symmetry between input and output modalities were intended to allow Rea to participate in human-computer conversation with equal footing. KQML was employed at the system level so that heterogeneous modules could interact with each other without constraints by exchanging messages in a common representation language.

After the success of Rea, we can find progress in the methodological aspects. The first advancement was script/markup languages for specifying the behavior of embodied conversational agents. Although markup languages such as AIML (Wallace 2003) had been proposed, they were for specifying text-based conversational systems. Early script languages, such as Microsoft Agent, CPL, and Multimodal Presentation Markup Language (MPML), were defined to specify the behaviors of conversational agents. Their interpreters allowed the scripts to be realized as character animations andutterances. Unfortunately, those early systems did not allow the programmer to refer to the body parts to specify the behavior of an animated character. Later script languages, such as STEP and BML, allowed for parametric representations of body motions. A tool like BEAT allows nonverbal behaviors to be generated from text data by identifying phrases that typically accompany a certain class of gestures, such as a beat.

The second type of progress can be referred to as interaction from observation, or corpus-based generation of behaviors in which the agent's behavior is generated in

two phases. In the first phase, one or more corpus is built by collecting and annotating a sufficient amount of observational data for the target communicative behaviors. In the second phase, parameter values are determined based on the analysis of the data accumulated in the corpora. We will discuss these concepts in more detail in the next chapter.

3.5 Story Understanding/Generation Systems

During the late 1970s to 1980s, intensive research was conducted by Schank and his students at Yale University (Schank and Abelson 1977; Schank and Riesbeck 1981; Schank 1990). In the beginning, they concentrated on using semantic representation, called *conceptual dependency*, to describe events and relations expressed by sentences without being significantly affected by superficial differences of linguistic expressions, such as active and passive mood, and subtle differences of vocabulary. For example, "eat" and "drink" were paraphrased as "ingest food" and "ingest liquid," respectively. They used conceptual dependency to answer questions, paraphrase given sentences, or generate stories.

After a while, they realized that a sufficient amount of knowledge is required to understand stories. As a result, a knowledge representation scheme called *script* was introduced to represent knowledge about stereotypical scenes, such as typical event sequences at a restaurant, used to build a system that required substantial background knowledge. Scripts were also used in a system that could skim stories. Plan was another knowledge representation scheme used to build a system that could understand less stereotypical, more goal oriented stories, such as news stories about international politics.

After building a series of knowledge-based story understanding systems, they started to address learning from stories. From the theory of *dynamic memory*, which results from intensive discussions on how people organize their memory from new stories, remembering similar stories and indexing new knowledge play an important role in the development of conversational systems. Memory oriented packages and theme oriented packages were hypothesized as memory organization schema (Schank 1982).

3.6 Cognitive Computing

Conversational agents need to be supported by a solid mechanism of underlying cognitive computing for lifelikeness, emotion, intentionality and agency, in order to gain an enough sense of presence.

Bates (1994) pointed out the importance of *believability* and *lifelikeness* in building interactive agents. He defined believable agents as those that can provide the illusion of life and permit the audience's *suspension of disbelief*. In other words, a

believable agent should allow the user to think "oh, this is like somebody I know," while keeping the user from being suspicious about their illusion to come back to themselves. Believability is critical to engage the user in aesthetic context, such as theater, film, drama, etc. In the Oz project, Bates addressed interactive drama as dramatically interesting virtual worlds inhabited by interactive characters within which the user experiences a story from a first person perspective. Personality, emotion, self-motivation, change, social relationship, and the illusion of life are identified as key concepts in realizing interactive drama.

The Artificial Life Interactive Video Environment (ALIVE) system allows the user to engage in entertainment interaction with an animal-like agent (Fig. 3.10) (Maes et al. 1995). A wireless, full-body interaction was realized to create natural and believable inter-actions. The "magic mirror" approach was employed to enable the user explicitly to grasp the context of interaction by seeing a representation of him/herself and his/her relationship to other objects in the world.

Silas The Dog is an autonomous pet agent (Blumberg 1997) used in the ALIVE system (Fig. 3.11). It was inspired by ethology to realize illusion of life, or lifelikeness. Behavior centered architecture was employed to creature-like motivation and intentions. The releasing mechanism for detecting stimuli meaningful to the creature agent and the lateral inhibition mechanism were also used to control the behavior.

Emotion plays a critical role in communication and decision making. An emotionless machine intelligence might be vulnerable; it may not recognize a danger that is not logically deducible, which even a young child could easily identify. In communication, the ability to express and interpret emotion is mandatory for an intelligent agent to engage empathically. Picard (1997) pointed out that emotional intelligence consists in the core of human intelligence, argued that computational intelligence needs the ability to recognize and express emotions in order to be a genuinely intelligent partner. Affective Reasoner (Elliott 1992) used the OCC model to realize emotional natural language dialogues as follows:

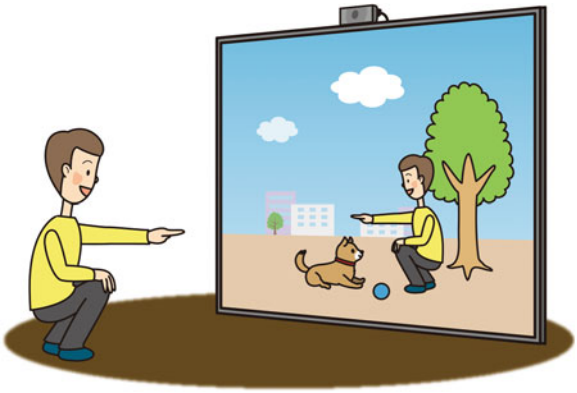


Fig. 3.10 ALIVE. Drawing inspired by Maes et al. (1995). © 2014, At, Inc. Reproduced with permission

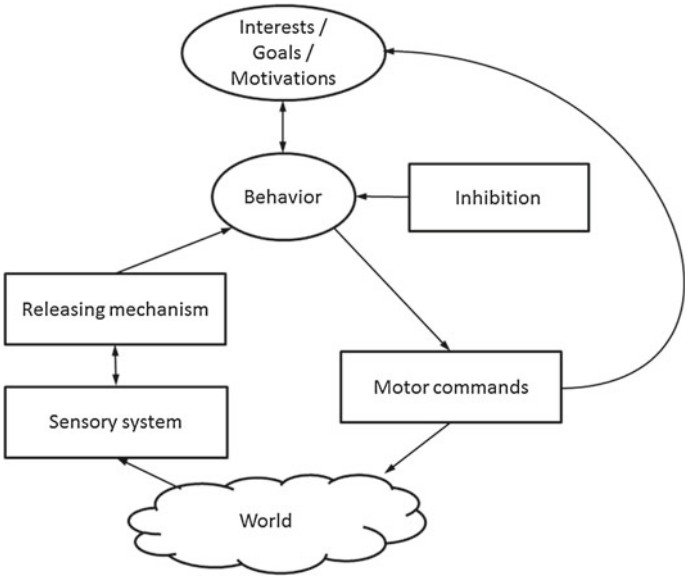


Fig. 3.11 The architecture employed to build Silas T. Dog. Diagram inspired by Blumberg (1997)

(The agent has previously learned that the word "Ruth" is a female friend of the user)
```
User:  Ruth is unhappy
Agent: I am sorry to hear that. You must feel sorry for her.
```
(days later)
```
Agent: How is Ruth?
User:  Ruth is jealous of her rival.
Agent: How much is Ruth feeling this resentment?
User:  Very much.
Agent: Perhaps you are worried about Ruth?
```

To respond to the first user utterance, the utterance of the agent (Affective Reasoner) was based on the inference that the user should feel sorry (or displeased about an undesirable event) for "Ruth" because the user and "Ruth" were friends and the user was told that "Ruth" is in the "unhappy" state.

To respond to the second user utterance, the agent inferred that "Ruth" was in a very negative state, based on the definition of jealousy as resentment about a desired mutually exclusive goal, and thought that the user was worried about "Ruth." Affective Reasoner is based on the OCC model. Affective Reasoner used the following features to extract emotion by eliciting condition relations from the user's utterances to produce emotional reactions:

- Self: the agent who possesses the emotional state
- Other: the agent who is the target of emotion of the self agent
- Desire-self: indicates whether the given situation is desirable for the self agent
- Desire-other: indicates whether the given situation is desirable for the other agent

Fig. 3.12 FearNot! is an anti-bullying. Drawing inspired by Aylett et al. (2005). © 2014, At, Inc. Reproduced with permission

- Pleased: indicates whether the emotion of the other agent is pleasant for the self agent
- Status: indicates whether the situation is as expected by the self agent
- Evaluation: indicates whether the situation is praiseworthy or blameworthy for the self agent
- Responsible Agent: the agent that the self agent holds responsible for a perceived praiseworthy or blameworthy act
- Appealingness: the self agent's assessment of the eliciting situation as containing an attractive or repulsive object. (Elliott 1992, p. 49)

FearNot! (Fun with Empathic Agents to Reach Novel Outcomes in Teaching) is a virtual learning environment (VLE) for helping school children learn a bullying scenario, by allowing them to explore what happens in bullying (Aylett et al. 2005) (Fig. 3.12). FearNot! integrates an appraisal-based emotion engine and a coping mechanism for handling the given problem. Emotions not only influence the agents' reactive behavior, but also guide the planning process. The child user watches the interaction and is asked to give an advice for the victim about how to deal with the situation (Aylett and Paiva 2012).

3.7 Towards Synergy

As we have seen so far in this chapter, it is evident that the three lines of research, i.e., interactional, transactional and cognitive, have already built their own basis landmarked by numerous monuments. At the same time, the maturity of the three veins manifested individual limitations.

The interactional approach originated from natural language question answer-ing systems has evolved into speech dialogue systems, multimodal dialogue systems, and finally embodied conversational agents or intelligent virtual agents. It allows for a computational intelligence to use synthetic characters as an interface to communicate with the user via verbal and nonverbal communication means. The agent-based

interface not only successfully provides the user with a "natural interface" that minimizes the cognitive overhead for consuming computational services, but also raises novel problems represented by questions and complaints such as "who are they?" and "they are boring".

In contrast, the transactional approach allows for analysis and synthesis of discourse and story that underlie conversation. On the one hand, it provides with methods for analyzing co-text references and synthesizing coherent stories with contextual expressions. On the other hand, it accounts for how stories are interpreted based on knowledge and memory is reorganized as a result of reading new stories. It also permits story telling programs to generate meaningful stories. It is evident that the transactional approach is complementary to the interactional approach in the sense that the former is about what to tell while the latter is how to tell.

However, neither of the interactional or transactional approach is concerned with the actor or agent who participates in conversation to produce or consume stories. Without the mental models for actors and agents, we cannot discuss how stories are produced as a result of experience or thought, or how stories make sense to intelligent agents. To put it another way, the whole life cycle of stories shared in conversations is out of scope unless we take the cognitive approach.

Although it appears evident that integration of the interactional, transactional and cognitive approaches is mandatory to make a breakthrough in conversational systems, it has not been clear at all exactly how we integrate the three approaches, for the more modules you incorporate into a system, the harder it becomes to control them both effectively and flexibly.

We take a data-intensive approach to explore a loose integration of the three approaches. It is expected that a data-intensive approach alleviates the complexity of control, by ascribing interaction and dependency complexity to abundance in data made available by measuring conversational behaviors. Cognitive aspects may be at least partly incorporated by taking physiological indices into account to estimate mental state, as we will show in Chaps. 5–11.

3.8 Summary

The history of conversational system development starts in the 1960s when attempts to develop text-based natural language question answering systems began. Soon after it, the line of research was split into two: one directed towards development of interactional systems and the other towards transactional systems. The former was looking for better human–computer interface that aims at providing natural user interfaces. Researchers on this line have attempted to introduce more modalities that make conversation more natural. Speech dialogue systems, multimodal interfaces and embodied conversational agents or intelligent virtual agents have been developed in this vein. Numerous epoch-making systems have been developed, including, among others, LUNAR, SHRDLU, ELIZA, HEARSAY-II, Put-That-There, Peedy, and Rea as well as generic techniques such as the blackboard systems, dialogue engines, and

scripts/markup languages. The Knowledge Navigator had a significant impact on sharing the concept of embodied conversational agents that provide personalized social service using multimodal interactions. After the initial implementation of the ideas introduced in The Knowledge Navigator video, methodological progress was made in scripts, markup languages, and corpus-based behavior generation. In the latter line of research on transactional systems, story understanding and generation shed light on the content. Knowledge based methods and the dynamic memory model were proposed. Third line of research has become active since the 1990s. Researches on this line believed that conversational agents need to be supported by a solid mechanism of underlying cognitive computing for lifelikeness, emotion, intentionality and agency, in order to gain an enough sense of presence. Mile stones include Affective Reasoner, ALIVE, and FearNot!

Chapter 4
Methodologies for Conversational System Development

Abstract When we build a conversational system, it is necessary to understand that a full-fledged system may become fairly complex if we are to address all the issues related to uncertainty and noise, coherency and consistency in strong time constraints, and a wide spectrum of phenomena across multiple levels of hierarchies. A strong methodological approach is necessary. In this chapter, we discuss a technical basis from a past research regarding architecture, scripting and markup languages, corpus-based approaches, and evaluations.

Keywords The architecture of conversational systems · Script language · Markup language · Corpus-based approach · Motif discovery · Constrained motif discovery · Evaluation of conversational systems

4.1 Introduction

It is mandatory to have a good grasp of the entire space of issues before you depart for a journey to develop a conversational system. A strong methodology needs to be employed to address the broad range of issues, as shown in Fig. 4.1.

The central consideration is the need to address the computational aspects of autonomous conversational agents that involve interpretations and generation of multimodal signals, dialogue management, affective computing, and theory of mind based on the design of internal representation, which in turn is based on knowledge and models of conversation. We also need to make decisions about designing the environment in which autonomous conversational agents are working with the users or their avatars. The language for describing the model of conversation to be employed by the conversational agents is critical to the development process. In addition, we must determine how to develop the system by considering model building and content management. Evaluation issues must be addressed in the research.

In the rest of this chapter, we discuss architecture, scripts and markup languages, corpus-based development of conversational agents, and evaluation.

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_4

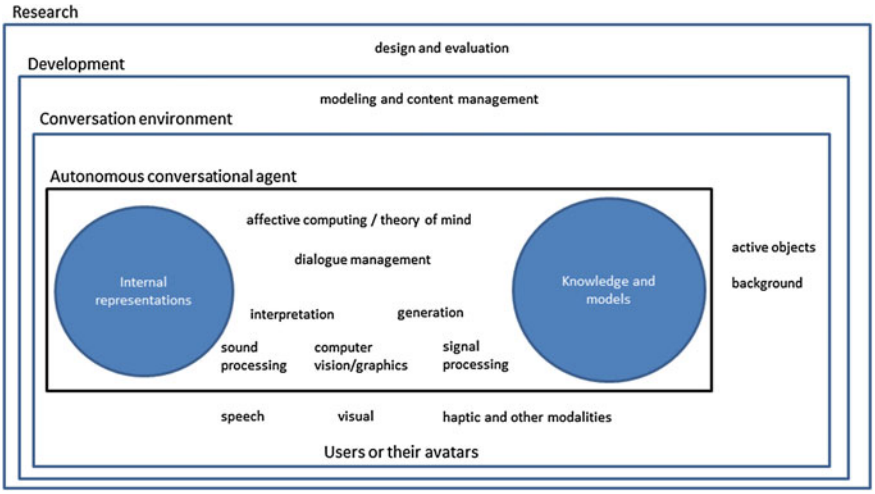


Fig. 4.1 The space of issues

4.2 Architecture

What kinds of components are necessary to build an autonomous conversational agent? The coarsest answer to this question might be to view the conversational agent as machinery for story understanding and generation as in Fig. 4.2. At this level of abstraction, an agent may have episodic and semantic memory. The *episodic memory* holds a collection of stories that the agent has created by itself or has been provided by other agents. The *semantic memory* stores knowledge that may be used to interpret incoming stories or produce stories. An advanced agent may have a dynamic memory component that can reorganize episodic memory. A learner module may update semantic memory by generalizing the lessons learned from experiences.

However, several questions arise. Where do stories come from? How should memory be reorganized? How should stories be generated? In the primordial soup of conversations, even a simple implementation of conversational agents may work. It may be interesting to implement a simple filter so that the conversational agents can choose and store only a given class of stories. We implemented this as the Public Opinion Channel (Nishida et al. 1999). Alternatively, one may introduce a story summarization engine, such as FRUMP (de Jong 1977), to produce summaries. As demonstrated by ELIZA and many chat/twitter bots, additional algorithms may make the behaviors of story understanders and generators interesting, without deep understanding of stories; however, they depend completely on the intelligence of the humans who interpret them.

We have a lot to learn from cognitive science regarding how knowledge and memory is used to understand and generate stories. Schank (1990) pointed out that indexing and reminding are critical for dynamic memory organization, which enables

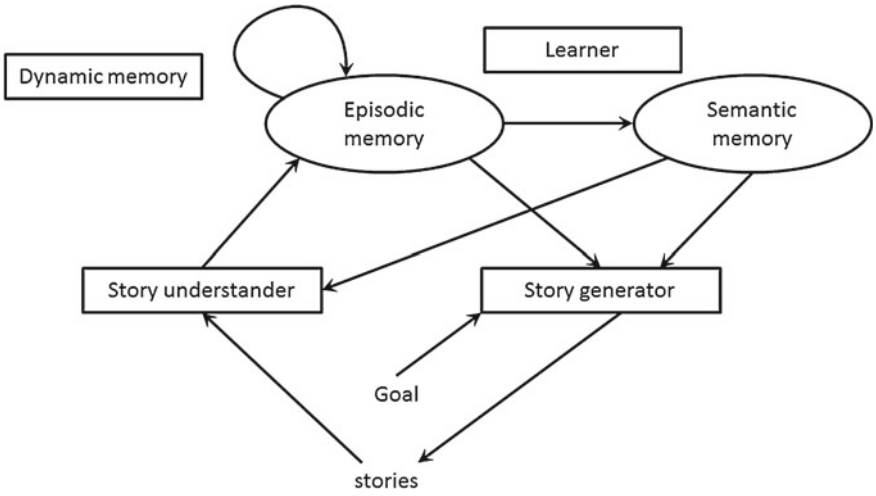


Fig. 4.2 Story understanding/generation systems

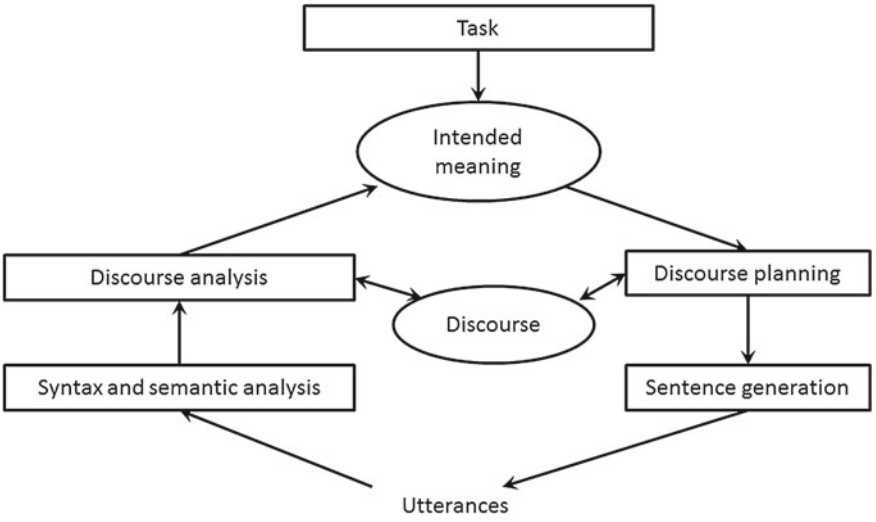


Fig. 4.3 Dialogue engine

the story understander to learn from experiences. Yet the implementation of such ideas may not succeed without substantial use of knowledge and natural language processing techniques, as is illustrated in Fig. 4.3.

A key component of a dialogue engine is the representation of discourse. This describes the information structure across the individual utterances involved in the discourse, specifying how topics are introduced and referred to therein. In addition, the representation of discourse identifies an intended social action. For

example, in a discourse where the availability of somebody is asked, a response "she has left" should be taken as a negative but informative answer.

Several technical difficulties may arise when building a dialogue engine. First, an utterance may be fairly ambiguous if it is simply analyzed in terms of syntax alone even though the most plausible interpretation may be obvious to people with sufficient knowledge. People are flexible in these respects and may become frustrated unless the conversational agent is equally competent. Second, logical representation may not embody subtle differences that are evident to people, such as word order, pause, pitch, or other nonverbal communication. Third, a natural language dialogue engine should be able to coordinate with components that handle other modalities to resolve ambiguities, deictic references, incorporate additional information, etc., which will complicate the control structure of the system. Fourth, building dictionaries and grammars are usually quite expensive because there are generally no clear definitions and boundaries for the system input. In particular, it might be quite expensive to develop a semantic division of a dictionary manually, and training data for associating natural language expression and internal representation pairs are rarely available.

To overcome some of these difficulties, one may employ a similarity-based method based on latent semantic analysis to build a dialogue system without explicit semantic representation. Such a system may allow one to take a data-intensive approach in building dialogue systems.

Besides natural language processing issues, conversational agents must engage in social interactions. In the multiagent research community, building autonomous agents that can participate in social interaction is a common goal. The *belief-desire-intention* (BDI) architecture has been proposed to realize rational agents in social contexts under real-time constraints. A BDI agent takes incoming signals by sensors as input and produces outgoing effects by effectors. The key issue here is abstract machinery that sets up and executes plans to produce rational social behaviors based on a mental model. The BDI model, as shown in Fig. 4.4, includes mental databases containing beliefs, desires, and intentions as basic elements of the mental model (Georgeff and Ingrand 1990), where a belief denotes a proposition the agent assumes to hold in the real world, a desire is a goal that the agent attempts to satisfy, and an intention is a sequence of actions the agent is committed to follow.

Affective computing must be incorporated if we are to realize life-like agents whose behaviors may be induced by more biology-inspired principles. Rosalin Picard took the models of emotion and proposed affective computing as a research field that addresses "computing that relates to, arises from, or deliberately influences emotions" (Picard 1997, p. 3). A typical example of affective computing is an affective tutor who can change the teaching mode according to the status of the learner.

Emotion can be simulated at various levels of abstraction. The simplest model might be to employ cognitive appraisal theories, such as the OCC model. To realize a further sophisticated mechanism for emotions that comply with Damasio's arguments, it is reasonable to have a model that consists of two levels, as is shown in Fig. 4.5. The lower level is for primary emotions comprising quick responses required to respond instantly to urgent problems, and the higher level is for secondary

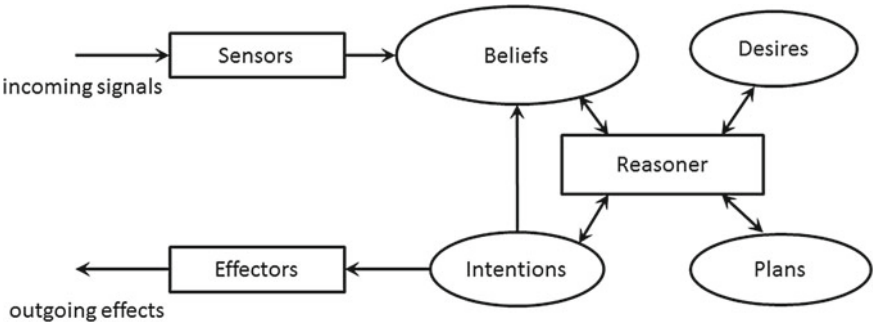


Fig. 4.4 BDI architecture. Diagram inspired by Georgeff and Ingrand (1990)

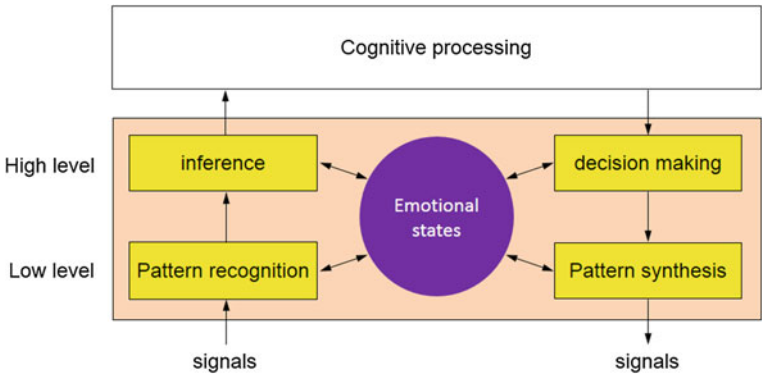


Fig. 4.5 An architecture for affective computing. Diagram inspired by Picard (1997)

emotions comprising slower deliberate processes regarding the purpose, expectation, or preferences.

Becker-Asano (2008) elaborated on this idea and proposed the WASABI architecture based on Mehrabian's PAD model, which consists of an emotion and a cognition module. The emotion module determines the PAD dynamics based on the valence and dominance values sent from the cognition module. The result is expressed as values of two variables representing the mood and awareness likelihood and is sent to the cognitive module. The cognitive module determines the behavior of the agent based on conscious or unconscious cognitive valences resulting from an evaluation of perception from the environment, social discourse, and the time series of previous activities.

Here, we will look at the FAtiMA model in detail. This model is a generic architecture that allows virtual characters to interact in a given setting by combining cognitive appraisal as well as reactive and planned coping behaviors. FAtiMA-PSI extends FAtiMA with PSI, which integrates cognition, emotion, and motivation for human action regulation and links to planning. The PSI component models biological aspects of an agent, including needs, i.e., survival needs, species-preserving

needs, need for affiliation, need for certainty, and need for competence. The PSI component also models motivational system or drives, including energy, integrity, affiliation, certainty, and competence (Lim et al. 2012). FAtiMA-PSI includes a mechanism to model other agents and their relationship to the individual agent. This mechanism can build and update a record of the motivational state of other agents according to perceived events. FAtiMA-PSI is used to drive ORIENT, which is an intelligent graphical-character-based system designed to enhance intercultural empathy. Symbols, rituals, and appraisals are used to represent cultural aspects in FAtiMA-PSI.

The above-mentioned methods do not appear significantly disparate. In fact, they may be collectively represented in a single diagram (Fig. 4.6). As observed, the entire system may consist of multiple levels, signal levels, and cognitive levels in particular. A shortcut might be needed at the signal level to cope with issues that should be handled by low-level but fast processing. The shortcut reaction may be monitored and eventually replaced by a slower but more deliberate process. Episodic and long-term memory must supplement each other, and a learning mechanism may be necessary to reorganize the episodic memory by considering the novel inputs.

At the platform level, the architecture of the system should allow the developer to handle complexity of the phenomena of the target conversation.

Common desiderata may include the following.

- The system should be able to run multiple processes simultaneously to handle multimodal signals.
- The system should allow the programmer to write a complex control structure without sacrificing real-time response.

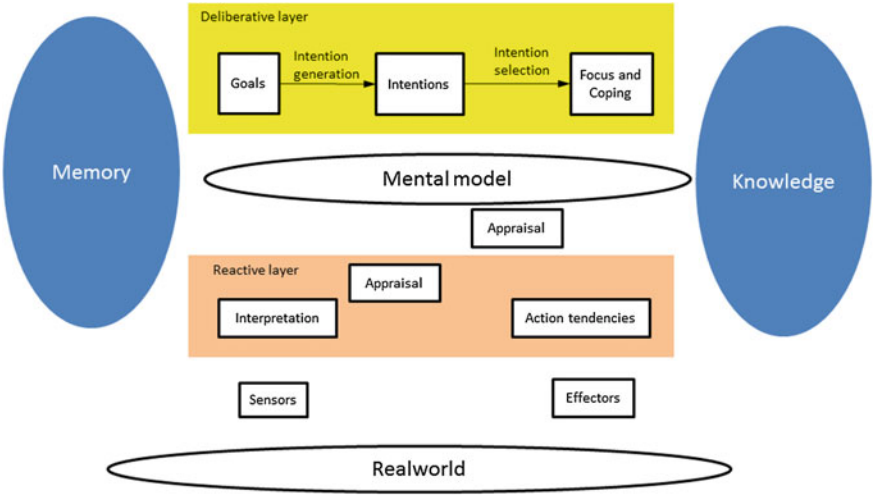


Fig. 4.6 Comprehensive architecture of conversational system

- The system should allow the programmer to write code across different levels of abstraction, ranging from signals to semantic representation.
- The system should allow the programmer to use high-level languages to produce their ideas in multiple levels of abstraction.
- The system architecture should allow the system to easily scale-up. For example, the system should be scalable to a system consisting of multiple processors, connected with each other through the global network, to integrate extensive knowledge to handle phenomena without introducing significant programming complexity.

In the early days of conversational system development, system architecture appears to be rather problem or task oriented for a given project. A typical example is the architecture of Peedy (p. 53).

The Generic ECA (GECA) is a generic framework for building an ECA system on multiple servers connected by a computer network (Fig. 4.7) (Huang et al. 2008). GECA allows mediating and transporting data streams and command messages among software modules. It provides a high-level protocol for exchanging XML messages among components, such as input sensors, inference engines, the emotion model, the personality model, the dialogue manager, the face and body animation engines, etc. An application programming interface is available for mainstream operating systems; thus, a programmer can easily adapt ECA software modules to the GECA platform. The blackboard model is employed as the backbone.

GECA has been implemented and applied for various applications, including a navigation agent, a quiz agent, and a pedagogical agent for teaching crosscultural communication.

When more efficient processing is required, blackboards might be divided to contain minimal sets of disjointed constraints so that they can be executed in parallel. This conforms to the generalization of the blackboard architecture into distributed artificial intelligence and multiagent systems.

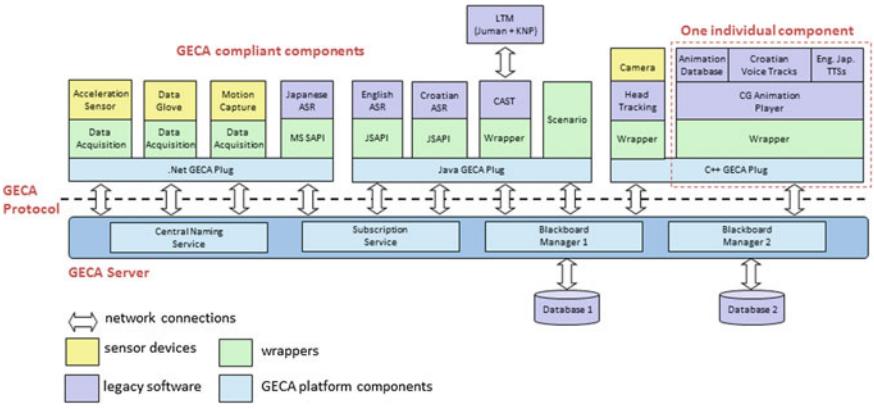


Fig. 4.7 The architecture of GECA (Huang et al. 2008)

4.3 Scripts and Markup Languages

Scripts and markup languages are used to specify the behavior of a conversational agent. They allow authoring conversational agent scenarios that can interact with the user or other agents without committing to implementation details.

A script language is more like a high-level programming language. An interpreter will take an expression in a given script language to produce animation for a given situation. In contrast, markup language is less procedural, allowing the target animation to be specified without complete procedural information, which introduces some flexibility. An action planner and an action realizer are required for procedural interpretation of markup language expressions to produce animation (Fig. 4.8).

The general requirements for script and markup languages are as follows.

- Synchronization of utterance, eye gaze, gesture, etc.
- Ability to express personality and information in terms of facial expression, utterances, etc.
- Ability to specify behaviors of more than one conversational agent.
- Ability to handle communication between the user and other agents.

The Artificial Intelligence Markup Language (AIML), developed between 1995 and 2000, is a markup language for specifying ELIZA-like dialogue patterns (Wallace 2003). As is shown in Fig. 4.9, its basic description element is a pattern-template pair that produces an output specified in the template section when an input that matches the pattern section is provided. The input sequence of words may be divided into segments according to the specified pattern. Thus, the output will be assembled using the segments included in the previous and current user inputs or using what the system has said to the user. The context mechanism, which allows referring to previous utterances and avoiding repetition of the same responses, is effective for increasing engagement. In the example shown in Fig. 4.9, the "that" tag is used to refer to the agent's previous utterance, which, as demonstrated, is a simple, yet useful trick.

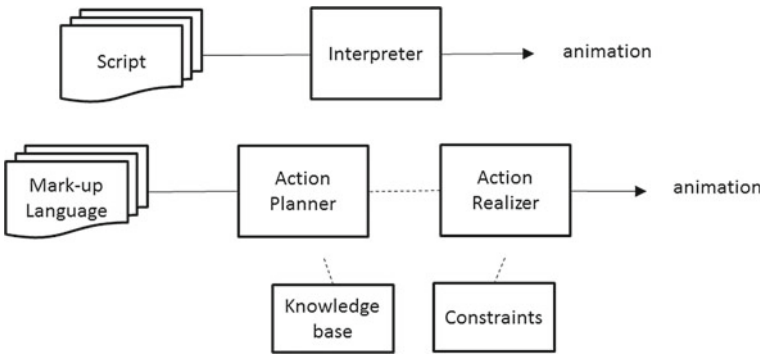


Fig. 4.8 Animation generation based from script/markup language

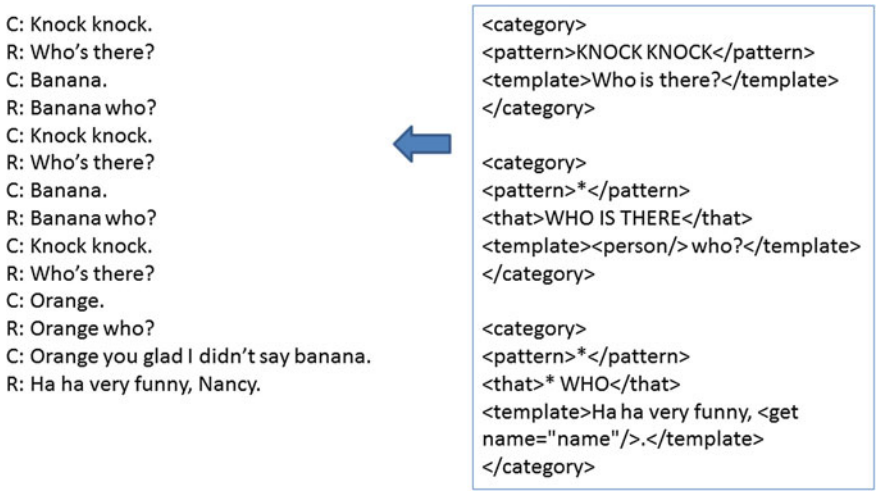


Fig. 4.9 AIML. (Wallace 2003)

Microsoft Agent is a programmable software service for authoring interactive presentation of animated characters, which was once provided as a part of the Windows operating system. The author was able to use Microsoft's speech recognition and text-to-speech engines as a part of user interaction. The author can control the behavior of visible or hidden agents using Visual Basic or VBScript in an intuitive fashion, as is shown in Fig. 4.10. There are other script languages, such as the Multimodal Presentation Markup Language (MPML) that allow for scripting a multimodal presentation in a similar manner (Ishizuka and Prendinger 2006).

With more advanced script languages and markup languages, such as STEP (Huang et al. 2004), the author can more precisely specify the details of agent behav-

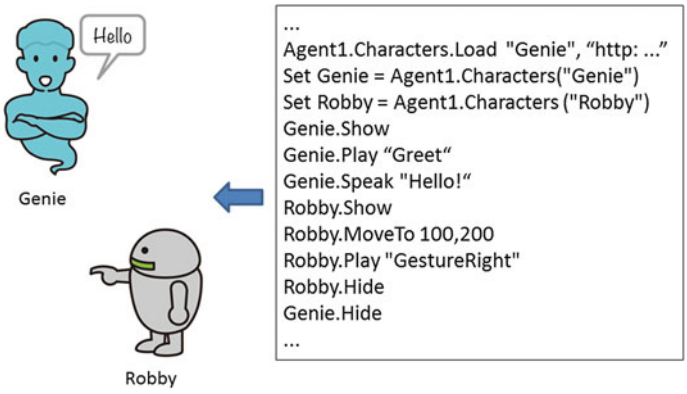


Fig. 4.10 Microsoft agent. Drawing based on (Microsoft Corporation, 1999) © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

iors, such as body movement. STEP features convenient specification of behaviors with an abstract level description, compositional semantics, reusability, parameterization, and easy description of interactions among various types of objects. STEP benefits from H-Anim, a standardization activity for specifying human body animation in VRML97. In H-Anim, a human body is modeled as a number of segments (e.g., forearm) that are connected by joints (e.g., elbow); thus, the author can refer to each segment and joint to alter the configuration (e.g., joint angle).

A parametric representation of the body, such as the one shown in Fig. 4.11, is also useful. Direction reference includes a collection of variables (X, Y, and Z) for specifying body orientation. The ranges for these variables are left/right, up/down, and front/back, respectively. Body reference involves a collection of variables for indicating various locations on the body, such as a joint node that connects one or more segment nodes for each body part, and a site node for a location at which one or more accessories, such as a hat, clothing, or jewelry, may be attached. Joint nodes are organized as a hierarchy. For example, a hand belongs to a subset of an elbow, which in turn belongs to a subset of a shoulder. If a body part is animated, its subsets follow, which allows easy programming. Body movements are specified around two main primitive actions, turn and move. The author can quantitatively specify the behavior using parameters for body movements. STEP is implemented in a distributed logic programming language. The programmer can use various facilities of the language to combine primitive animations to produce more complex character animations.

A more declarative method was proposed in the SAIBA framework, which consists of three stages to produce the behaviors of a conversational agent; i.e., intent planning, behavior planning, and behavior realization (Kopp et al. 2006). The Function Markup

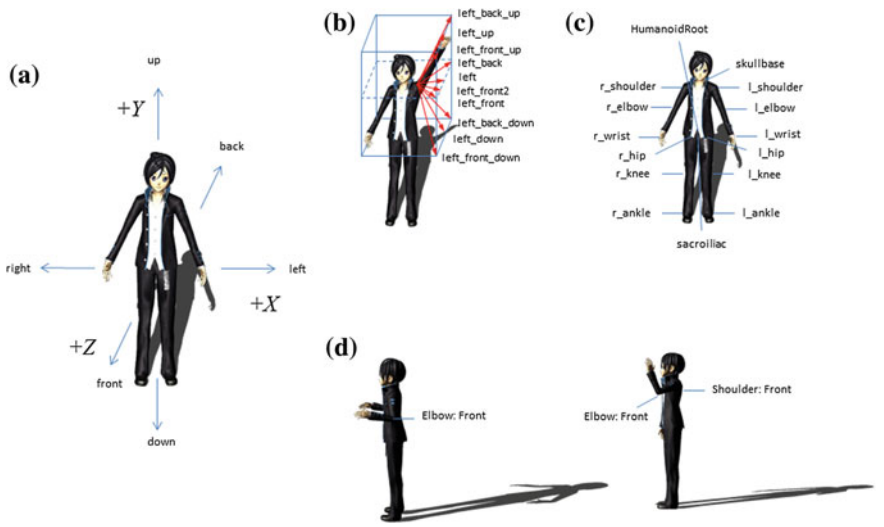


Fig. 4.11 Parameterized description of body parts. Drawing inspired by Huang et al. (2004). **a** Direction reference for humanoid **b** Combination of the directions for *left arm* **c** Typical joints for humaniod **d** Elbow joint in different situations

Language (FML) can be specified to describe an intent-level description without referring to physical details, and the Behavior Markup Language (BML) can be specified to describe a parametric and physical-level description of the behavior (Fig. 4.12). The goal of the project was to provide a powerful and unified model for representations of multimodal behavior generation in an application-independent and graphics-model independent fashion to present a clear-cut separation between functions and behaviors.

An initial version of BML was published with reference implementations (Kopp et al. 2006), while FML seems to be at an early stage of development (Heylen et al. 2008). BML uses Kendon's hierarchical framework for describing gestures at three levels (g-units, g-phrases, and g-phases), as is shown in Fig. 4.13. The g-unit encompasses a movement from the moment the articulators begin to depart from a position of relaxation until the moment they return to that position. Each gesture consists of one or more g-phrases that contain one g-phase, referred to as a stroke, in which a salient shape and dynamics that characterize the gesture are the most clearly manifested. In each g-phrase, a stroke g-phase is preceded by a preparation g-phase and followed by a hold g-phase.

One of the prominent features of BML is functional facilities to specify synchronization as a temporal constraint using behavior IDs and behavior synchronization points. A behavior ID is a unique identifier for an instance of behavior. A behavior synchronization point, or a *syncpoint*, uniquely refers to a significant point of alignment between gesture phases, as is shown in Fig. 4.14. This allows the author to annotate semantically the behaviors without considering the quantitative details of the behaviors, such as the time when an intended phase of the gesture starts or terminates. For example, the author can synchronize the speech "hello" exactly when the agent begins to wave a hand.

A fragment of BML specification for synchronization is illustrated in Fig. 4.15, wherein the wait behavior is used to describe a syncpoint implicitly determined by

Fig. 4.12 SAIBA model. Diagram inspired by Kopp et al. (2006)

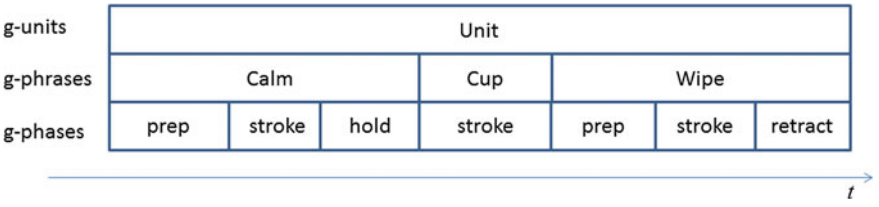


Fig. 4.13 Terminology of describing gesture segments proposed by Kendon (2004)

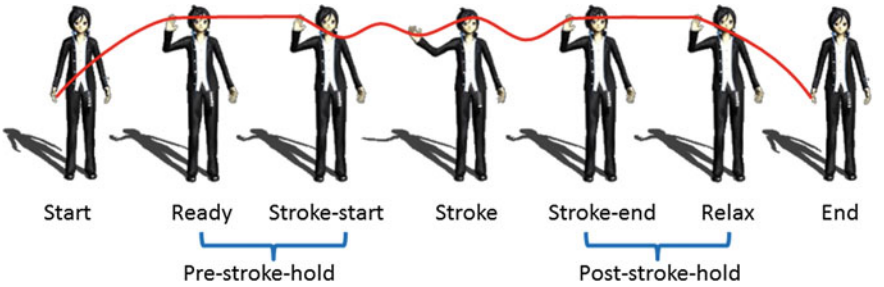


Fig. 4.14 Gesture phases used in BML which is based on Kendon's terminology. Drawing inspired by Kopp et al. (2007)

```
<bml>
   <gesture id="g1" type="point" target="object1"/>
   <body id="b1" posture="sit"/>
   <wait id="w1" condition="g1:end AND b1:end"/>
   <gaze target="object2" start="w1:end"/>
</bml>
```

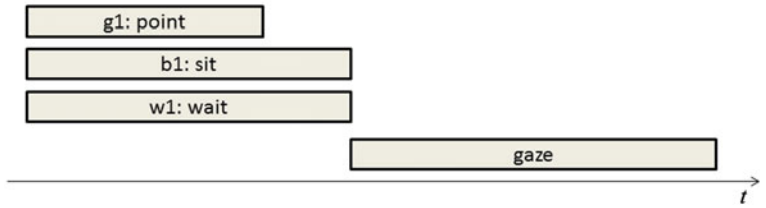


Fig. 4.15 Specification of a synchronization using a "wait" behavior (Kopp et al. 2007)

the end of another action to initiate a gaze behavior after the completion of both pointing and sitting behaviors.

Another example, shown in Fig. 4.16, illustrates how a start point of a gesture phase can be symbolically specified to follow immediately after the end of another event "w1," which will occur after five seconds of waiting once the speech "s1" is initiated. Speech "s2" will be initiated right after the five second waiting period is completed. As a result, the second speech might overlap the first speech if the first speech takes longer than five seconds.

For the BML segment shown in Fig. 4.17, syncpoints with unique identifiers are used to specify constraints among multiple behavior components. In this example, three actions, "s1", "d1", and "d2," are introduced. The constraints on syncpoints indicate that the start point of "s1" should be synchronized with the stroke starting point, the relaxing point of "d1" should be synchronized with the ready point of "d2," and the speech "that" of "s1" should start at the same time as the stroke point of "d2."

As the above examples suggest, the markup language simply specifies the constraints on the timing of involved actions; the markup language interpreter will

```
<bml>
   <speech id="s1" type="text/plain">
      First sentence
   <speech/>
   <event start="s1:end" emit="ACT1_COMPLETE" />
</bml>

<bml>
   <wait id="w1" event="ACT_COMPLETE" duration="5.0"/>
   <speech type="text/plain" start="w1:end">
      Second sentence
   </speech>
</bml>
```

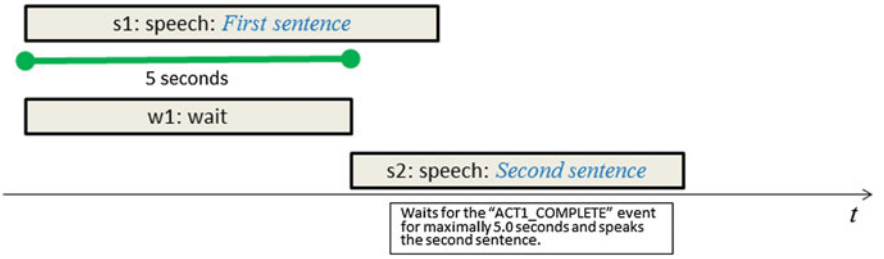


Fig. 4.16 Start point of the first speech s2 is set at five second after the first speech s1 (Kopp et al. 2007)

```
<speech id="s1"> <sync id="1"/>This or <sync id="2"/>that.</speech>

<gesture id="d2" type="DEICTIC" ... stroke="s1:2" />

<gesture id="d1" type="DEICTIC" ... stroke="s1:1" relax="d2:ready" />
```

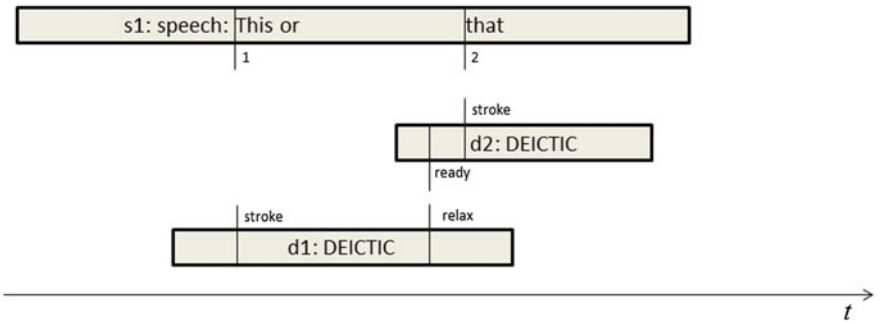


Fig. 4.17 Syncpoints with an identifier are used to specify synchronization (Kopp et al. 2007)

use the expressions to generate the behaviors by determining timing that satisfies the given constraints.

To generate behaviors from markup language specifications, a knowledge-based intelligent constraint solver (Fig. 4.18) is required to generate an actual event sequence for the physical realization level that may satisfy the given set of temporal constraints (Kipp et al. 2007).

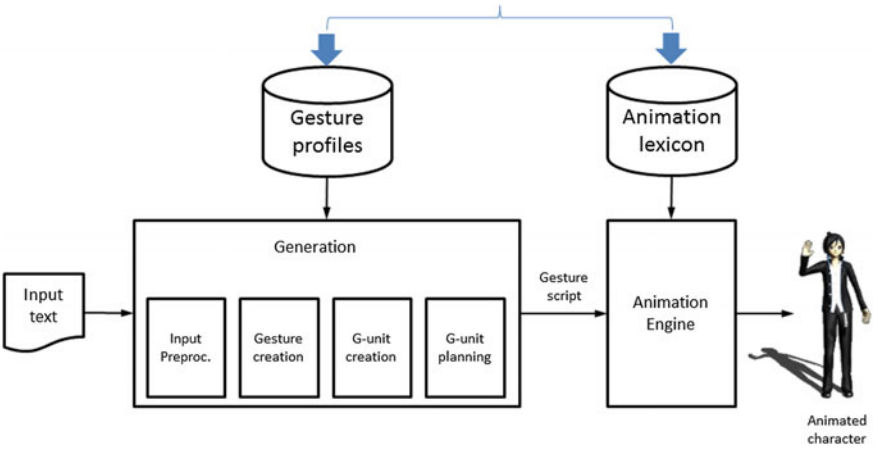


Fig. 4.18 The online phase. Drawing inspired by Kipp et al. (2007)

4.4 Basing Behaviors on Conversation Corpus

In the early days of conversational agent development, programmers often relied on their intuition to write and modify code for animating characters. Such a coding style often resulted in poor quality characters that were subject to criticism, such as being "wooden," because uniqueness and limitless variety were expected (Kipp et al. 2007).

A corpus-based approach realizes a data-driven approach and bases target behaviors of conversational agents on varieties observed in existing conversations. In corpus-based generation of conversation agent behaviors, one first creates an interaction corpus or a database containing instances of actual conversation behaviors, from which specification of interactions is generated in a markup language. By examining the behaviors of humans quantitatively, one might be able to determine the key features that differentiate appealing gestures from unappealing gestures. For example, it has been observed that figures in famous TV programs have significantly longer g-units than the laypersons reported in the literature (Kipp et al. 2007).

Building interaction corpora has been addressed in several projects. AMI (Carletta et al. 2006) aims to build a meeting corpus and focuses on conversation analysis. In this project, they constructed a corpus containing 70 h of scenario conversation and 30 h of free conversation. They attached a significant amount of metadata, such as time, transcripts of conversations, topic segmentation, summary, mental model of the participants, head and hand gestures, gaze directions, camera image, recorded voice, projected images, and white board strokes. The goal of CHIL (Waibel and Stiefelhagen 2009) was to extract human nonverbal behaviors automatically using machine learning methods. The CHIL corpus provides multimodal and multisensory recordings of realistic human behavior and interaction in lecture and meeting scenarios. Such data is useful to detect, classify, understand, learn, and

adapt to human activity. VACE (Chen et al. 2006) automatically collects and analyzes the visual content of meetings. The VACE corpus was recorded in real-world war game and military exercise scenarios (five to eight civilians, military personnel, or mixed). Chen et al. (2006) collected multimodal data, such as word transcriptions and prosodic features, and 3D head, torso, and hand motions. They focused equally on nonverbal and verbal interactions.

The main task of a corpus-based approach (Kipp 2004) consists of annotation and modeling, as is shown in Fig. 4.19. First, an interaction corpus is constructed, i.e., a dataset for a target phenomenon that is often an archive of videos systematically shot or collected. Second, the annotation phase follows, in which annotations (text notes) are associated with intervals of speech, gestures, or other data tracks in the collected video. The results of the annotation phase are annotation files and/or an inventory of tags used in the annotation phase, such as a gesture lexicon. Third, animation profiles and optionally an animation lexicon are generated to animate conversational agents as a result of modeling. Model parameters are calculated to quantitatively explain measures obtained from formalized observations represented as annotated tags.

In the annotation process, a collection of structured descriptions, referred to as *annotations* or *tags*, are associated with segments of transcription records. Each annotation to be associated with an interval of a record may consist of a variable and a value, e.g., HEAD=RaUHD ("the subject raises head") and TRUNK=LF ("the subject leans forward"). Here, HEAD and TRUNK are variables, and RaUHD and TRUNK are values. A coding scheme must be designed to specify the inventory of variables and the meanings of possible values.

Such a coding scheme can only be designed after performing further research that focuses on conversations. Such research could include conversation analysis, ethnography, social psychology, communication science, natural language analysis, and spoken language analysis. Systems for describing conversation-related phenomena include Kendon's hierarchical conceptualization of gesture, consisting of the g-unit, g-phrase, and g-phase (Fig. 4.20); McNeill's terminology for describing a hand gesture (Fig. 4.20) (McNeill 2005); Bull's posture and body movement scoring

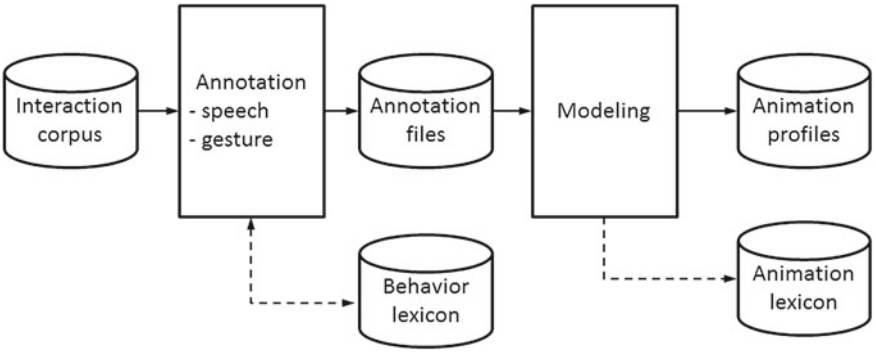


Fig. 4.19 A corpus-based approach (Kipp 2004)

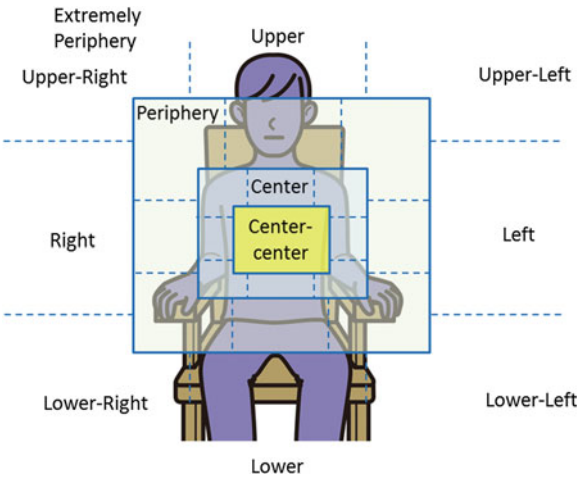


Fig. 4.20 McNeill's space manikin; drawing inspired by McNeill (2005). © 2014, Toyoaki Nishida and At, Inc. 2014. Reproduced with permission

systems (Bull 1987); and tones and break indices, which is a standard for labeling English prosody.

The annotation process is often accompanied by very costly manual labor performed by multiple human annotators. Investigators need to write a manual that contains specifications and definitions of lexicons so that the annotators are able to describe findings in a uniform fashion.

Annotation tools can be used to create an interaction corpus. Figure 4.21 illustrates a generic conceptual framework for ANVIL (Kipp 2004). The central concept of the framework is a coding scheme, i.e., a particular form with respect to structure and vocabulary that all annotations are expected to follow. A coding scheme that reflects the specific objectives of a research project must be defined prior performing annotative coding. The meta scheme provided by ANVIL is used to specify an individual coding scheme.

To obtain reliable annotations, it is necessary to employ more than one annotator. As mentioned previously, annotation is a time-consuming task, and cross validation, which ensures quality, introduces another cost.

Coding reliability may be examined in terms of segmentation and classification (Kipp 2004). Segmentation can be checked by determining the extent of agreement in the identification of meaningful time intervals for tags obtained among annotators. The measurement can be represented as

$$\frac{a}{a-d},$$

where a and d denote agreement and disagreement, respectively. Reliability of classification can be calculated from a confusion matrix.

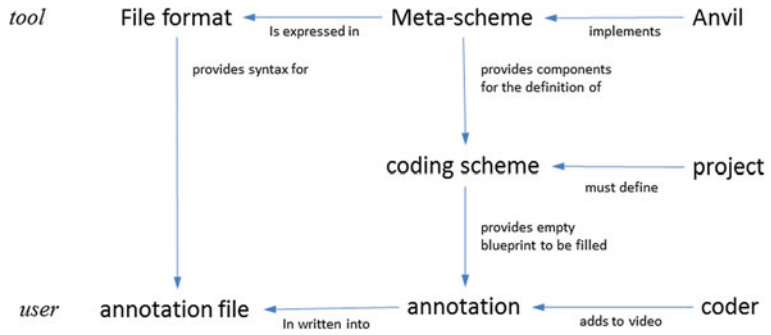


Fig. 4.21 Anvil annotation framework. Diagram inspired by Kipp (2004)

$$\begin{array}{ccccccc} & c_1 & c_2 & \cdots & c_n & & \\ c_1 & f_{1,1} & f_{1,2} & \cdots & f_{1,n} & f_1 \\ c_2 & f_{2,1} & f_{2,2} & \cdots & f_{2,n} & f_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdot \\ c_n & f_{n,1} & f_{n,2} & \cdots & f_{n,n} & f_n \\ & f_1 & f_2 & \cdots & f_n & N \end{array}$$

Here, c_i and $f_{i,j}$ denote i-th category and frequency of labels that are classified as c_i by the i-th annotator and c_j by the j-th annotator, respectively. The κ value measures the degree of disagreement beyond chance (Kipp 2004) and can be expressed as follows:

$$\kappa = \frac{p_a - p_e}{1 - p_e}.$$

Here, p_a denotes the ratio of agreement among coders, and p_e denotes the hypothetical probability of chance agreement. p_a and p_e are expressed as follows:

$$p_a = \frac{\sum_{i=1}^{n} f_{i,i}}{N},$$
$$p_e = \frac{\sum_{i=1}^{n} f_{i\cdot} f_{\cdot i}}{N^2}.$$

κ values range from -1 to 1. $\kappa = 1$ if the results of two annotators agree completely; κ approaches *zero* if agreement results are obtained only by chance; and $\kappa = -1$ if there is no agreement. The annotation data cannot be trusted unless the κ value is small; however, in practice the actual threshold depends on the text.

Cross validation (Fig. 4.22) can be used to validate the induced model (Kohavi 1995). The accuracy is calculated as follows:

$$acc_{\mathrm{CV}} = \frac{1}{n} \sum_{\langle v_i, y_i \rangle \in \mathcal{D}} \delta(\mathcal{I}(\mathcal{D}_t - \mathcal{D}_{(i)}, v_i), y_i),$$

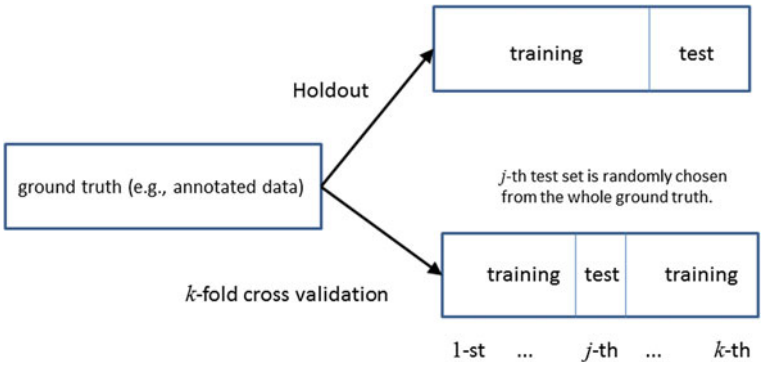


Fig. 4.22 k-fold cross validation contrasted with holdout. Drawing inspired by Kohavi (1995)

where $\mathcal{D} = \{\langle v_1, y_1\rangle, \langle v_2, y_2\rangle, \cdots \langle v_n, y_n\rangle\}$ stands for a labelled dataset, $\mathcal{D}_j$ stands for the j-th mutually exclusive subset of $\mathcal{D}$, $\mathcal{D}_{(i)}$ stands for the subset that includes the $\langle v_i, y_i\rangle$, $\mathcal{I}(\mathcal{X}, v)$ stands for the label assigned to an unlabelled instance v by the classifier the inducer (or an annotator) $\mathcal{I}$ has built based on dataset $\mathcal{X}$.

Cross validation may be compared to a holdout set, where accuracy is calculated as follows:

$$acc_{\mathrm{H}} = \frac{1}{h} \sum_{\langle v_i, y_i\rangle \in \mathcal{D}} \delta(\mathcal{I}(\mathcal{D}_t, v_i), y_i).$$

The advantage of cross validation over a holdout set is that cross validation utilizes ground truth data, which are often quite expensive, more efficiently.

Thoughtful selection of labels can lead to interesting findings. Nakano et al. (2003) shed light on nonverbal grounding by attentional behaviors toward the physical world. A map task was used to observe interactional behaviors of participants where one participant was trying to guide another to a destination using a shared floor map of a building. Nakano investigated the correlation between key social acknowledgment actions, i.e., answer, information request, assertion, and four nonverbal behaviors, i.e., gaze at partner (gP), gaze at map (gM), gaze elsewhere (gE), and head nod (Nod). Nakano found that speaker assertion accompanied by listener gaze during utterance was often followed by speaker elaboration (73 % of the time), whereas an assertion that was not accompanied by listener gaze was less frequently followed by elaboration (30 % of the time). In addition, when the listener was looking at the map, the speaker's next action was to inform the listener to continue looking (52 % of the time); otherwise, the speaker changed explanation behavior (73 % of the time). From those observations, Nakano suggested that the speaker interpreted continuous listener gaze at the speaker as the not-understanding signal and that it was often followed by the repair behavior by the speaker, as shown in Fig. 4.23.

CUBE-G is a project aimed at modeling culture-dependent conversation behaviors (Rehm et al. 2009). A dimensional model inspired by Hofstede was used to paramet-

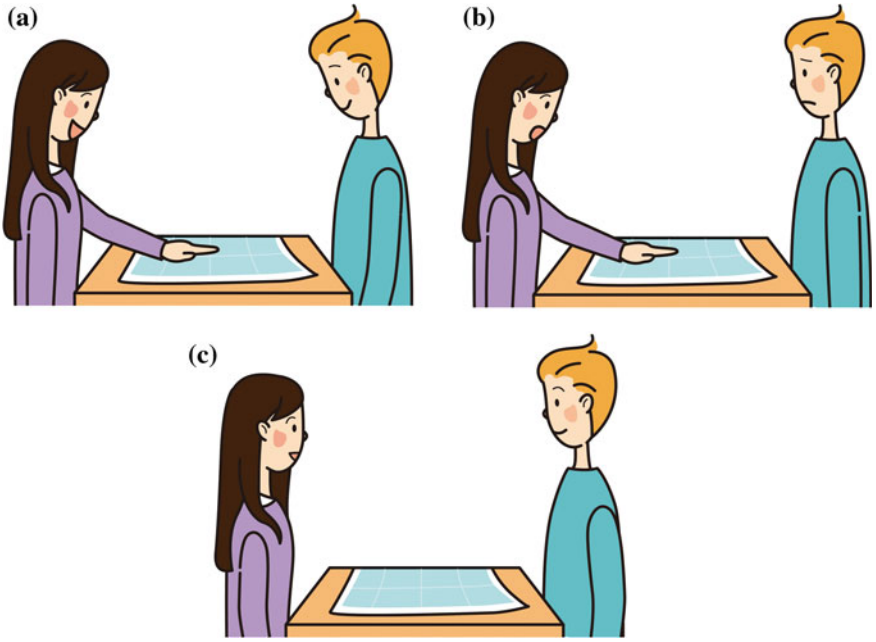


Fig. 4.23 Typical interaction pattern in the map task. **a** Understood, **b** Failure, **c** Repair. Diagram inspired by Nakano et al. (2003). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

rically specify the dependency of interaction styles on national culture by hierarchy (difference in social status), identity (individual-oriented versus group-oriented), gender, uncertainty (tendency to avoid anxiety originating from uncertainty), and orientation (long-term versus short-term).

A corpus containing three prototypical interaction scenarios was built to compare differences between German and Japanese interaction styles. The prototypical dialogues involved meeting someone for the first time, negotiating, and interacting with an authority. Significant differences were observed between participants from the two cultural backgrounds. For example, many Japanese participants performed gestures only with the lower arms. Generally, Japanese gestures were spatially confined while the opposite was observed for German participants, as depicted in Fig. 4.24. A Bayesian network model was created to represent cultural adaptation resulting from the two cultural backgrounds parametrically.

Recent research into embodied conversational agents has focused on corpus-based approaches. The employed annotations focused on particular behaviors of an embodied conversational agent to produce feasible and well-defined annotations. For example, Kopp et al. (2007) focused on illustrating gestures using *image description features* (IDF) (Fig. 4.25).

IDFs represent geometric and spatial features of referenced objects, as shown in Fig. 4.26. An interaction corpus was annotated with IDFs and used to generate the

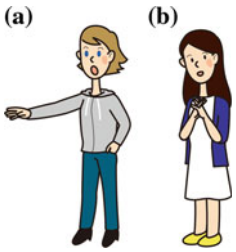


Fig. 4.24 Contrastive gestures of German and Japanese participants found in the CUBE-G project. Drawings inspired by Rehm et al. (2009). **a** Typical German gesture **b** Typical Japanese gesture. © 2014, At, Inc. Reproduced with permission

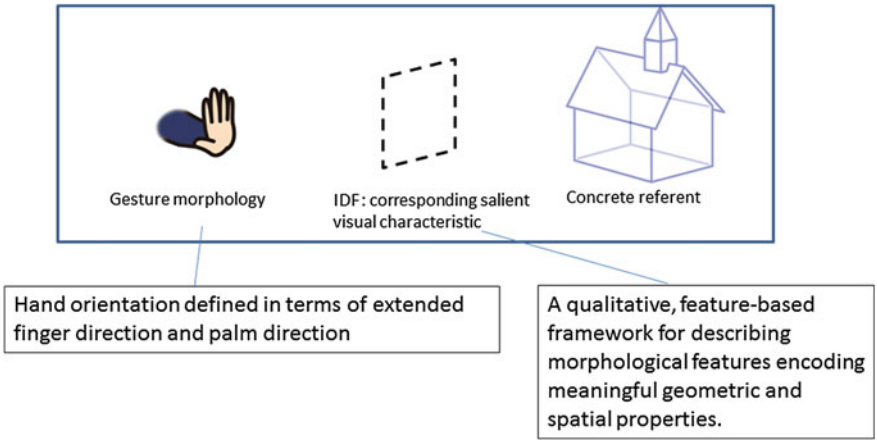


Fig. 4.25 Modeling illustrating gestures using image description features. Drawing inspired by Kopp et al. (2007). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

behavior of an embodied conversational agent system for campus navigation tasks, i.e., NUMACK.

4.5 Behavior Learning Using Motif Discovery

Machine learning techniques can be used to automatically or semi-automatically generate the interactional behaviors of conversational agents. The role of machine learning is essentially to induce a generic mechanism of producing communication acts from a collection of records of conversation (Fig. 4.27). Machine learning techniques allow for both reducing the programming cost and increasing the quality of the codes by basing the result on observation.

Machine learning techniques are classified into supervised learning and unsupervised learning. Supervised learning generalizes an input-output mapping function from training data, while unsupervised learning clusters a given collection data based

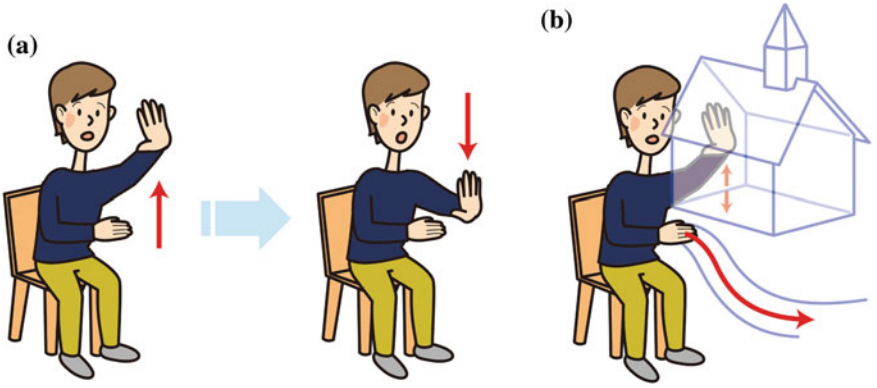


Fig. 4.26 Example of illustrating gestures modeled by IDF features. Drawing inspired by Kopp et al. (2007). **a** Gesture **b** Gesture space. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

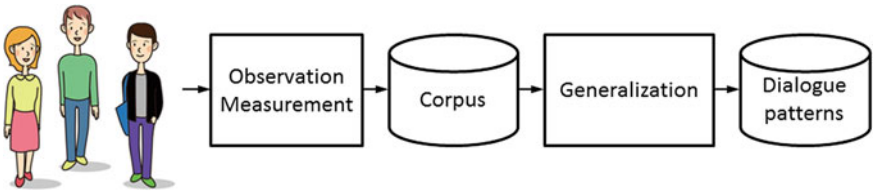


Fig. 4.27 Using machine learning. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

on a given function that computes similarity between data items. Supervised learning allows us to generalize the observation more reliably but at more cost. To benefit from supervised machine learning, we need to accumulate a large amount of data of conversation annotated with basic units and their structure. It motivates another venue of utilizing unsupervised learning.

Let us see in more details how to induce dialogue and interaction patterns from data. The problem can be formulated as in Fig. 4.28. There are three major challenges. The first is to distinguish basic communication acts from noises. The second is to find causality among communication acts. The third is to induce a generic mechanism underlying the structured set of communication acts among agents.

Let us see now how we can identify signals in a given time series. Given the assumption that the signal is recurring in a time series, the problem can be formulated as a motif discovery problem, where a motif is defined as a pattern that can be repeated in a time series almost in the same form.

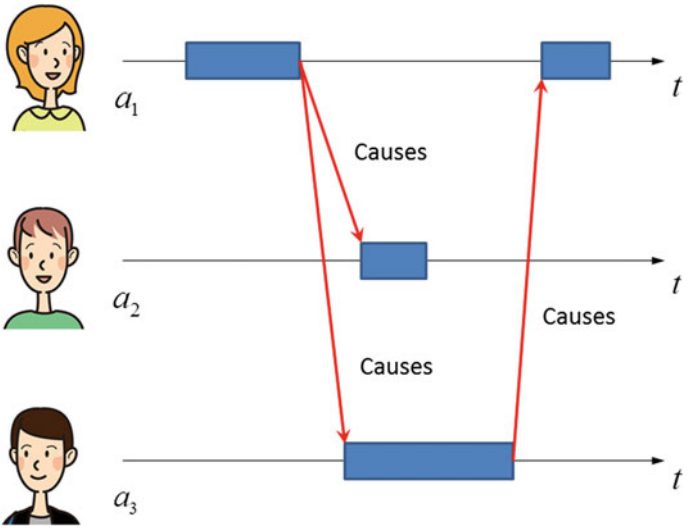


Fig. 4.28 Conversation as a structured collection of communication acts. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

4.5.1 Motif Discovery in Discrete Sequences

Discovering communicative acts from interaction time-series is a special case of a known data mining problem called *motif discovery*. In this section, we introduce the motif discovery problem in its historical context focusing on its versions related to the discovery of patterns in interactions and conversation.

Motif discovery was first discussed in discrete sequences with applications primary to DNA, RNA and protein sequences by the late 80s and the early 90s (e.g., Staden, 1989). One of the first occurrences of the term was by Waterman, Arratia, and Galas where it was called *consensus motifs* (Waterman et al. 1984). The main interest of these early algorithms was discovering recurring patterns in DNA, RNA and protein sequences. These sequences are discrete in nature with a small finite alphabet size (4 nucleic-bases in case of DNA {A,T,C,G} and RNA {U,T,C,G} and 20 amino-acids in case of protein {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}).

Few definitions are needed to state any form of the motif discovery problem in this context. An alphabet Σ is a finite set of symbols with cardinality $|\Sigma|$. A sequence s of an alphabet Σ is a string (ordered list) of symbols where each symbol is a member of the associated alphabet (Σ).

A subsequence $s_{i,l}$ is the sequence of l characters starting at position i to position $i + l - 1$ of the original S sequence (when l is known from the context we will drop it and use s_i to mean $s_{i,l}$). When we are working with more than one input sequence, we will use superscripts to indicate the sequence index and subscript to

indicate subsequences. For example, $s_{i,l}^{(j)}$ is the subsequence of length l starting at position i in the sequence number j.

A distance function $D_l(p, q)$ is a mapping $|\Sigma|^l \times |\Sigma|^l \to \mathbb{R}$ that measures how different are these two sequences. Examples of distance functions are the *Hamming* distance which measures the number of places in which the two sequences differ and the *Levinstein* distance which measures the number of edit operations (mutation, insertion and deletion) that are required to match the two sequences. Notice that both of these distances generate integers rather than real numbers but we will stick with using $\mathbb{R}$ as the range for the sake of generality. Again when l is known from the context we will drop it and use $D(\cdot, \cdot)$.

The distance between two sequences of different length is defined as:

$$\mathrm{D}_{|p|}\,(\mathrm{p},\,\mathrm{q}) = \mathrm{D}_{|p|}\,(\mathrm{q},\,\mathrm{p}) = \min_i D_{|p|}\left(p, q_{i,|p|}\right), \tag{4.1}$$

where (without loss of generality) we assume that $|p| \leq |q|$. A sequence p is said to occur (or occur exactly) in another sequence q (where $|p| \leq |q|$) when $D_{|p|}\,(p, q) = 0$. A sequence p is said to occur approximately within a range of $\tau \in \mathbb{R}$ in another sequence q (where $|p| \leq |q|$) when $D_{|p|}\,(p, q) \leq \tau$. The sequence p is said to occur approximately in q at locations $\{i\}$ when $D_{|p|}\,(p, q_i) \leq \tau$.

Given these definitions we can define two related problems that were among the first attempts at motif discovery.

Definition 4.1 *Single-Length Pattern-discovery problem* (*PDP*): Given a set of t sequences $\{s^{(j)}\}$ $(1 \leq j \leq t)$, a motif length l, a distance function D and a threshold value τ, find all sequences that occur approximately with range τ in at least q out of the t sequences.

Pavesi et al. (2001) introduced a slightly different form of this problem. In this case PDP is applied to all lengths from l to the maximum possible limit: $\min_j \left|s^{(j)}\right|$ (Pavesi et al. 2001). This problem is motivated by the practical problem of finding binding sites in unaligned DNA sequences like the ribosome binding site problem (Tompa, 1999). Notice that in PDP, the resulting sequences may not occur in *any* of the input sequences exactly. This problem goes also with the name *common motif discovery problem* (Sagot 1998).

There are in general two approaches to solve this problem: the first approach starts from possible sequences of length l and finds out which satisfy the condition in the PDP definition. Algorithms taking this approach are called pattern-driven methods in literature (e.g., Pavesi et al. 2001). The second approach is to start from the input sequences s^k and use them to infer subsequences that satisfy this condition. We call algorithms taking this approach sequence-driven methods.

The simplest pattern-driven method is to generate a table called P of all patterns of length l which will have a length of $|\Sigma|^l$. For each member of this table p_i, we calculate $D_l\left(p_i, s^{(j)}\right)$ for all sequences $\{s^{(j)}\}$ $(1 \leq j \leq t)$ and count the number of times where $D_l\left(p_i, s^{(j)}\right) = 0$. These counts are stored in a list called C with members c_i storing the count for sequence p_i. For each pattern $p_i \in P$, we then

calculate a new count (c'_i) which sums the counts of all other member of P that occur approximately within range τ from p_i (i.e., $c'_i = \sum_{j|D_l(p_i,p_j)\leqslant\tau} c_j$). Finally from each subset of P that is within range τ from each other, we keep only the member with maximum c'_i value. This approach was used with slight modification by Staden (1989). A toy example of this algorithm with $\tau = 1$ and $l = 2$ is given in Fig. 4.29. For more realistic examples and analysis of this algorithm, the reader is referred to Staden (1989).

There are two major shortcomings with this simple approach to the PDP problem. Firstly, the approach is only practical in small values of l and $|\Sigma|$ because the sizes of Q, and C as well as execution time are exponential in the alphabet size (i.e., $O\left(|\Sigma|^l\right)$). For DNA applications, $|\Sigma| = 4$ and for the ribosome binding site problem $l = 7$. This made it possible to use a similar approach for solving this problem (Tompa, 1999). The second problem of this simple approach is that only recurrence of the pattern is not enough to make it significant because some patterns may appear (specially at small values of l) with high frequency just because of the high probability of their generation given the frequencies of character appearance in the background signal. For this reason another significance measure is usually used to keep only the patterns that are significant (i.e., having a small probability of being generated randomly in the sequences). The most used such measure is the relative entropy between character

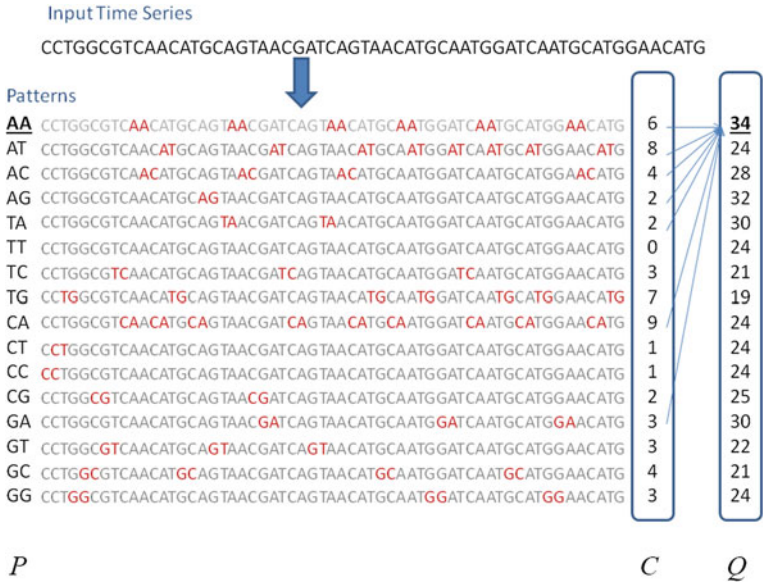


Fig. 4.29 A simple example of the pattern-driven method for solving the PDP problem. All possible patterns of the given length are generated, and their occurrences in the input string are found in the list C. The occurrence-count of all patterns within τ distance from each pattern is then calculated in the list Q (the ones for the pattern 'AA' are shown by *arrows*). The maximum(s) of Q are then selected as motifs (Here only one candidate: *AA*)

frequencies in the pattern p_i and character frequencies in the whole sequence set $\{s^{(j)}\}$ $(1 \leq j \leq t)$ (Pavesi et al. 2001).

Another related problem that was formulated around the same time is the planted (l, d)-motif problem (Buhler and Tompa 2002). This problem best illustrates the sequence-driven approach.

Definition 4.2 *Planted* (l, d)*-Motif Problem* (*PMP*): Let M be a fixed but unknown sequence (the motif) of length l. Suppose that M occurs once in each of t background sequences $s^{(j)}$ $(1 \leq j \leq t)$ of common length n, but that each occurrence of M is corrupted by exactly d point substitutions in positions chosen independently at random. Given the t sequences, recover the motif occurrences and the motif M.

The earliest algorithms for solving this form of the problem employed local search including CONSENSUS, GibbsDNA, and MEME. Pevzner et al. (2000) developed two algorithms, WINNOWER and SP-STAR, to more reliably solve the planted (l, d)-motif problem. Briefly, WINNOWER constructs a graph whose vertices correspond to all subsequences of length l present in the t input sequences, with an edge connecting two vertices if and only if the corresponding subsequences differ in at most $2d$ positions and do not both come from the same sequence. WINNOWER then looks for a clique of size l in this graph. This guarantees that these cliques represent all the occurrences of M because no other subsequence can be approximately occurring in any of them within a range of $2d$. The choice of $2d$ rather than d is due to the fact that given exactly d point substitutions in two sequences, the maximum *hamming* distance between them is $2d$ which happens when the sequences has no distortions in the same position in both.

The second algorithm, SP-STAR, is a local search method that starts in turn from each individual subsequence of length l in the input, chooses the closest match to this subsequence from every other input sequence, and uses a sum-of-pairs score and iterative refinement to converge on a good motif (Pevzner et al. 2000). Both SP-STAR and WINNOWER algorithms are shown in Fig. 4.30.

PMP and PDP problems differ in several points that are relevant to our discussion for learning recurring patterns in interaction or conversation. Firstly, PMP assumes a specific distance function (the hamming distance) while PDP is a more general form of motif discovery that can utilize any distance function. This translates exactly to the discovery of recurring interaction patterns. Secondly, PMP assumes a specific distance to the unknown implanted motif while PDP only assumes a distance limit (maximum acceptable distance). This makes PDP a better representative of the problem of discovering interaction patterns (and in fact discovering patterns of DNA sequences). For these reasons PMP can be thought of as an idealized form of PDP.

Definition 4.3 *Repeated Motif Discovery Problem* (*RMD*): Given a sequence s, a motif length l, a real number τ, and a positive integer $n > 1$, find all sequences of length l that occur in s approximately within τ at—at least—n different locations.

Sagot (1998) employed this form where he proposed a suffix tree based algorithm for discovering these motifs. The RMD problem is an extension of the planted (l, d)-motif problem because it allows defects (substitutions) in less than d points.

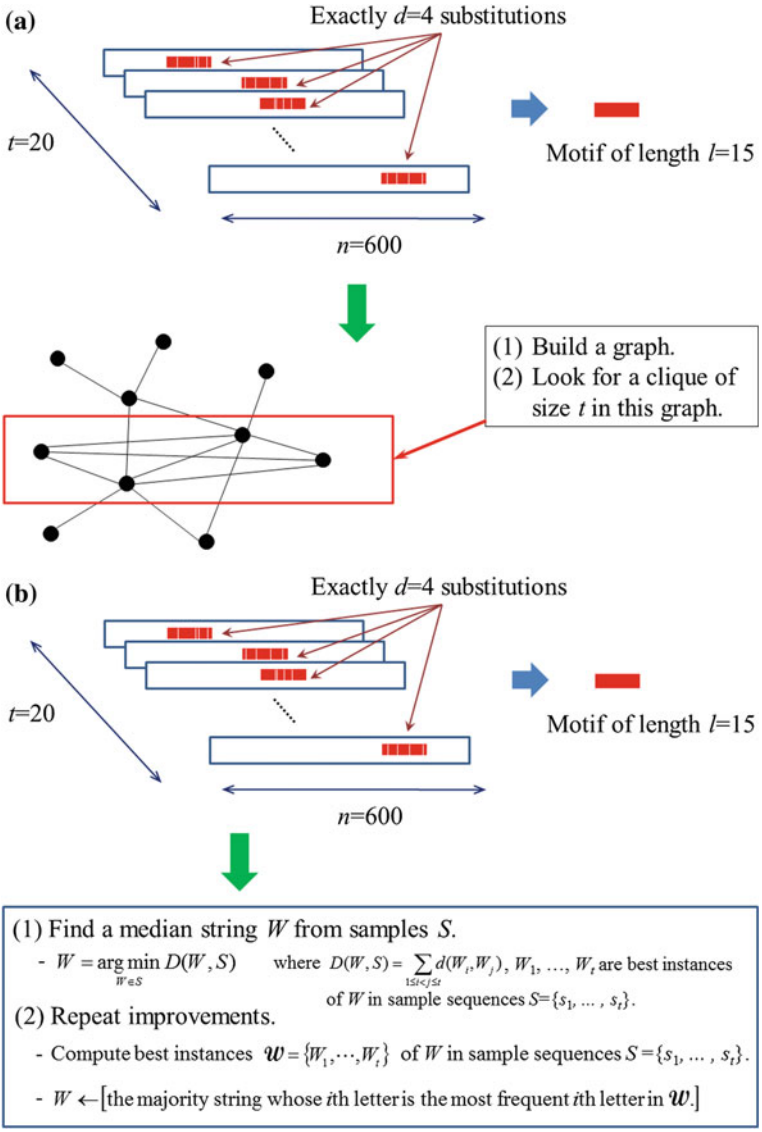


Fig. 4.30 The WINNOWER and SP-STAR algorithms as examples of sequence-driven approaches to discrete motif discovery. In both cases, there is no need to build the table of all possible patterns. **a** The WINNOWER algorithm **b** THe SP-STAR algorithm

Buhler and Tompa (2002) proposed an important approximate solution to the planted $(l, 2)$-motif problem (that can also be used to solve both RMD and PDP problems) called the PROJECTIONS algorithm. This algorithm was later utilized in discovering motifs in real world time-series by several researchers which makes it

of direct interest to our goal of discovering interaction patterns from logs of human-human conversations or interactions.

PROJECTIONS is a seeding step that aims at discovering probable candidates for the hidden motif M. The main idea of PROJECTIONS is to select several random hash functions $f_m\left(s_i^{(j)}\right)$ $(1 \leq m \leq r)$ and use them to hash the input sequences. Occurrences of the hidden motif are expected to hash frequently to the same value (called *bucket*) with a small proportion of background noise. Noise sequences on the other hand are not expected to hash to similar values. If the hash functions are different enough (and complex enough), we can use the buckets with largest hits as representing occurrences of the motif. This is an initialization step that can then be refined using the EM algorithm in order to recover the full motif. These steps are shown in Fig. 4.31.

There are several internal parameters to PROJECTIONS that need to be set. Firstly, the hashing functions f_m must be determined and their number r must be decided. Secondly, it is not expected that all occurrences of the same motif will hash to the same bucket for every function so we need an estimate of the number of hits to count as significant for each bucket (h).

PROJECTIONS use hashing functions that balance ease of computation and versatility. Each function is a mapping from sequences of l characters (the input size) to sequences of k characters where $k < l$. The hashing function $f_m\left(s; \langle l_1, l_2, ..., l_k \rangle\right)$ simply selects the characters of s at the k positions and concatenates them to create a sequence. This leads to $|\Sigma|^k$ buckets. Notice that we need not store all of these

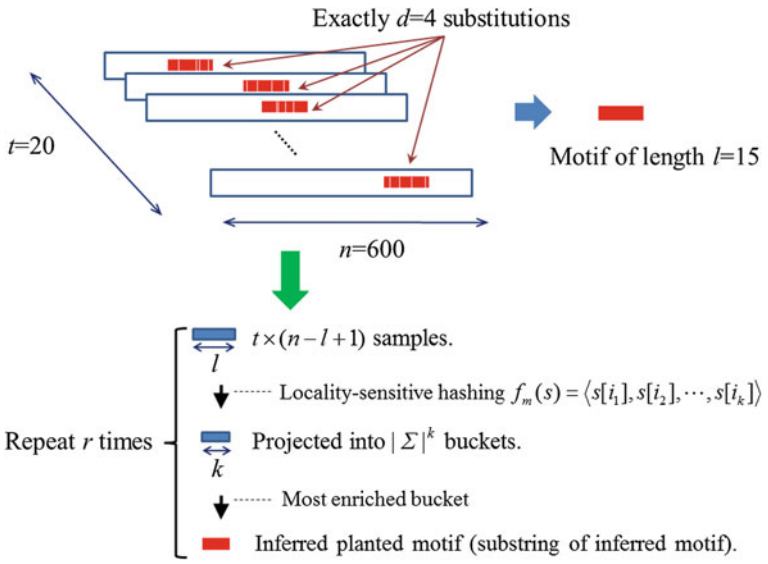


Fig. 4.31 The PROJECTIONS algorithm for finding *candidate* motifs. The algorithm hashes all subsequences using θ different but locations sensitive hashing functions into *buckets*. The bucket with the largest number of hits is selected as the candidate motif

buckets as most of them will be empty which makes the bucket list a sparse matrix for large enough values of k.

Now that we have decided the form of the hashing functions, we need to decide how to select them from all possible hashing functions of that form (there are $k!$ possible such functions). We simply select r random functions of this function set.

Two questions now remain concerning the hashing functions: How to select k and r. A simple way to select k is to select a value that minimizes the probability that two random sequences of length n will hash to the same bucket. This can be satisfied by selecting $K \geqslant \log_{|\Sigma|}((t\,(n-l+1))/e)$ where e is the allowed expected number of random sequences that hash into the same bucket (usually selected as some number less than 1). The guiding principle in selecting r is that we want to continue hashing until the number of occurrences of the hidden motif that hash to the same bucket are expected to be greater than some cutoff value θ that can then be used to decide which buckets contain possible occurrences of the motif we are after. Buhler and Tompa (2002) have shown that we can select r such that:

$$r = \left\lceil \frac{\log(1-q)}{\log(B_{l,p}(s))} \right\rceil \tag{4.2}$$

where q is the probability that the planted bucket contains θ or more planted motif instances in at least one of the r trials, and $B_{l,p}(s)$ is the probability that there are fewer than θ success at l independent Bernoulli trials each having probability p of success, and p is the probability that an occurrence will hash to a specific bucket which can be calculated as:

$$p = \binom{l-d}{k} \Big/ \binom{l}{k}. \tag{4.3}$$

The final piece of the puzzle is how to select the cutoff number of hits to a bucket to be considered as a candidate occurrence set of a motif (θ) which played an important rule in selecting the value of r. Unfortunately, there is no agreed upon method to select θ but practical tests have shown that as long as it is near the value of d usually good results can be achieved, e.g., 3–4 for the case of the (20, 4)-motif case (Buhler and Tompa 2002).

Even though, all of the formulations and algorithms discussed so far focus on discrete sequences, it can be directly extended to real world time series by first discretizing the time series then just using one of these algorithms.

Both PDP and RMD can have direct applications to interaction data mining. For example, consider the case of discovering signs of attention in the nonverbal behavior of students (or the nonverbal aspects of their vocalization) given a set of class-room interactions in which we know that the subjects were showing attentive behavior (based on subsequence results in evaluation tests). This problem can easily be formulated as a pattern-discovery problem (PDP) where the set of sequences are discretized versions of the interactions and the hidden motif is the sign of attention in this behavior of students. One immediate problem here is that the motif length is

not known a-priori but a simple solution is to search for motifs at multiple lengths then utilize a human expert to confirm or reject them (leading to a semi-supervised approach).

This example hides some of the complications that result from the fact that our data is originally continuous and in most cases multidimensional. These complications will be discussed in details in Chap. 9. In what follows, we will introduce the problem of motif discovery in real-valued time-series focusing on the single dimensional case and provide some introduction to the techniques employed in solving it.

4.5.2 Motif Discovery in Real-Valued Time-Series

Consider the case of gesture discovery: Fig. 4.32a shows three occurrences of a waving gesture within a longer time-series. This time-series was generated by having a human subject perform several movements including the highlighted three waving gestures in front of a Kinect sensor. Skeletal data from the sensor are then passed through an inverse-kinematics system (Mohammad and Nishida 2013b) that returns four joint angles for each arm at every time step (these are shown in Fig. 4.32a for the left arm). The four time-series for the left arm are then combined using Principal Component Analysis (PCA) to form a single time-series that is shown in Fig. 4.32b. The three highlighted parts of the time series correspond to three occurrences of a single motif.

A time series $x\,(t)$ is an ordered set of T real values. A subsequence $x_{i,j} = [x\,(i) : x\,(j)]$ is a contiguous part of a time series x. Given two subsequences $\alpha_{i,j}$ and $\beta_{k,l}$, a distance function $D\,(\cdot, \cdot)$ and a positive real value R, we say the two subsequences *match up to* R if and only if $D\left(\alpha_{i,j}, \beta_{k,l}\right) < R$. We call R the range following Lin et al. (2002). In most cases, the distance between overlapping subsequences is considered to be infinitely high to avoid assuming that two sequences are matching just because they are shifted versions of each other, called trivial motifs (Keogh et al. 2005).

Lin et al. (2002) provided one of the earliest discussions of motif discovery in the context of real-valued time-series from which we borrow the following definition of a motif:

Definition 4.4 *K-Motifs*: Given a time series x, a subsequence length l and a range R, the most significant motif in x (called thereafter *1-Motif*) is the subsequence C_1 that has the highest count of non-trivial matches (ties are broken by choosing the motif whose matches have the lower variance). The K^{th} most significant motif in x (called thereafter *K-motif*) is the subsequence C_K that has the highest count of non-trivial matches, and satisfies $D\,(C_K, C_i) > 2R$, for all $1 \leq i < K$.

We utilize a slightly different definition which is more appropriate to our goals of discovering interaction patterns.

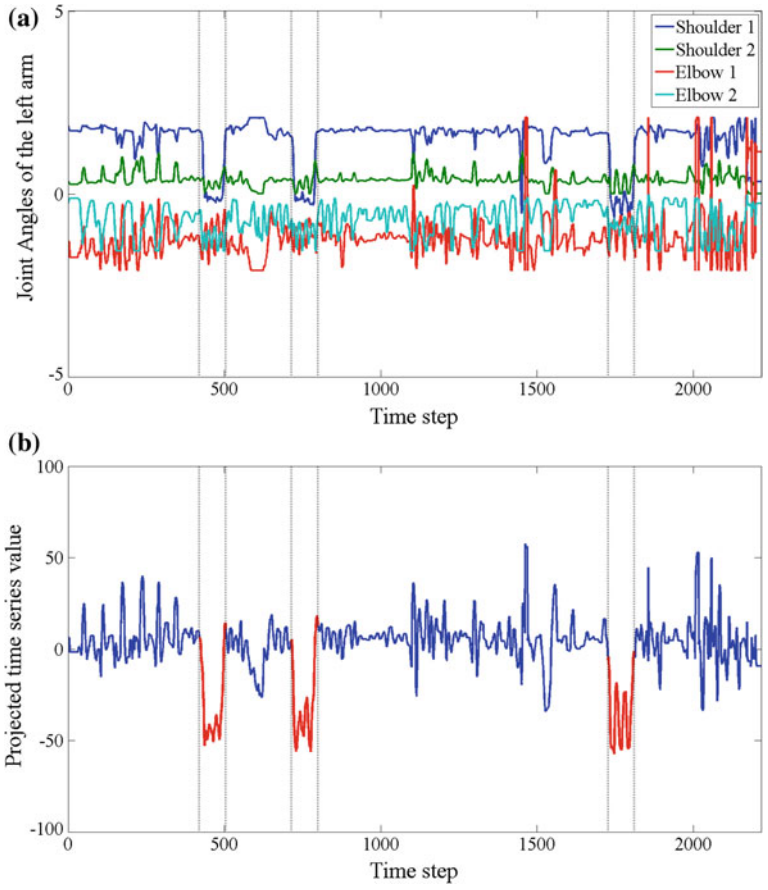


Fig. 4.32 Three occurrences of an approximately recurrent motif (ARM) in the stream of joint angles representing three waving gestures collected by a Kinect sensor. **a** Joint angles of the left arm **b** PCA projected time-series

Definition 4.5 *Approximately Recurrent Motif with connectivity C* (ARM_c): Given a time-series x of length T, a lower and an upper limit on motif lengths (l_{min} and l_{max}), a range R, and a connectivity level C where $0 < C \leq 1$, an ARM with connectivity c (M_c) is a set of n subsequences ($\{m_1, m_2, ..., m_n\}$) where for each $i \in [1, n]$ there exists K numbers $\{j_k\}$ where $j_k \in [1, n]$, $j_k \neq i$, $D\left(m_i, m_{j_k}\right) < R$, and $K \geq c \times n$. Each $m_i \in M$ is called an *ARM occurrence* or just *occurrence*. M is considered an ARM_c only if it is maximal in the sense that extending any of its occurrences breaks one of the aforementioned conditions on M.

Definition 4.6 *Approximately Recurrent Motif* (*ARM*): an ARM is a shorthand for an ARM with connectivity of 1.0.

Definition 4.7 *ARM Cardinality* ($Card\,(M)$): Given an ARM (M), its cardinality is the cardinality of the set (M).

Definition 4.8 *ARM Discovery* (*ARMD*): Given a time-series x of length T, a lower and an upper limit on motif lengths (l_{min} and l_{max}), a range R, a connectivity level c and minimum recurrence count RC, and maximum overlap MO, find all ARM_cs (M_i) that satisfy Card(M_i) $\leq RC$ and maximum overlap between occurrences of any two ARM_cs is less than MO.

Chiu et al. (2003) generalized the definition of a K-motif (Definition 4.4) to account for *don't care* subsections. It is clear that K-motifs are not ARMs. For example it is likely that the 1-Motif and 2-Motif are actually members of the same ARM. Also a K-motif is a single subsequence from the time series while an ARM is a list of occurrences that can be used for modeling (say using an HMM or an ARMA process).

Generally speaking, ARMs are what robots and infants look for when trying to discover important events in their perceptual spaces, while K-motifs are more mathematically tractable entities that are of more interest to data miners looking for interesting patterns in the time-series.

Nevertheless, discovering ARMs and K-motifs are related problems. Any algorithm that can discover K-motifs can be used to discover ARMs by simply finding several K-motifs then inducing an undirected graph where each vertex represents a K-motif and an edge connects two vertices iff the distance between their corresponding subsequences is less than the range. Once this graph is induced, each connected component represents an ARM_c for some value of c and each clique represents an ARM.

Several algorithms were suggested to discover K-motifs (e.g., Mohammad and Nishida 2009a; Chiu et al. 2003; Oates 2002; Jensen et al. 2006; Lin et al. 2002; Minnen et al. 2007; Tang and Liao 2008). Many of these algorithms are based on the PROJECTIONS algorithm proposed by Tompa and Buhler (2001) which uses hashing of random projections to approximate the problem of comparing all pairwise distances between n subsequences to have linear rather than quadratic space and time complexities. Because this algorithm works only with discrete spaces, the time series must be discretized before applying any of PROJECTIONS variants to it. A common discretization algorithm employed for this purpose is the SAX (Lin et al. 2002).

Minnen et al. (2007) proposed An unsupervised method for finding a sensible range parameter for these algorithms. MCFull (Mohammad and Nishida 2009a) differs from all of these approaches (even with automatic range estimation) in requiring no discretization step and being able to discover motifs in a range of lengths rather than a single length. This algorithm also has adjustable space and time complexity and is linear in the worst case, while all PROJECTIONS based algorithms require good selection of the discretization process parameters to lead to sparse collision matrices in order to avoid being quadratic.

Catalano et al. (2006) proposed another approach for finding these motifs that uses random sampling from the time series (without any disretization). This algorithm

requires an upper limit on the motif length and also is not guaranteed to discover any K-motifs or to discover them in order. An explicit assumption of this algorithm is that the motifs are frequent enough that random sampling will has a high probability of sampling two complete occurrences in candidate and comparison windows of lengths just above the maximum motif length.

The sampling process was improved in MCFull by utilizing a change point discovery algorithm to guide the sampling process with reported significant increase in discovery rate (Mohammad and Nishida 2009a).

Even though no clear definition of what is actually discovered by these last two algorithms (other than being frequent), they actually discover ARMs.

Definition 4.9 *Exact Motif*: (Mueen et al. 2009) An Exact Motif is a pair of subsequences $x_{i,l}, x_{j,l}$ of a time series x that are most similar. More formally, $\forall a, b, i, j$ the pair $\{x_{i,l}, x_{j,l}\}$ is the exact motif iff $D\left(x_{i,l}, x_{j,l}\right) \leq D\left(x_{a,l}, x_{b,l}\right)$, $|i - j| \geq w$ and $|a - b| \geq w$ for $w > 0$.

Exact Motifs as defined in Definition 4.9 (Mueen et al. 2009) are more similar to ARMs than K-motifs but it still keeps only two members of each motif which reduces the usability of the system for motif modeling (e.g., using HMMs). An algorithm for discovering exact motifs according to this definition (and the definition itself) was given by Mueen et al. (2009). In this work, the Euclidean distance between pairs of subsequences of length l was used to rank motif candidates. This has the problem of requiring a predefined motif length. It is also sensitive to short bursts of noise that can affect the distance.

Oates (2002) gives yet another definition of *recurrent patterns* in time-series that uses probabilistic modeling and provides an algorithm (PERUSE) that can be used to discover these patterns. This is the nearest definition to ARM we found in literature but it assumes a probabilistic interpretation of the generation process which is not assumed by Definitions 4.5 and 4.6. PERUSE also assumes that the patterns to be found are frequent enough that random sampling from the time series will yield to complete patterns. This is the same assumption used by Catalano et al. (2006).

Hereafter, we will use the word *motif* and ARM interchangeably as long as the context is clear.

VLMD (Nunthanid et al. 2011) was recently proposed to find variable-length motifs in time-series. The algorithm uses exhaustive search of all possible motif lengths and applies an exact motif discovery algorithm at each length. Even though the authors have shown that the algorithm can find a few number of *interesting* motif lengths, the algorithm is not expected to scale well for longer time series due to its exhaustive search strategy. Li and Lin (2010) proposed using the Sequitur (Nevill-Manning and Witten 1997) algorithm for discovering motifs of variable length using grammar inference after discretization using SAX (Lin et al. 2007). This technique can discover variable length motifs but it requires discretization. Another problem with this approach is that a small burst of outliers in a single motif occurrence will result in dividing this motif into two disjoint motifs.

Given the large number of algorithms that were devised to solve ARM, K-motif, and exact-motif discovery problems exactly and approximately, we will discuss a single representative of each of these three approaches.

4.5.3 Symbolization Approaches

Arguably, the simplest approach to ARM discovery is to symbolize the time-series then apply one of the aforementioned discrete motif discovery algorithms to it then report the locations in the original real-valued time-series that correspond to discovered discrete motifs.

Lin et al. (2003) proposed one of the most widely used symbolization algorithms which is known as SAX (Symbolic Aggregate approXimation). SAX has two attractive features for symbolization of time-series with the goal of discovering ARMs. Firstly, it reduces the dimensionality of the time-series considerably. This is of special interest when mining long conversation or interaction data which may contain millions of data points. Secondly, it was proven by Lin et al. (2003) that the distance between SAX representations of two time-series lower bounds the true Euclidean distances between the original time-series. This enables early pruning when searching for motifs and makes the representation appropriate for mining large datasets. SAX was also designed to bias the symbolization process into generating all symbols with equal probabilities.

The first step of SAX is to apply piecewise aggregate approximation (PAA) to reduce the length of the time-series from T to N. As will be clear later, SAX assumes that the time-series to be transformed is zscore normalized. Given an input time-series $X(t)$ we zscore normalize it by subtracting the mean and dividing by its standard deviation $x(t) = (X(t) - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the time-series in order.

Given the zscore normalized time-series $x(t)$ of length T, the corresponding PAA representation (called $\bar{x}(n)$ of length N is obtained by replacing each T/N points of the original time-series with their mean:

$$\bar{x}(n) = \frac{N}{T} \sum_{t=\frac{T}{N}(n-1)+1}^{\frac{Tn}{N}} x(t). \tag{4.4}$$

This piecewise constant representation is then converted into a character string from an alphabet with M characters. The main constraint on this operation is to have roughly equal probability of producing any symbol(character) in the alphabet. Lin et al. have found empirically using 50 different datasets that normalized time-series have highly Gaussian distributions (Lin et al. 2003). To achieve equiprobable production of symbols, the distribution of values in the time-series is assumed to be Gaussian and the Gaussian distribution is divided into M bins with equal probability

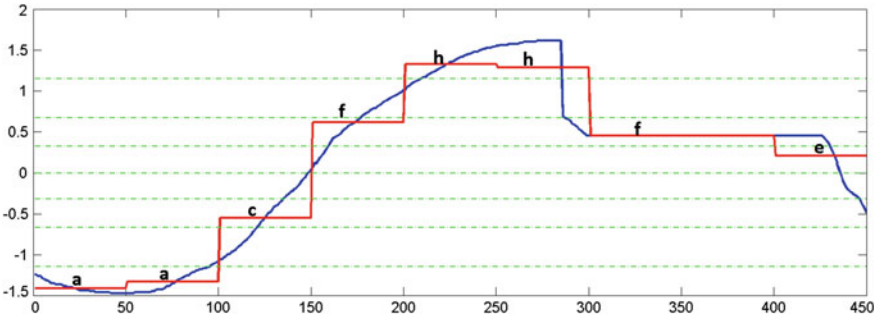


Fig. 4.33 Discretizing a time-series using SAX

at every bin. This can be done offline to generate a set of break points corresponding to every symbol in the alphabet.

Given the PAA representation ($\bar{x}$) and the breakpoints, each point in $\bar{x}$ is mapped into the symbol of the alphabet within which its value falls.

The original SAX algorithm was designed to handle very long time-series. For this reason, the input time-series is divided into subsequences that is each zscore normalized before applying the PAA representation step to it.

Given the SAX representation of any time-series (or any other suitable symbolization algorithm), there are several ways to discover motifs.

The simplest approach is to just apply the PROJECTIONS algorithm explained earlier to the SAX representation of the time series. Chiu et al. (2003) used this approach and it is the basis of several other symbolization approaches since that time (e.g., Minnen et al. 2007; Tang and Liao 2008; Yankov et al. 2007; Canelas et al. 2013, etc). The discretization step of this approach is shown in Fig. 4.33.

4.5.4 Exact Motif Discovery Approaches

A promising approach to solve the ARM problem is to use an algorithm that finds *exactly* the K time-series subsequence pairs (called 2-motifs) of maximal similarity then use them as the basis for discovering recurrent patterns which by definition must have maximal similarity between its pairs. The naive algorithm for solving this problem exactly for a time series of length n and motifs of lengths between l_1 and $l_2 = l_1 + l$ has a time complexity of $O\left(n^2 K l\right)$. This quadruple complexity makes it impractical to apply this algorithm except for short time-series, short motifs, and short motif length ranges.

The simpler problem of exact discovery of the top 2-motif of a given length in a time-series was defined by Mueen et al. (2009) and an efficient exact solution with amortized linear complexity was proposed (called the MK algorithm). This algorithm reduced the amortized time complexity from quadratic to linear which

makes it practical to apply it to moderately long time-series. The MK algorithm uses the Euclidean distance between zscore normalized subsequences as a dissimilarity measure (Mueen et al. 2009).

$$D_{zscore}(x, y) = \sum_{k=0}^{L} \left(\frac{x(k) - \mu_x}{\sigma_x} - \frac{y(k) - \mu_y}{\sigma_y} \right)^2 \tag{4.5}$$

where μ_i and σ_i are the mean and standard deviation of time-series i.

The main advantage of this distance function is that it is offset and scale invariant. It was also shown that it can provide a comparable performance to Dynamic Time Wrapping (Ding et al. 2008).

Mohammad and Nishida (2012b) proposed MK+ which is an efficient extension of MK to discover top K 2-motifs of a given length using the same distance function and showed that it outperforms iterative application of the MK algorithm. MK+ was further extended by Mohammad and Nishida (2014a) (MK++) to discover top K 2-motifs of a range of lengths but assuming that the distance between two subsequences of the time-series cannot decreased with increased length. This assumption is true of the Euclidean distance and Euclidean distance between mean-shifted subsequences but is not true for zscore normalized subsequences. This means that MK++ cannot be used to discover scale-invariant 2-motifs which means that it cannot be a basis for scale invariant ARM discovery.

Recently, Mueen (2013) proposed MOEN for solving the scale invariant version of the problem. The main idea of MOEN is to calculate a lower bound on the distance between any two subsequences at length l given this distance at length $l - 1$. Using this lower bound, it is possible to efficiently discover 2-motifs at different lengths. The goal of MOEN is to find subsequences at which the noise in the time series is minimal at different lengths that can be used to better understand the phenomena generating the time-series.

The MK algorithm finds the top 2-occurrences motif in a time series. The main idea behind MK algorithm is to use the triangular inequality to prune large distances without the need for calculating them (Mueen et al. 2009). For metrics $D(\cdot, \cdot)$ (including the Euclidean distance), the triangular inequality can be stated as:

$$D(A, B) - D(C, B) \leq D(A, C). \tag{4.6}$$

Assume that we have an upper limit on the distance between the two occurrences of the motif we are after (th) and we have the distance between two subsequences A and C and some reference point B. If subtracting the two distances results in a value greater than th, we know that A and C cannot be the motif we are after without ever calculating their distance. By careful selection of the order of distance calculations, MK algorithm can prune away most of the distance calculations required by a brute-force quadratic motif discovery algorithm. The availability of the upper limit on motif distance (th), is also used to stop the calculation of any Euclidean distance once it

exceeds this limit. Combining these two factors, 60 folds speedup was reported by Mueen et al. (2009) compared with the brute-force approach.

The inputs to the algorithm are the time series x, its total length T, motif length L, and the number of reference points N_r.The algorithm starts by selecting a random set of N_r reference points. The algorithm works in two phases:

The first phase (called hereafter referencing phase) is used to calculate both the upper limit on best motif distance and a lower limit on distances of all possible pairs. During this phase, distances between the subsequences of length L starting at the N_r reference points and all other $T - L + 1$ points in the time series are calculated resulting in a distance matrix of dimensions $N_r \times (T - L + 1)$. The smallest distance encountered (D_{best}) and the corresponding subsequence locations are updated at every distance calculation. This D_{best} value provides an upper limit on the distance between the two occurrences of the best motif. The reference point with the highest variance in distances is then selected and the subsequences are ordered by their distances to this point. This phase is clearly linear in both the length of the time series, the number of reference points, and motif length.

The second phase of the algorithm (called scanning phase) scans all pairs of subsequences in the order calculated in the referencing phase to ensure pruning most of the calculations. The scan progresses by comparing sequences that are k steps from each other in this ordered list and use the triangular inequality to calculate distances only if needed updating D_{best}. The value of k is increased from 1 to $T - L + 1$. Mueen et al. (2009) showed that this ensures that all subsequence pairs are considered and that each is considered exactly once. Once a complete pass over the list is done with no update to d_{best}, it is safe to ignore all remaining pairs of subsequences and announce the pair corresponding to D_{best} to be the *exact* motif. This typically happens with small values of k resulting in an amortized subquadratic complexity.

The algorithm also receives a parameter (wMO) representing the allowed fraction of overlap between pairs that can constitute 2-occurrence motifs or *within-motif overlap*. In mining conversational records, we can safely assume that this value is zero allowing no overlap between pairs but enforcing no minimum delay between their appearance in the time series. This choice was dictated by the nature of human interaction logs where recurrent activities are ubiquitous and the motifs we are after are only useful when they are disjoint.

4.5.5 Constrained Motif Discovery Approaches

Constrained motif discovery is a final approach to motif discovery that is of special interest to conversational informatics because it allows the integration of information from domain-knowledge and behavior of the interaction partners into the discovery process. Examples of this approach were given by Mohammad and Nishida (2009a). In this section, we only provide a simplified version of this approach.

A constrained motif discovery (CMD) algorithm is a motif discovery algorithm that tries to first find possible locations of motifs using another technology (for

example change point discovery as will be discussed in Chap. 9) then applies motif discovery to the subsequences near expected motif locations.

A simple CMD algorithm is given in this section. The algorithm has four arguments, namely, the input time series, the input constrained, the minimum length of a motif and the maximum length of a motif. The motif length limits need not be tight but choosing very small minimum motif length can lead to slow operation. The CMD algorithm goes as follows:

1. Find the optimal threshold (T_{cons}) of the constraint input over which the corresponding time series is considered to have a candidate motif occurrence by using the L method:
   (a) Apply a thinning operation to the constraint to keep only local maxima.
   (b) Sort the thinned constraint series.
   (c) Find the best two-lines fit that minimizes the sum of squared errors.
   (d) The constraint value at the intersection of these two lines is considered the optimal threshold.
2. Find the points in the time series at which the constrained is larger than T_{cons}. The list of subsequences of length *minimum_length* that end at these points is called C. If the available space is limited only a subset of this list can be used in the remaining steps of the algorithm
3. Build a full distance matrix between members of C. Any distance measure can be used here. We use $1 - cos\,(\theta)$ where θ is the angle between the subspaces representing the largest l Eigen vectors of the Hankel matrix associated with the subsequence.
4. Find the best distance threshold to distinguish near and far subsequences in the list C by using the L method again. This threshold is called T_{dist}.
5. Construct a graph from the distance matrix after making all entries greater than T_{dist}. Each clique in this graph represents a stem of a motif type.
6. For every motif stem try to extend the motif occurrences by adding one point at the time from the time series before and after the members of the motif stem until the variance of the time series values at the next point is larger than the average variance at every point of the motif stems or the *maximum_length* limit is reached.
7. For every extended motif stem scan the original time series to detect all other occurrences of the motif.
8. Combine motifs the occurrences of which are always coming together in the same order.

As described, the CMD algorithm is a single dimension motif discovery algorithm. To use it with multidimensional data the algorithm is applied to every dimension of the data. The resulting motifs in different dimensions are then combined if the pearson correlation coefficient between their occurrences exceeded a threshold. This technique is different than the only available subdimensional motif discovery algorithm described by Minnen et al. (2007) in that it does not use early detection of *distraction* dimensions and this allows it to discover multidimensional motifs with overlapping occurrences.

4.6 Evaluation

Evaluation is indispensable to establish a solid understanding technology. Conversational systems are not easy to evaluate because they are developed from both exploratory and empirical studies where vague questions and hypotheses are refined in a trial-and-error fashion (Cohen 1995). Conversational systems can possess several unique and highly abstract aspects, such as facial display, anthropomorphism, embodiment, communication modalities, personality, emotions, and sociability, for which evaluation methods have not necessarily been established. Evaluation can also be influenced by the user's background, such as culture, gender, age, ethnicity, language, education, computer skills, and familiarity with artificial intelligence technologies. Consequently, evaluations must be carefully managed. Therefore, it is necessary to combine multiple approaches to draw reliable conclusions from evaluation data.

Generally, evaluation criteria can be classified as usability and user perception (Ruttkay and Pelachaud 2004). *Usability* is related to task performance and can be investigated in terms of learnability, efficiency, and error. We can regard conversational agents as a type of intelligent user interface and rank them according to the ease in learning them, the efficiency with which the user can achieve a predetermined goal, and the number of errors expected while using them. On the other hand, *user perception* is concerned with the way a user perceives the conversational system. User perception may be judged with respect to satisfaction, engagement, helpfulness, trust, believability, likeability, and entertainment. In contrast to usability, user perception appears more subjective and is likely to depend on individual user backgrounds. Consequently, a method that relies on self-reporting, such as a questionnaire, to measure user perception may be unreliable unless it is combined with other methods, such as video analysis or physiological measurements.

In addition, an effectively designed evaluation may need to consider the specifics of the evaluated procedure. Although a literature study might be useful to evaluate conversational system technology in general, an empirical study based on data collection and statistical analysis will result in much deeper understanding. However, empirical studies are generally restricted to a very narrow scope and sufficient care should be taken to ensure that the analysis is properly controlled. Data collection includes qualitative and quantitative methods. Qualitative methods include interviews and focus groups, and informal or descriptive observations; quantitative methods include questionnaires, systematic observation, log files, heuristic evaluation, and biological measurements.

Although the questionnaire method is relatively inexpensive, the data can be biased. For example, bias can result from participants' tendency to provide what they believe to be the desired feedback rather than their own actual perceptions. Obtaining data from physiological sensors or measuring subconscious attitudes may relax self-reporting biases. The *implicit association test* (IAT) (Greenwald et al. 2003) is a method that estimates an *implicit attitude* on a concept by measuring priming time or latency time that are assumed to be caused by the attitude. IAT is based on the

accessibility effect, i.e., if the participant has access to a word before a task, access speed to the word and related words will increase (Fazio et al. 1986).

During the IAT procedure, each participant is repeatedly shown a stimulus word in the center and two or four words on the top left and top right corners on the screen and asked to press the "left" or "right" button as quickly as possible, depending on his or her rapid judgment on which of the words shown on the top corners is closer to the given stimulus (Fig. 4.34). The shorter the average response time of the participant is for one combination of a concept and an attitude represented by a collection of words or representative images, the stronger association the participant is estimated to have for it over the contrasted combination. For example, if the participant exhibited a shorter average response time for displays for a combination "flower and pleasant" than those for "insect and unpleasant", he or she would be estimated to have a more positive attitude towards flowers than insects.

Occasionally, it is desirable to build a model that captures the essence of a given behavior to better understand the collected data. Linear regression uses a predictor variable and an estimated linear function as an approximation of the observed value. This allows researchers to examine the value of a coefficient to determine the amount of contribution of the predictor variable to the resulting observation. Multiple linear regression uses more than one predictor variable to account for the observed value.Structural equation modeling (SEM) may be useful if one is interested in the causal dependency of variables that represent key features of a given system (Cohen 1995).

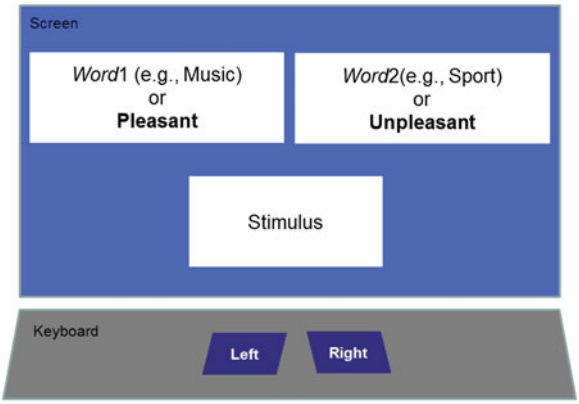


Fig. 4.34 Implicit association test in action. Drawing inspired by Greenwald et al. (2003)

4.7 Summary

In this chapter, we have discussed methodological aspects for conversational system development methodologies. We shed light on the technical aspects of existing techniques, and highlighted five aspects: the architecture, scripts and markup languages, corpus-based approaches, behavior learning using machine learning techniques, and the evaluation methodology. As for the architecture, identifying major components of conversational systems and their organization is the key. Insights from existing systems allows us to make a rather comprehensive list of components that may be arranged in a hierarchical fashion depending on the level of abstraction. Meanwhile, a generic control structure such as the blackboard system is mandatory to support a complex structure of invocation and coordination among components. As for the scripts and markup languages, general requirements include easy specification of the behavior of conversational systems. Although scripts are procedural, markup languages are declarative. Markup languages allow for partial specification of behaviors, which may be easy for system builders while increase computational cost. BML and FML resulted from a standardization forum to allow researchers to share data and codes. A corpus-based approach realizes a data-driven approach and bases target behaviors of conversational agents on varieties observed in existing conversations. The impact of a corpus-based approach depends on the quality of a corpus and resulting model. Machine learning techniques may be applied to corpora to generate interactional behaviors of conversational agents. As social signals in conversation are mostly captured as real-valued data, motif-discovery algorithms either consist of symbolization and discrete motif discovery or directly work on time series consisting of continuous values. As for evaluation of conversational systems, major aspects involve usability and user perception. Techniques such as the implicit association test allows for estimating participants' implicit attitudes in evaluation.

Chapter 5
Conversation Quantization

Abstract Conversation quantization is a conceptual framework used for capturing and reusing shared meanings and expressions in a conversation. As a generic framework, it encompasses different implementations ranging in granularity, depth and breadth of annotation, representational fidelity, and generality. In this chapter, we discuss the scope and requirements of conversation quantization, the range of basic functions necessary for individual implementation, and the space of potential implementation. We also sketch out the idea of a portable conversation space for maximizing the potential of conversation quantization.

Keywords Conversation quantization · Representation scheme · Production/Consumption scheme · Circulation scheme · Shared conversation space

5.1 Framework of Conversation Quantization

A conversation is essentially a local phenomenon that entails a focused interaction within a gathering of individuals to discuss selected issues. Although certain issues may be prepared in advance by some of the participants, more often, a majority are improvised on the spot unless the conversation structure is strongly controlled or stylized. Conversation is useful for our intellectual life, not just as an opportunity for exchanging information and knowledge, but also because new and unanticipated meanings, interpretations, aspects, and concepts often arise suddenly during the course of the conversation.

How then does a new meaning come to bear in a conversation? When it is generated during the interaction, either accidentally or on purpose, and is noticed by one of the participants, he or she may propose to take it up by commenting on it or, more subtly, by emitting a small back channel signal and temporarily embracing it. Sustained meanings may manifest as behaviors that can be perceived by other participants so that they can be shared for criticism, or contribute additional meaning. Although meanings may arise when we are alone, they are so subtle and fragile that they dissipate and eventually disappear in the sea of entropy unless they are meticulously recorded. By contrast, a conversational setting allows participants to collaboratively

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_5

filter out useless inputs so that each participant can single out his or her own most meaningful insight from the conversation and express it optimally in a language.

Consider the hypothetical conversation at Waikiki Beach depicted in Fig. 5.1. In this example, facts such as "Diamond Head is a volcano," or "C went up there with his or her family in 1985" may have remained untold had A, B, and C not come across each other and engaged in conversation. Even though B's knowledge is trivial from her/his own perspective, this might not have been so for A or C. C's experience of Diamond Head might have not been a shared one with A or B, and this is clearly why C made that utterance in the first place. B's and C's shared utterances were derived from A's utterance in the conversation.

Unfortunately, a conversation's utility in terms of its contribution to individual and collective intelligence is severely limited, as conversation is often volatile. The memory of a new meaning that manifested in a conversation is often only retained in the minds of the participants, and, as time elapses, becomes distorted, dissipates, and fades out.

Although such properties may be advantageous in an evolutionary sense for selecting only useful memes, the process of diffusing natural information is often inefficient and time consuming. New meanings might eventually reach people who develop a need for them only after many generations of repeating similar conversations on the subject. As a result of this inefficiency, some people even regard conversation as being irrelevant to the development of organizational knowledge.

Traditional methods of preserving expressions and meanings are either to keep talking or to record them. Unfortunately, both of these methods are expensive as they require nontrivial amounts of time and labor, and hence, are not widely practiced. Recent advances in social media have resulted in drastic improvements by significantly decreasing the cost of spontaneously recording and distributing meanings on the fly. In addition, search engines greatly contribute to the content market.

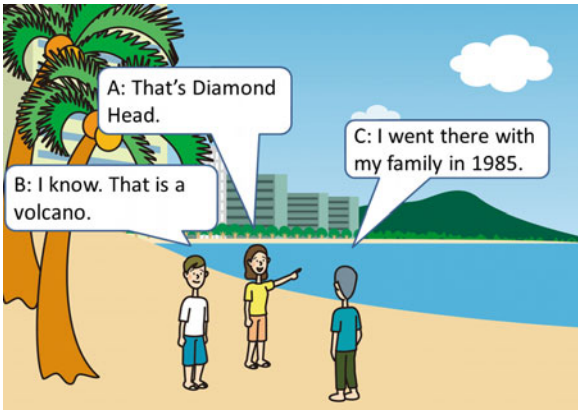


Fig. 5.1 Talking about Diamond Head on Waikiki Beach. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

However, current social media are still quite limited in scope as a device for capturing and reusing meaning. They require the conscious efforts of participants, and their expressive power for capturing meanings is limited to texts together with audio-visual records. Moreover, their reusability is limited.

In the Waikiki Beach example illustrated above, the conversation could be recorded as text in which reference to the physical star would obviously be lost. Thus, the reader would have to recover this reference, based on his or her knowledge of associated linguistic expressions with the referent being the physical environment. The inclusion of accompanying photos, diagrams, or other means of facilitating the search for the star in the sky would alleviate difficulties in recovering the physical reference, depending on the referential aid that is made available.

It is also important to record as many circumstantial cues as possible which may contribute to replicating the process of meaning emergence. Examples include aspects of the physical atmosphere such as temperature and wind, the audio-visual scene, each participant's verbal and nonverbal expressions, their first-hand perceptions of the environment, or even their physiological states, together with precise time and geological stamps.

Conversation quantization is a conceptual framework that can significantly contribute to enhancing the evolutionary process of conversation circulation within a community. Specifically, by capturing shared meanings and expressions in conversation, it may provide a basis for subsequent intellectual development. Figure 5.2 provides a bird's eye view of how conversation quantization operates in practice. There are four main ideas underlying conversation quantization.

The first entails the use of an information package known as a *conversation quantum*. This encapsulates as much information as is relevant to the meaning and expressions of a significant segment of the conversation. Both meanings and expressions are retained as only one of these falls short in replicating the original conversation. A conversation quantum retains information as a representation, which is complementary to the way information manifests as interaction within conversation. Thus, conversation quantization encompasses the reciprocal transformation of information between the processes of interaction and representation. Materialization refers to the process whereby conversation quanta are created for interaction. It essentially entails converting the interaction into a data structure that substantiates the interaction. Dematerialization, however, refers to the reverse process of creating an interaction based on conversation quanta.

The second key idea is that a long conversational series can be divided into one or more segments with associated conversation quanta required for its reproduction. We can assume that some kind of schematic system is employed as a guide on how information is collected and packaged for a given conversation scene. Thus, our approach can be partly schema-based, and neither purely bottom–up nor restricted to data-recording. The concept might become fully automated in practice if the schematic system can be made fully computational, or if a high quality approximation can be implemented. By contrast, we might well be able to build a semi-automatic system in which the user supplies part or all of the schemata.

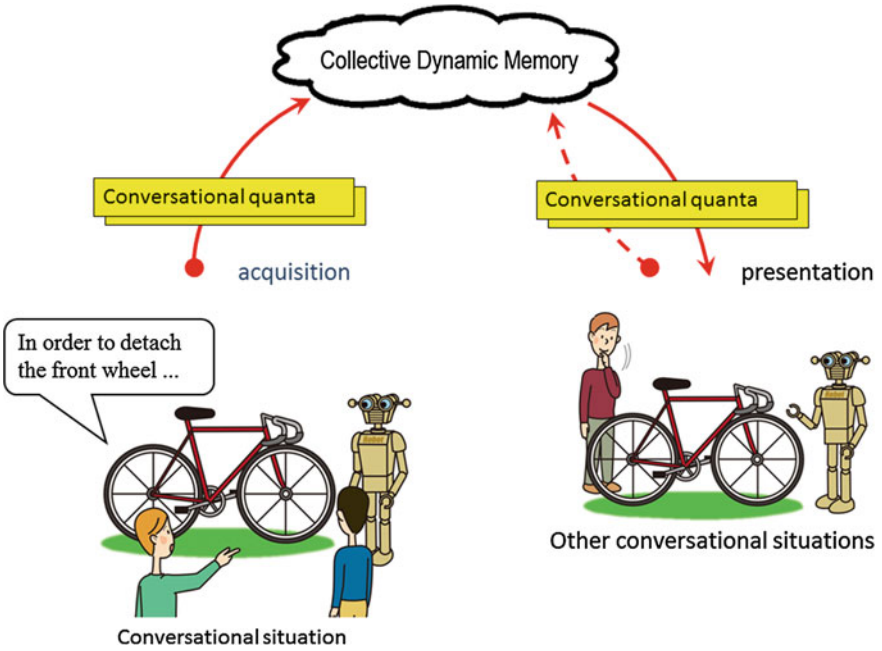


Fig. 5.2 Conversation quantization as a conceptual framework for evolutionary conversation circulation in community. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

The third key point is that once one or more conversation quanta are obtained from a conversation scene, they are subjected to manipulation and distribution for replication. The user may search for interesting items from a market of conversation quanta and use the selected conversation quanta for his or her own purpose. Moreover, the user may be interested in creating his or her own collage of conversation quanta and displaying this in a gallery according to his or her story. We envision a community of prosumers in which each user not only consumes conversation quanta, but also produces new conversation quanta by compiling and editing conversation quanta from many sources to create something altogether new.

Last, full automation might be ideal from an artificial intelligence (AI) perspective and useful if realized. However, it is not mandatory for a data-intensive approach, providing that the entire process is effective and contributes to an increase in the primordial soup of conversation. Indeed, professionals may often voluntarily contribute to manual production of conversation quanta that represent their intellectual assets. In educational environments, school teachers might be interested in creating their own course materials in the above way. Students may compose notebooks of their course materials, related materials, and their personal notes. Local people might be interested in authoring local guides. A product manager might create a user guide. A designer might create test case scenarios, and a politician might formulate a policy by editing citizens' voices.

Conversation quantization theory entails five important aspects. First, it specifies a representation scheme for describing a conversation scene. It not only enables the recording of a conversation scene for reproduction, but also for using a description for multiple purposes. These may include extracting particular kinds of information according to the conversation structure, or even its use as a source of knowledge by a conversation agent for producing utterances in similar conversational situations. Second, conversation quantization specifies a production/consumption scheme for producing conversation quanta from conversations, and the reverse situation using conversation quanta to produce agents' interactional behaviors within conversations. We will discuss the cognitive procedures involved in the production and consumption of conversation quanta. Third, conversation quantization involves a manipulation scheme which specifies basic functions that allow for the combination of one or more conversation quanta to create a new conversation quantum. We will examine how a large accumulation of conversation quanta constitutes a shared knowledge medium. Fourth, conversation quantization implies a circulation scheme, suggesting how a collection of conversation quanta may evolve as a result of circulation within a community. We will consider how the notion of dynamic memory may be applied to learning and organizing a large chunk of conversation quanta. Finally, conversation quantization provides engineers with the possibility of designing a conversationally enhanced space whereby users not only leverage a dense collection of conversation quanta, embedded in the environment, to acquire their predecessors' wisdom in an interactive fashion, but also contribute to the future.

The use of conversation quantization depends on such features as:

- Granularity: This refers to the number of conversation quanta required to describe the discourse. The greater the granularity of conversation quanta, the easier it is to reproduce the discourse while simultaneously reducing its reusability.
- Depth and breadth of annotation: This refers to the quantity of semantic annotation required to annotate the situation. The greater the amount of available annotation, the more reusable it is, while at the same time being more expensive.
- Representational fidelity: Often human communication signals cannot be fully restored for many reasons such as noise or error. The greater the fidelity, the better the quality, while also increasing the expense.
- Generality: This refers to how much abstraction is required to describe the discourse. The greater the generality, the greater its reusability, while increasing the cost of content production.

The extent to which we can actually record and replicate the meanings and expressions of a conversation depends entirely on the available technology. We, therefore, consider a generic conceptual framework encompassing the space of potential representations rather than a specific one which is highly technology-dependent. The designer needs to know the trade-off between the use and the cost of designing an actual conversation system. Conversation quantization conceptually defines a design space for capturing and reusing meanings arising within conversational settings.

In general, there is a trade-off between the costs of acquisition and reusability, and the quality of interaction. If we wish to decrease the acquisition cost, then reuse

of the content tends to become expensive. Conversely, to reduce the cost of reuse, we have to invest in acquisition, especially when the quality matters. Although using deep semantic representation might be thought of as ideal for increasing reusability, usually semantic representation, as a source of nonverbal behavior, does not embody complete information. Thus, additional care should be taken to assure the quality of interaction. As such, the most reasonable resolution might be to adopt an annotation-based approach that will preserve the original expression of interaction while making additional semantic information available for reuse.

In the following discussion, we focus on the four aspects of conversation quantization based on a hypothetical annotation-based approach.

5.2 The Representation Scheme

The role of conversation quantization is to provide a conceptual means for representing the expressions and meaning that arise in a conversation. Intuitively, conversation quantum is expected to carry adequate information to reproduce the original interaction as representationally equivalent to interactions in a conversation scene. In addition, the representation should be generic enough to be applied to conversation scenes that differ from the original scene in which the conversation quantum was created, and amenable for information extraction and manipulation. Although it appears to be impossible to devise a perfect representation that permits a complete reproduction of the original conversation scene, it is possible to encode key conversational aspects to enable the reproduction of essential information.

Basically, a conversation quantum represents an ideal observer's record of conversation, as illustrated in Fig. 5.3. In other words, we do not assume a purely objective description of a conversation scene, as it is practically impossible to obtain such a description, even with precise measurements. This is because we would also like to include a participant's mental state which is often mandatory in describing unobservable meanings and expressions. We believe that a purely objective description of a conversation scene is not only infeasible but also useless. Even in a situation wherein one or more participants voluntarily transcribe the conversation scene, we would ask the volunteer to pretend to be a third observer to externally describe the conversation scene, possibly with reference to her/his own mental state. This design decision was made to retain the reusability of the conversation quantum. However, it should be noted that even as a result of this design decision, it is possible to reproduce the conversation scene from a subjective, first-person viewpoint.

A conversation quantum package provides representations of the *ground*, *discourse*, and *interaction*. The ground consists of descriptions about entities and of participants in particular that are explicitly or implicitly referred to in the conversation. In the Diamond Head example, the ground section of a conversation quantum may contain descriptions about the participants, the Diamond Head as a referent, and other potential referents such as the Waikiki beach, the palm trees, hotels, beach people, and others. The resolution or granularity of the ground section determines

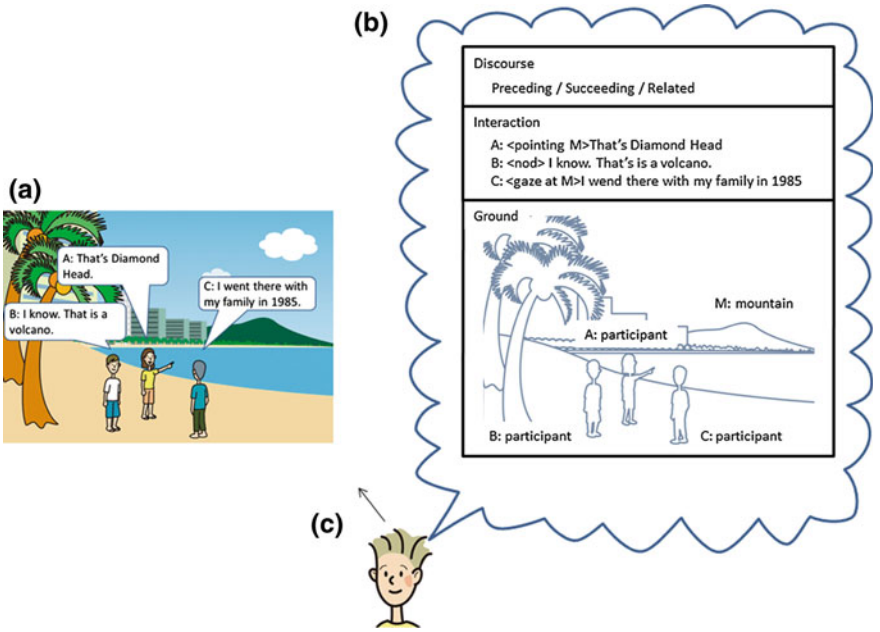


Fig. 5.3 Using conversation quantum to describe a conversation scene. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission. **a** Conversation scene. **b** Conversation quantum. **c** (Hypothetical) observer

the quality and quantity of the situational information to be incorporated in a conversation quantum. The finer the resolution of the ground section, the more situated the conversation quantum.

The discourse section describes how the given conversation quantum is related to the conversation scenes, and to their elements described in other conversation quanta. Cues for co-textual references, based on endophoric relations as opposed to exophoric relations, are described in this section. The user will, however, need to look for a referent in other conversation quanta. This formulation allows us to use conversation quanta as building blocks used in multiple discourses despite an increase in the overhead.

Embedded discourse can be represented using the discourse section. We use one conversation quantum to represent one conversation scene. When a conversation scene embeds another referential conversation scene, we instantiate a new conversation quantum for the embedded conversation scene. For example, in the conversational scene shown in Fig. 5.4, P is talking about his dream for which a conversation quantum D2 is instantiated and referred to from the main conversation quantum, D1.

Rhetoric may be partly represented in a similar fashion. Figure 5.5 shows a conversation scene in which a girl pretends to use a banana as a phone to talk with her mother. In this case, a conversation quantum is allocated for her imaginary world and another for the content of the talk over the imaginary phone. The contrast between

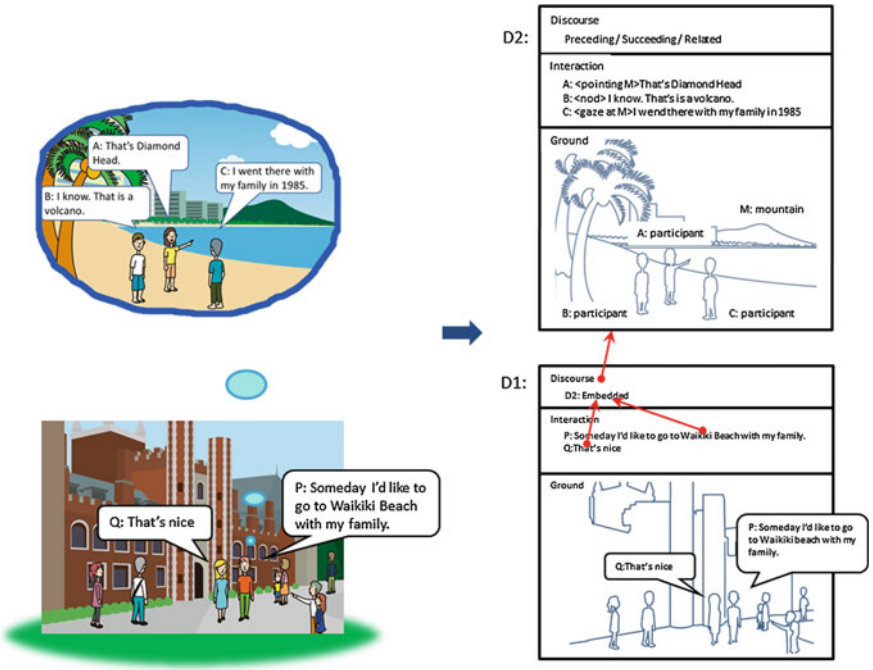


Fig. 5.4 Conversation quantization for an embedded discourse. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

the physical banana and the imaginary phone is made in the conversation quantum for the main discourse.

The interaction section contains the transaction record of the conversation scene. Although the transcript of verbal interactions might be a minimal requirement, it is highly desirable to include the transcript of nonverbal interactions and to associate them with verbal expressions. This is because they are deemed an integral part of the interaction and cannot be recovered solely from textual information.

It should be noted that the above discussion constitutes a kind of a textbook standard. Appropriate descriptive levels in each constituent may and should vary depending on what is needed and available in the actual implementation. Conversations that occur on the spot and refer to entities in the environment around the participants appear to be rare. It is more often the case that the conversations are secondary and detached from the original referents. Participants are more interested in the narratives or logical structure of events and relations, or even in the maxims they have heard from other people or read in books, magazines, or whatever sources, than in how those stories are grounded in the history of the real world. It is probable that as with listening to music, participants are doing so to look for a story that they may use to entertain themselves or their friends. Second-hand stories are an effective means of sustaining a conversation to maintain a relationship with colleagues, as evidenced by Schank (1990).

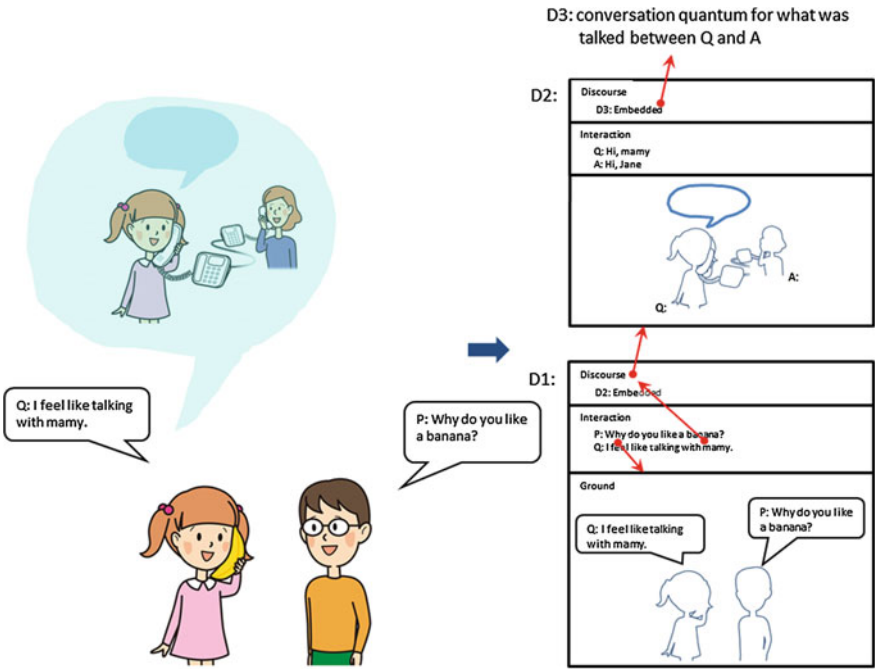


Fig. 5.5 Conversation quantization for a metaphor. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

If conversation is just for exchanging stories, the grounding part and nonverbal expressions used to associate expressions with environmental referents may be somewhat redundant. The discourse section may be strong enough to endow the conversation agent, who makes use of conversation quantization, with ample talent as a story teller.

A different situation arises if the content of the conversation quantum is used spontaneously, or even by an intelligent robot, as a source of knowledge to achieve a mission through the conduct of some meaningful action. In this case, all the details needed to recognize entities in the environment, and to perform and monitor an action to affect the environment, should be grounded. Nonverbal cues that are used to refer to entities in the environment might be reused to recognize or produce communicative behaviors.

5.3 The Production/Consumption Scheme

How can we generate and use conversation quanta in a conversation? The answer depends on a technical choice. We emphasize the balancing of the production and consumption costs of conversation quanta. As we have selected an annotation-based approach, acquisition is mostly related to segmentation and semantic annotation,

which are not inhibitive if not easy, while the cost of reuse depends on the quality of segmentation and semantic annotation.

Manual dictation, either by one or more participants or by an auditor, might be feasible provided that a usable tool is available for auditors. Alternatively, we can directly create conversation quanta for consumption even without conducting conversations, similar to tweeting on Twitter. Although dictating may often be a painstaking task, except for some special occasions, converting either short or long texts into conversation quanta is not overly difficult provided that grounding is not required, as we discuss in the next section. In this section, we look in more detail at materialization and dematerialization, or the situational creation and use of conversation quanta.

Our conceptual approach assumes various hypothetical devices for describing what needs to be done either to produce conversation quanta from conversations, or to reproduce conversations from conversation quanta.

We consider how a conceptual device, known as a*schema*, plays a central role both in producing and consuming conversation quanta. A schema serves as a template or prototype to identify salient features in the environment. In the Diamond Head conversation, a schema might be considered a "conversation at a scenic place" wherein participants are at a location from which points of interest might be visible (Fig. 5.6). The scheme might also contain people other than the participants who just happen to come by. The *dictionary of schemata* may contain more or less specific schemata for a given situation.

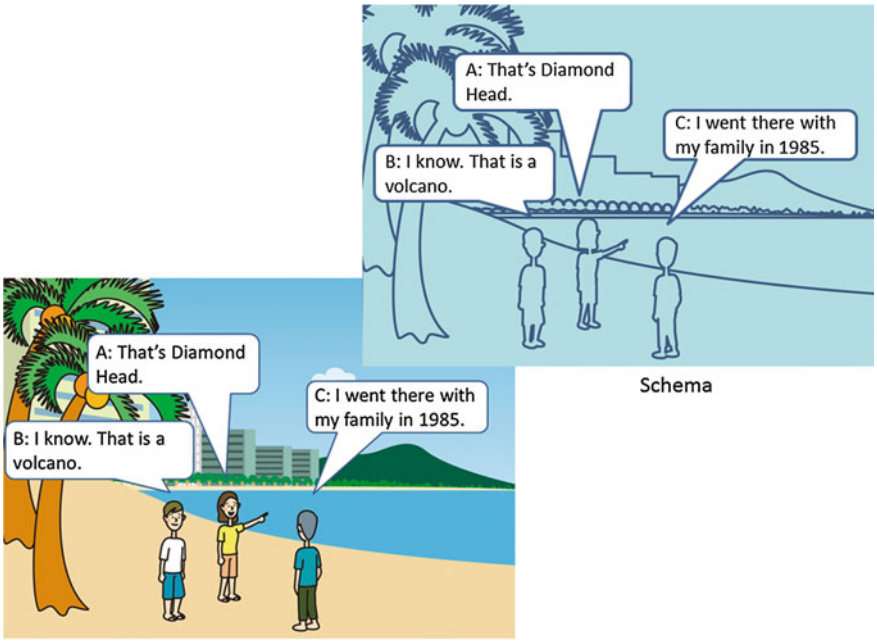


Fig. 5.6 Using schema to generate a ground. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Figure 5.7 shows how either a natural or artificial agent might produce a conversation quantum. A natural, that is, human agent is expected to have already acquired a rich dictionary of schemata, and to, therefore, be able to properly articulate the given conversation scene using a schema to produce a conversation quanta. Humans might innately possess *schemata-based recognizer* that matches an appropriate schema to a given conversation scene, and applies it to ascertain how each component of the schema is instantiated in a given conversation scene. During the segmentation phase in what may be a long conversation record, meaning segments will be identified and the conversation quantum that has been created will be incorporated into a network of conversation quanta.

Regarding consumption, the most straightforward use of conversation quanta is simply to reproduce recorded conversation scenes. Conversation quantization, in contrast to a simple recorded conversation, accommodates a certain level of flexibility. This may include a change in viewpoint, for example, from that of an auditor to that of a participant in the conversation scene, employing only part of a conversation scene for reproduction, or even editing the content of the conversation.

More advanced usage of conversation quanta entails producing the behaviors of the conversing agents. In this case, the conversation agent takes on a role described in a given conversation quantum. The hypothetical devices required for this process are a schemata-based recognizer, a dialogue manager, and an agent controller. The role

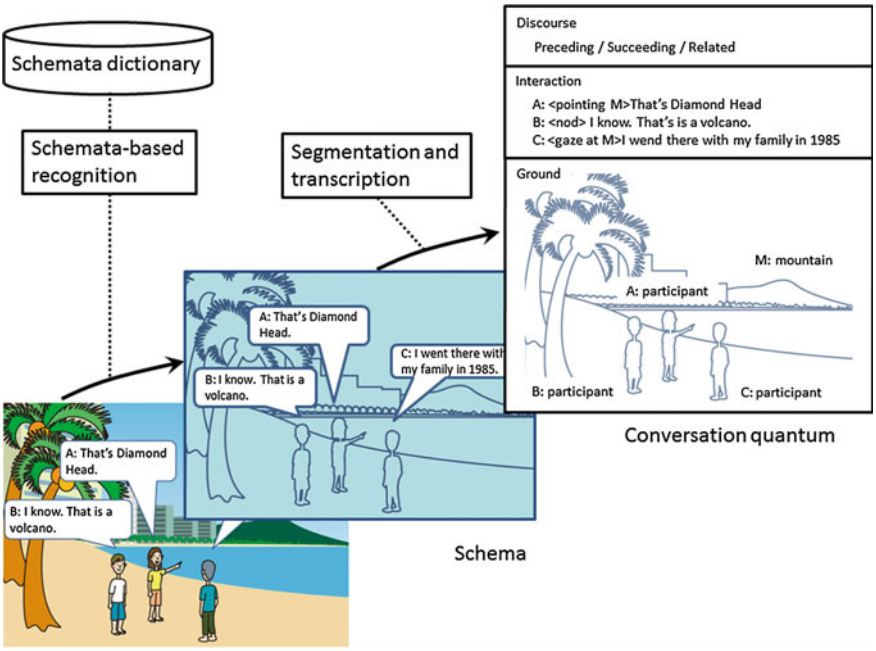


Fig. 5.7 Producing a conversation quantum. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

of a schemata-based recognizer is to search within the environment for a referent of an exophoric reference and for salient features noted in a given conversation quantum. A dialogue manager uses the elicited schema and a conversation quantum to generate the communicative behavior of a conversation agent. The content stored in the conversation quantum needs to be adapted to the current situation.

For example, only a relevant portion of a conversation quantum may be extracted. Figure 5.8 illustrates this process whereby robot Q's utterance is produced in response to P's previous utterance, which itself is a response to the same utterance in the same situation stored in the conversation quantum. The role of an agent controller is to steer the body of the conversation agent to reproduce the communicative behaviors described in the conversation quantum.

Automating the process described above requires the overcoming of several challenges. While the development of a dictionary of schemata and a schemata-based recognizer remain highly challenging, and are yet to be achieved, they are among the most effective solutions. However, in the current era of "big data," it may not be long before we witness their first implementation, probably through the application of pattern-finding algorithms on a large collection of instances. According to our research strategy, it is likely that we can collect a large amount of useful conversation data from a strategically developed primordial soup of conversation as a means of realizing conversational intelligence.

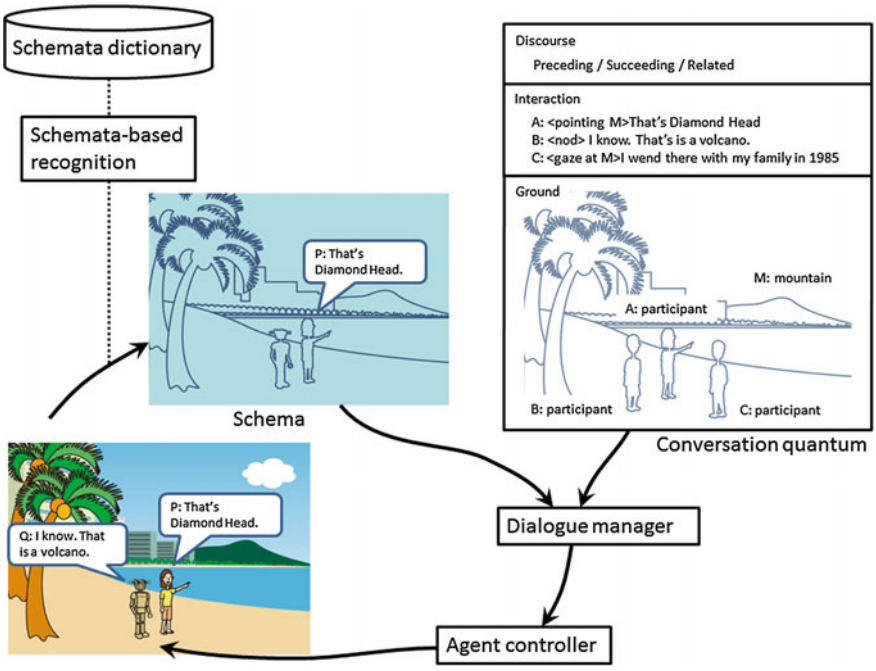


Fig. 5.8 Consuming a conversation quantum. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Creating an audio-visual recognizer for recognizing objects and speech is another challenging problem. However, it should be noted that even without an object and speech recognizer, which appear to be necessary for the task, we might be able to heuristically approximate this task, with possible assistance from the user. Schemata might be sought using heuristic cues such as a GPS or auxiliary cues such as a keyword in a schedule book. We might find a referent by combining gestures and salient features in the image. Moreover, we could simply make an audio recording of speech and use a keyword-spotting technique to recognize critical words. Even with a recognizer that has limited capabilities, it may still be possible to produce conversation quanta of moderate quality.

From time to time, the speaker may create an imaginary gesture space in front of him or her and use gestures, as shown in Fig. 5.9, to visually illustrate the referent. An automated system would need to recognize such verbal–nonverbal behaviors for depicting and referring to an imaginary object in a gesture space to produce an associated representation in a conversation quantum. It should simultaneously be able to generate appropriate behaviors of a synthetic character for this type of expression using a gesture-space, if this is judged to be effective during the conversation. There are further challenges involved in producing and consuming conversation quanta for higher-order conversational behaviors such as metaphors and jokes.

Although full automation of a conversation quantizer might be useful, its implementation may not be advisable until the required technology has matured further. At this point in time, a semi-automated or human-assisted conversation quantizer may be equally useful for building a useful enterprise as an integral part of a primordial soup of conversation.

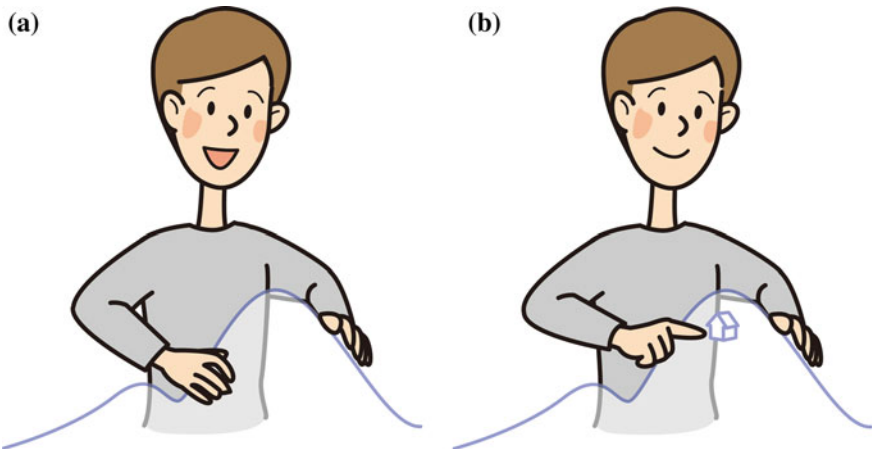


Fig. 5.9 Recognizing gesture space. **a** The speaker is talking as if there were a mountain in front. **b** The speaker pointed a small shrine on the imaginary mountain. © 2014, At, Inc. Reproduced with permission

5.4 Manipulation Scheme

The biggest advantage of materialization is that it allows for the changing of stories that spontaneously arise in a conversation into something that is external to the participants and can, therefore, be manipulated not only by them, but also by other people, and even by artificial intelligence. The manipulation scheme specifies basic functions that permit combinations of one or more conversation quanta to create a new conversation quantum. The structured representations of information that are stored in conversation quanta provide a handle for easily editing the content, possibly by artificial means.

There are two categories of operations on conversation quanta: syntactic and semantic. Syntactic operations are derived from the structure embedded within conversation quanta. If sufficient information is included in the ground section of a conversation quantum, the spatial, temporal, and social relationships among the entities mentioned in a conversation quantum can be obtained instantaneously by following a prescribed procedure for analyzing the structure of the representation. The converse applies when inadequate information is provided in the ground section. An example of this is when the user of a conversation quantum has to conduct a semantic operation to integrate available cues and available methods of inference and conjecture to obtain tacit information. This severely constrains the automatic manipulation of conversation quanta and hence their use. However, this may be unavoidable because of our limited knowledge and constraints in the real world that do not allow for information representations that fulfill all possible requirements.

Thus, at this stage of development, we are yet to identify syntactic operations on conversation quanta beyond those that depend on low-level meta information such as time, place, and potential referents in the ground section, annotated information in the interaction section, and co-textual links in the discourse section. Although these operations are very basic, we suspect that there are no other purely syntactic operations.

Other operations are highly semantic. Although their execution depends on human intelligence, we cannot expect perfect results. Powerful methods of intelligent assistance and experiential learning are available for future research. However, it may be useful, to some degree, to discuss a list of desirables at this point in time.

Concatenate and *partition*. Whereas the *concatenate* operation connects one or more conversation quanta to form one large conversation quantum, the partition operation divides a given conversation into multiple conversation quanta.

Summarize and *elaborate*. Although similar to the *concatenate* and *partition* operations, the *summarize* and *elaborate* operations require further semantic processing. For the *summarize* operation, critical expressions need to be extracted from the whole, or a lengthy expression may even need to be replaced by a more concise one. For an *elaborate* operation, on the other hand, new information is added to flesh out a given conversation quantum.

Revise. The *revise* operation replaces the content of a given conversation quantum by one that is considered better by the operator.

Embed. This operation embeds a conversation quantum as an utterance of a participant in the conversation.

Merge and *split*. The *merge* operation converts dialogues into monologues, whereas the *split* operation does the reverse. For example, a conversation can be converted into a monologue through a *merge* operation by compressing the utterances of multiple participants into one utterance, as shown in Fig. 5.10. By contrast, through a *split* operation, a monologue can be converted into a dialogue by editing linguistic and non-linguistic expressions.

Although full automation of the above operations may be both challenging and attractive from the viewpoint of artificial intelligence research, we believe that motivated human involvement in the process, and an extensive accumulation of effective conversational quanta, may essentially lead to enhanced quality of conversation quanta through the incorporation of the various contributions and interpretations of human participants.

The Sustainable Knowledge Globe (SKG) (Kubota et al. 2007) is a visual tool used for editing a large collection of conversation quanta. Here, it is important to support the entire life cycle of the accumulated conversation quanta, rather than merely helping the user to manipulate a large amount of data. An extensive collection of conversation quanta is clearly not produced in an instant. Rather it is accumulated over a long period of time. In the beginning, the collection might not have a structure. However, conversation quanta may subsequently be grouped together, based on certain similarities, to constitute a cluster such that the entire collection might be composed of a number of clusters. A structure may then emerge for organizing the clusters or their content in a hierarchical, ordinal, or other manner. SKG enables the user to visually grasp a relatively large collection of conversation quanta by means of a landscape. By zooming out, the user may grasp the whole landscape

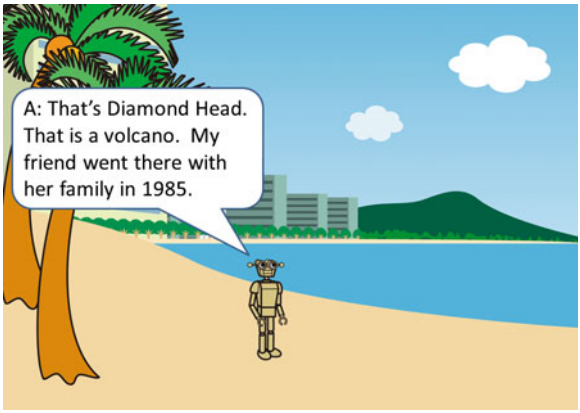


Fig. 5.10 A monologue adapted from a conversation in Fig. 5.1. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

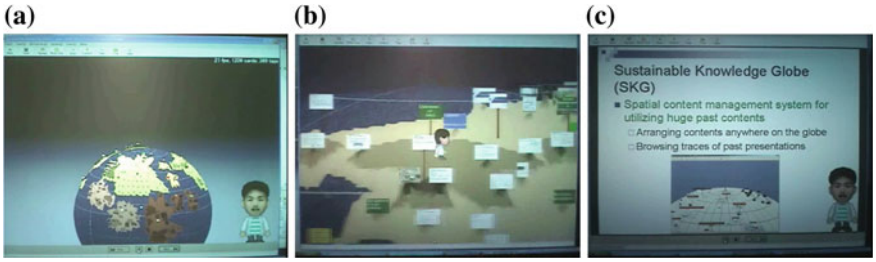


Fig. 5.11 Sustainable Knowledge globe. **a** Intial. **b** Moving. **c** Presentation. (Kubota et al. 2007)

(Fig. 5.11a), whereas by zooming in, s/he can manipulate the details of conversation quanta (Fig. 5.11b, c).

Importers and exporters who can convert conversation quanta to/from existing media, such as web documents, can and should contribute to increasing the collection size of the of conversation quanta. Tools can effectively contribute to this process, providing that they can support the dissemination of group communication so that the community of participants can share knowledge or information.

5.5 Circulation Scheme

We believe that the circulation of conversation quanta within the community is the most effective means of improving both the quantity and quality of the collection of conversation quanta in an evolutionary fashion.

Both community and individual functions are critical. The community's role is to collect conversation quanta from its members, index them according to their content and social aspects, and disseminate them to interested members. These functions contribute to an evolutionary process by selecting conversation quanta that are liked or supported by a majority of the participants. The individual's role is to organize a personal collection of conversational quanta according to the personal traits of each individual. As shown in Fig. 5.12, an evolutionary process requires the combining of community and individual functions. Community functions have been widely discussed in the context of social informatics, and, especially, knowledge management. In the remainder of this book, we will, therefore, focus on individual functions that involve numerous pending issues.

The individual function is essential to enable incoming information to expand the ensemble of conversation quanta. This is because individuals add their own experiences to the collective memory, or create new stories by combining and editing existing conversation quanta based on their own thoughts. This individual function needs to be inherently dynamic to play a vital role by fostering personal thoughts.

The theory of dynamic memory can be employed for building up autobiographical memory by empowering the individual memory process. Schank (1982)

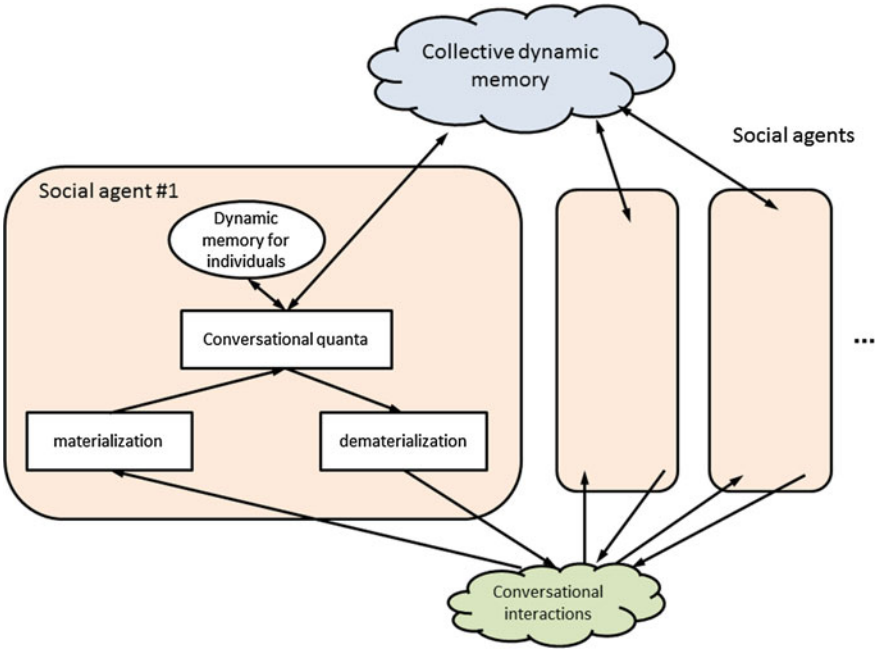


Fig. 5.12 Dynamic memory as a model for an individual agent

suggested that dynamic memory centers on reminding, indexing, and generalization. The reminding mechanism enables recall of the memory relevant to the given situation. The indexing mechanism enables the personal memory to be organized through the grouping together of similar stories. Schank emphasizes the use of themes or even maxims as centers of story clusters. The last mechanism of generalization mechanism, though still requiring extensive research, contributes to the extraction of essences from stories.

Diverging from dynamic memory theory, we emphasize that memory units are in fact conversation quanta that can be roughly classified into two classes. The first class of conversation quanta contains first-person narratives. This means that there is one character in a conversation quantum representing the owner of the conversation quantum who relates her/his own experiences and thoughts as an actor. By contrast, the second class of conversation quanta merely includes imported stories from external sources. Although the first class of conversation quanta is valuable for its representation of the owner's direct experiences, these quanta are expensive. This is because they are not solely records, but require consistent reflection on the owner's interpretation and expression in ways that other people can make sense of. Conversation quanta of the second class are usually much less expensive, serving as a means of supporting personal memory.

5.6 Augmenting Conversation Through Conversation Quantization

As depicted in Fig. 5.13, conversation augmentation is key to the realization of the primordial soup of conversation through the application of conversation quantization. At the center is a *shared conversation space* wherein all of the virtualized participants from two types of conversation terminals, namely, *immersive conversation theaters* and *open conversation places*, can share the conversational environment, including virtualized referents.

In an immersive conversation theater, one or more participants can converse with other participants through projections of images of these other participants, and of referents, on the surrounding walls. The body image and behaviors of the participants in an immersive conversation theater are sensed and transmitted to the shared conversation space for communication with other participants. When anonymous participation is desired, a participant may be incarnated as an avatar or a robot.

In an open conversation place, participants interact with other participants within an augmented reality in which real robots serve as physical avatars of the other participants. The behaviors of the participants in an open conversation place may be projected into the shared conversation place so that other participants may become aware of them for further conversation.

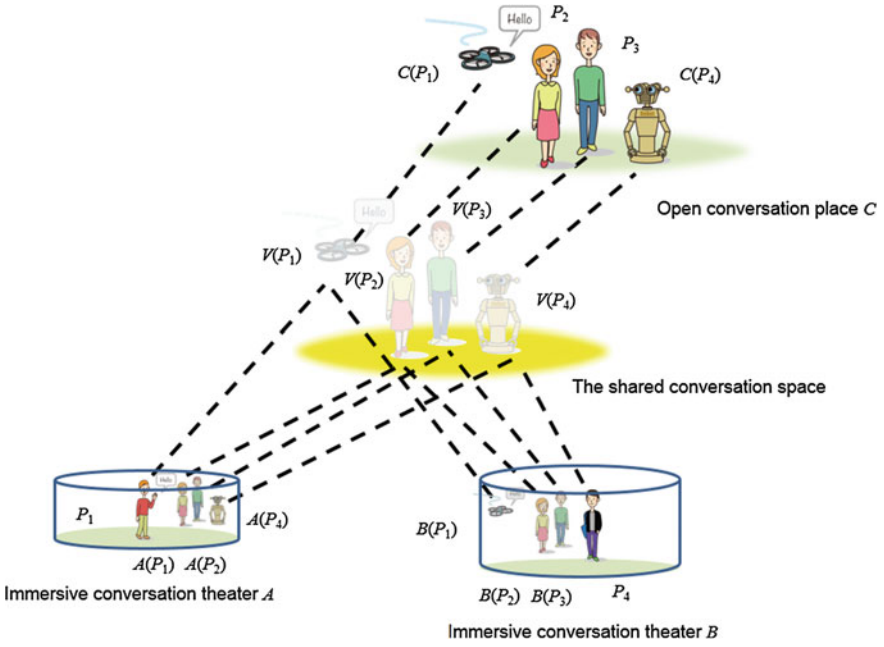


Fig. 5.13 Augmented conversation centered on the shared conversation space. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

These schemes of production/consumption, manipulation, and circulation are defined and embedded into open or immersive conversations spaces, depending on the nature of the task. The concept of an augmented conversation environment can be manifested and applied to various kinds of tasks. In what follows in the remainder of this section, we will elaborate on three typical examples: shared virtual meetings, interaction games, and tele-existence.

5.6.1 Shared Virtual Meeting Space

Individuals may participate in a meeting from one or more immersive conversation theaters, as shown in Fig. 5.14. Even without the application of any advanced computational intelligence, they may simply participate in a virtual meeting space from different physical places across the globe to enjoy conversation in this shared place. The background of the conversation might be any panoramic image that may be obtained from the internet. Alternatively, one of the participants may use an omnidirectional camera to capture an on-the-spot panoramic image. An omnidirectional camera may even be attached to a mobile robot that can not only run along a road, but also fly high in the sky or dive deep into the sea. The participants can choose

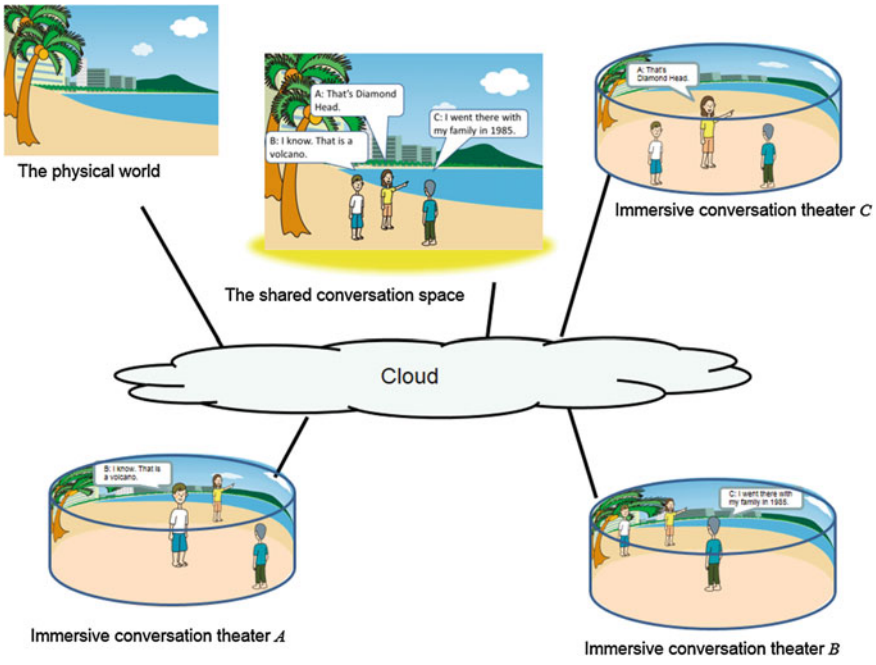


Fig. 5.14 Shared virtual meeting space. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

the place by controlling the camera or robot, or asking the cameraman to do so, and can move around virtually by controlling the background image. From time to time, a simple technique may be needed to insert the participants' images into the background image.

Even in the case of the above augmented conversational environment, all the conversations exchanged may be recorded, together with the background image which can serve as an invaluable resource of conversation quanta.

A smart sensing technique enables us to detect deictic gestures of the participants and record conversation together with referential information. Alternatively, we can segment conversations by recognizing nonverbal expressions, such as turning around, which may work as a discourse marker to delimit the conversation discourse.

Extensive natural language processing may also be necessary if complex linguistic expressions are used. Further difficulties may arise when a local gesture space is created for a structured space of imaginary referents during the conversation. The system needs to be able to distinguish whether a given nonverbal/verbal expression refers to an object in the projected background on the display surface surrounding the user, or to one present in the local gesture space. This perceptual information processing enables us to semi-automatically produce much higher quality conversation quanta at little or no cost than would be otherwise possible.

A more challenging scenario entails the introduction of one or more guides or navigator agents into the augmented conversation environment. If a sufficient amount of conversation content has been accumulated as a collection of conversation quanta for conversation scenes, it is relatively easy to create informative behaviors of guides/navigation agents. This may be one of the most feasible ways of creating the primordial soup of conversation. Autonomous agents are powerful enough to not only convey conversations beyond temporal constraints, but also to accumulate pieces of conversation into integral components of well-organized stories. Indeed, we believe that the richness of the underlying stories that are accumulated contributes to the engagement of participants.

5.6.2 Virtual Interaction Game

A game is a structured interaction among participants competing to achieve certain goals. Although competition may be a basic element of a game, cooperation may also be an important feature in the case of team games. A structured interaction may be viewed as a game, or interaction game, if participants appear to follow certain rules and to cohere to achieve some explicit or implicit goal. The collective action of composing a story, or interactive storytelling, can often be viewed as an interaction game, as it may be accompanied by tacit rules for cooperation and competition to bring about the sustained commitment of the participants. As games often induce participants' engagement, an interaction game may facilitate the composition of the primordial conversation soup on being introduced into the augmented conversation environment.

Virtual basketball (Lala and Nishida 2013) is an interaction game within an augmented conversation environment, as depicted in Fig. 5.15. It involves the participants, who are either humans or agents, sharing a virtual common place to play a game of basketball. An integrated sensing system, employing depth-sensors and pressure sensors projected onto a shared conversation space, senses the behaviors of the human participants. Agents, characterized as non-playing characters, may enable a game to be played, even when a sufficient number of human players are not available. Although basketball games are not usually considered as conversations, they share essential aspects of communication in common with conversations. Participants in both use various kinds of social signals to coordinate their behaviors. A key challenge here is to endow agents with enough social presence so that human players regard them as intelligent partners with whom high-level social signals can be exchanged.

Another key challenge is mutual adaptation between humans and agents so that new communication protocols can be sought and established in a heuristic fashion. It is natural to assume that a communication protocol is dynamically generated as it is unlikely to be provided in advance. The context of the game may create certain constraints for participants to formulate a communication protocol without explicit collaboration.

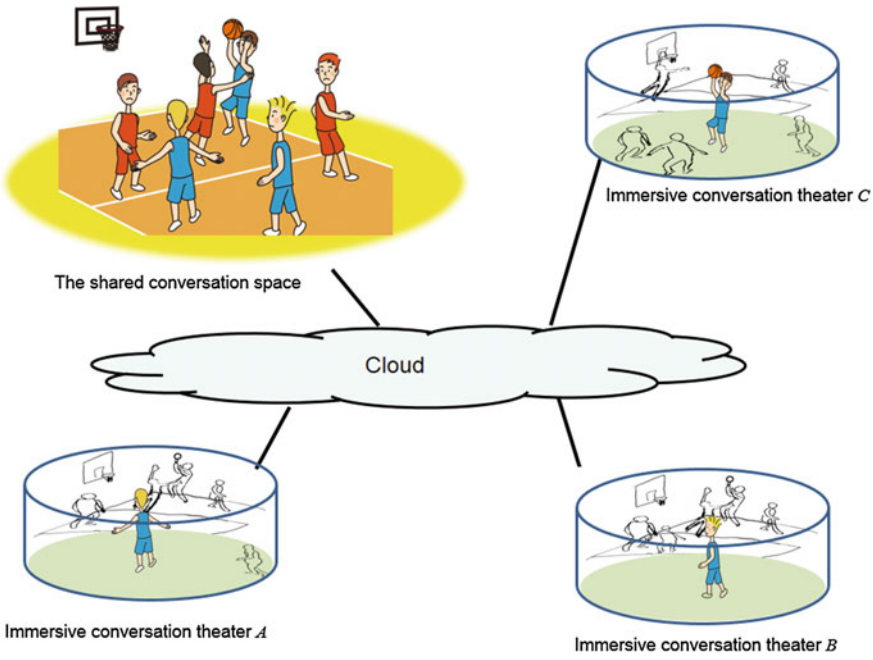


Fig. 5.15 Virtual basketball game. Image inspired by Lala and Nishida (2013). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

5.6.3 Tele-presence

As shown in Fig. 5.16, by creating a tele-present setting, one or more individuals in distant locations are able to participate in a conversation that takes place within a physical environment. In other words, technology equates a physical conversation space with a shared virtual space. The conversational situation within a physical environment is audio-visually projected on the wall surface of each immersive conversation theater so that remote users are able to perceive what is happening in the physical space and produce situated behaviors proactively or reactively. The behaviors of a user in an immersive conversation theater are sensed and projected onto a robot that represents the user in the shared physical conversation place.

There are numerous practical applications of tele-presence such as tele-shopping and tele-operations. In the context of conversation informatics, it may also serve as an immersive "Wizard of Oz" facility for collecting data on human-agent interaction for analysis, or for producing the communicative behavior of robotic agents. In this immersive facility, the user can conceive of a situation in the remote physical conversation space as if he or she were incarnated as the robot out there. His or her conceptions about an ideal communicative behavior in the given situation may be

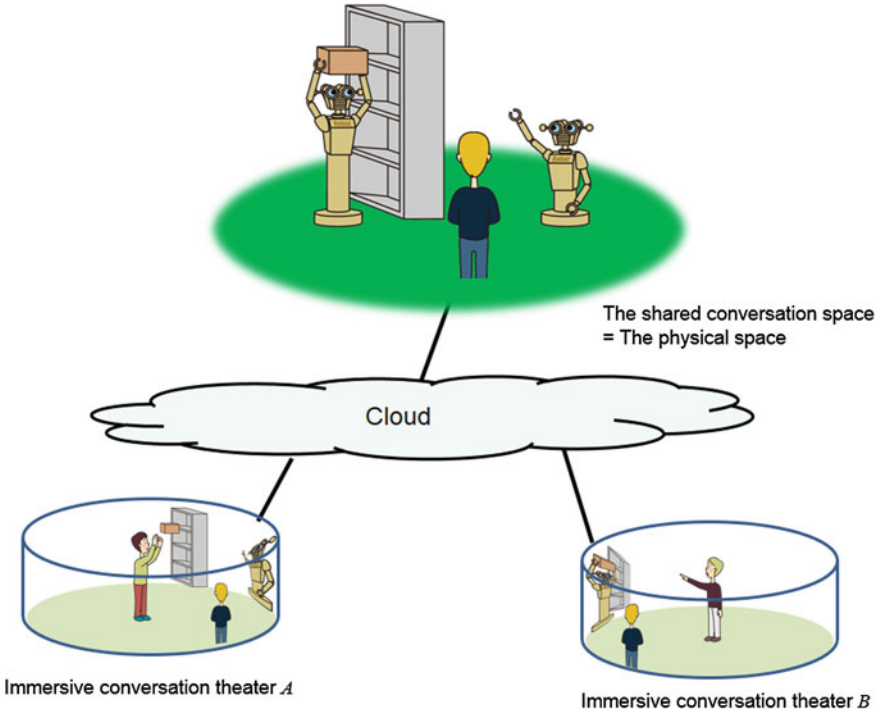


Fig. 5.16 Tele-presence. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

observed as an action that is projected onto the robot in the physical environment so that the participants can sustain the interaction. The data collected can not only be used to analyze interactions, but can also be exploited as a resource for producing the communicative behaviors of robotic agents, using digital learning techniques.

5.7 Historical Notes

The idea of conversation quantization dates back to 1998 when we implemented the first prototype of circulating conversation records in a networked community named CoMeMo to increase awareness (Nishida et al. 1998). We employed a metaphor of talking alter-egos within this community. The user could initiate and observe virtual conversations among alter egos of herself/himself or/and others. The CoMeMo-community consisted of two major components. One was an alter-ego that retained the externalized memory of a person. Each personal memory was represented as a directed hypergraph with each node denoting a key word. The other was a conversation place where alter-egos took turns making utterances according to the rules of conversation. In each conversation session, participating alter-egos in the conversation place collaborated with each other to generate a story by alternately reproducing memory fragments from the personal memory embedded in each alter-ego (Fig. 5.17).

A conversation agent was used to present conversational content (Kubota et al. 2000). Monologue-dialogue conversion was applied to increase productivity (Nishida 2002). The concept of a circulating conversation was implemented as a Public Opinion Channel (POC) (Nishida et al. 1999) and as an "EgoChat" system (Kubota and Nishida 2002). A POC is a participatory broadcasting system that continuously elicits messages from people within a community and feeds edited messages back to them. It maintains the circulation of small talks consisting of stories that reflect questions, beliefs, and opinions arising within the community. Card-oriented representations of content, or knowledge cards, are employed to represent the information content of a small talk. Each knowledge card consists of text of around a hundred words, and possibly a pictorial image reflecting a small talk occurring within a community. This knowledge card representation was the origin of conversation quanta.

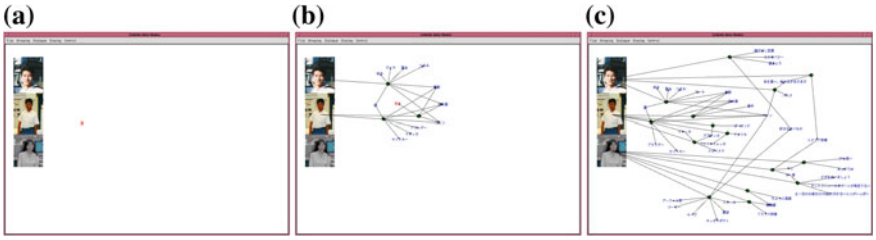


Fig. 5.17 CoMeMo-community. **a** Initial. **b** Growth. **c** Further growth. (Nishida et al. 1998)

Our POC system, as a whole, consisted of a POC server and two kinds of terminals known as a POC TV and a POC communicator. The POC server was similar to a list server except for its method of keyword base message retrieval. The POC TV provided the user with an easy access to the POC broadcasting. A couple of conversation agents appeared on the POC TV screen so that a POC message could be introduced to the user as a conversation occurring between two agents (Fig. 5.18). On being given a POC message, one agent (the caster agent) introduced the subject, and another agent (the announcer agent) talked about the content, though from time to time, the first agent could interrupt the second by inserting comments or questions. The POC communicator was a PC-based system that enabled the user to submit or browse messages. It maintained a user profile so that the message streams could be customized by users.

A Stream-Oriented Public Opinion Channel (SPOC) was an extension of POC TV in which a web-based multimedia environment enabled novice users to embody a story as multimedia content and distribute it on the internet (Matsumura et al. 2005). A sophisticated presentation was generated from the plain-text representation of conversation quanta specifying the utterances of participants in the conversation. The system produced both digital images and agent animations according to the availability of linguistic information in a given natural language text. Immersive

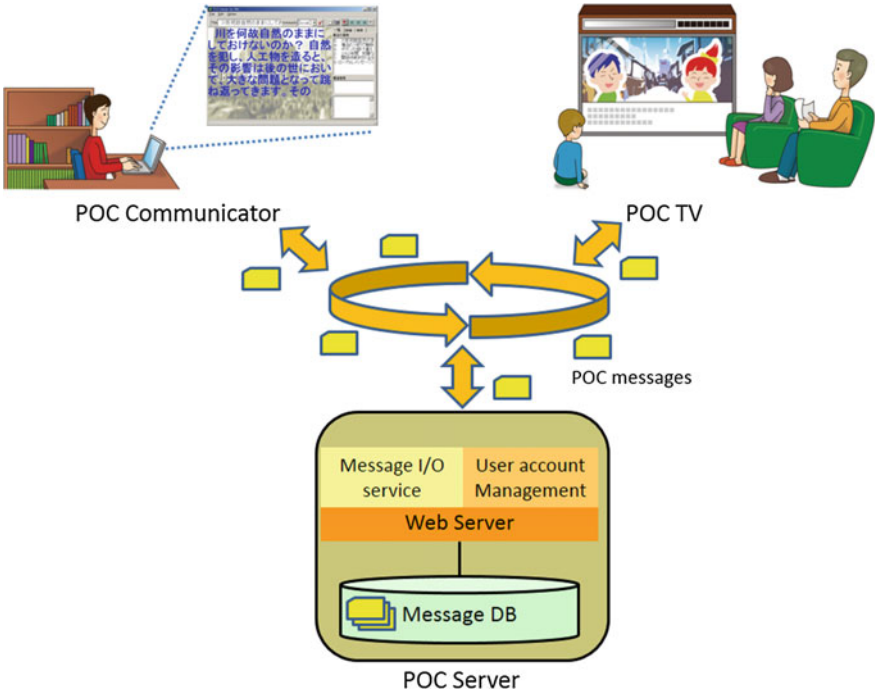


Fig. 5.18 POC TV. Adapted from (Nishida 2002). © 2014, Toyoaki Nishida, and At, Inc. Reproduced with permission

Public Opinion Channel (IPOC) (Nakano et al. 2006) is SPOC's successor. It enables an expansion of conversation quanta within a virtual immersive environment. Users can interact with conversation agents in a story-space, which is a panoramic pictorial background into which stories are embedded.

We constructed Sustainable Knowledge Globe (SKG) for editing and sharing a large-scale accumulation of knowledge cards (Kubota et al. 2007). In this environment, the spherical surface of SKG was projected onto a large screen. The users could then converse in front of a large screen. We built a conversation quanta acquisition tool that used SKG as a shared electronic whiteboard for group discussion (Saito et al. 2007) (Fig. 5.19). The conversation content could be captured in the human-assisted conversation content capture environment. We introduced a button device so that participants were able to specify when they thought that conversation should be recorded as a conversation quantum. In addition to the button devices, a touch screen served as a capture device whereby the user could touch the screen to signal her/his intentions. SKG enabled participants to edit the content of the captured conversation quanta and organize them into a structure on the fly.

Nishida (2007a) applied conversation quantization to in-vehicle conversation-sharing. In in-vehicle conversations, the participants' behaviors were expected to fall into a relatively small number of typical patterns. Simple sensing techniques sufficed to capture the conversational situations. Nishida (2007a) developed an augmented conversational environment for a driving simulator. By analyzing the pointing gestures of participants, the system could ground the conversation in events observed through the simulated window of the vehicle (Fig. 5.20). To encourage user

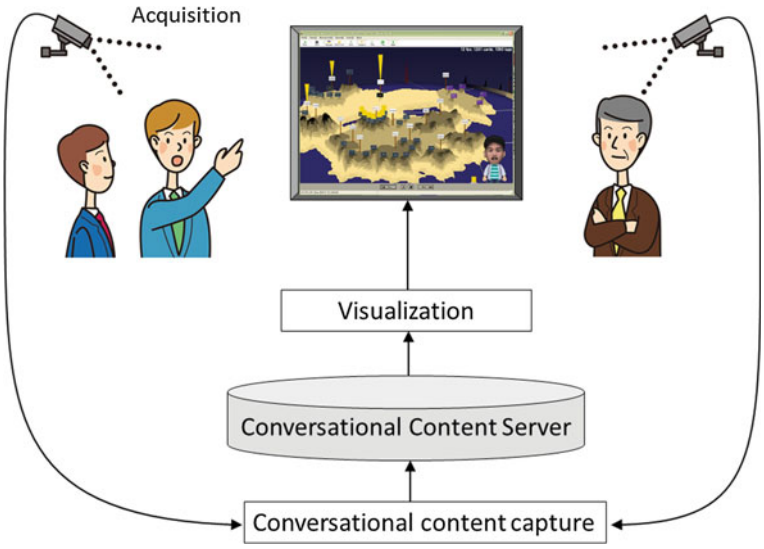


Fig. 5.19 SKG meeting (Saito et al. 2007). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

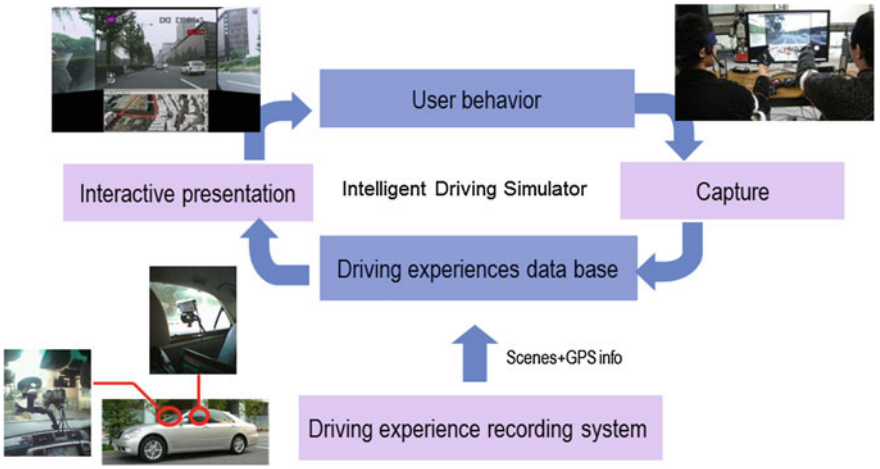


Fig. 5.20 In-vehicle conversation circulation (Nishida 2007a)

engagement by increasing opportunities for active participation, we introduced three features. First, we incorporated a 3D gaze generator into the system so that it could attract users' attention and create awareness about the system's internal status to enhance the sense of interaction. Second, we introduced a function that permitted the user to switch between immersive and bird-eye views. The former was used to follow the driving experience of other people, while the latter was used to actively obtain global information about a region such as the distribution of frequently discussed topics. Third, we developed facilities for capturing the user's communicative status. Some of these passively measured the user's nonverbal behaviors, while others were designed to facilitate the user's spontaneous utterances and information provision.

For human–robot interactions, we employed conversation quantization for circulating knowledge using conversational robots (Nishida et al. 2006). With the aim of prototyping the concept of robots as embodied knowledge media, we constructed listener and presenter robots on top of the three layer model. The pair of robots served as a means of communicating embodied knowledge (Fig. 5.21). The listener robot first interacted with a knowledgeable human to acquire knowledge quanta. The presenter robot, equipped with a small display, then interacted with a human to show the appropriate video in appropriate situations where this knowledge was considered necessary. Conversation quanta served to encode knowledge transferred from the listener to the presenter robot.

The listener robot was designed to exhibit appropriate nonverbal behavior while the presenter was explaining and producing a series of conversation quanta as records of the conversation. When the listener robot was first introduced, the subject domain was furniture assembly. We subsequently extended the framework to bicycle assembly and disassembly. We designed the response behavior of the listener robot based on

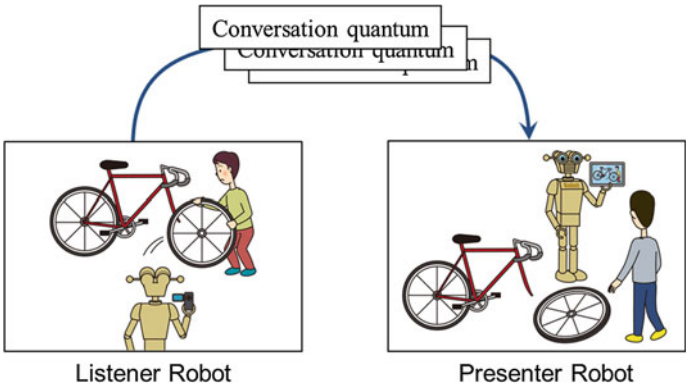


Fig. 5.21 Listener–Presenter robots (Nishida et al. 2006). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

its ability to analyze motion capture data of the user and to identify four action modes of the human instructor: talking-to, talking-about, confirming, and busy modes. The presenter robot was designed to approach the task area when it detected the human listener's need for assistance. It would then adjust the position and angle of the display according to the listener's position and posture, and show the video to the listener.

5.8 Summary

In this chapter, we have described the concept of conversation quantization, which conceptually supports a data-intensive approach. There are four main ideas underlying conversation quantization: conversation quantum as a package of information for describing the meaning and expressions of a significant segment of conversation, the use of conversation quantum as a component of complex conversational scenes, generic procedures for obtaining, manipulating, and utilizing conversation quanta, and amenability of implementation at different levels of automation. The actual implementation may vary depending on granularity, depth and breadth of annotation, representational fidelity and generality. We discussed four aspects of conversation quantization: representation, production and consumption, manipulation, and circulation. The representation scheme involves how conversation quantum is associated with conversation scenes, how salient objects and participants are characterized, annotated transactional record of interaction, and how a given conversation quantum is related to other conversation quanta to make up a larger discourse. In the representation scheme, the discourse link of a conversational quantum allows for representing embedded discourse. The key component is a schemata-based recognizer that matches an appropriate schema to a given conversation scene, and applies it to ascertain how each component of the schema is instantiated in a given conversation

scene. In the manipulation scheme, we identified a generic semantic operations such as concatenate, partition, summarize, elaborate, revise, embed, merge, and split. In the circulation scheme, dynamic memory organization is critical both in organizing individual and collective memory. The application of conversation quantization centers on the implementation of a shared conversation space using the mixed reality technology. We discussed shared virtual meeting space, virtual interaction game, and tele-presence as typical examples. Finally, we have included a historical note, as the idea of conversation quantization has slowly evolved over a decade.

Chapter 6
Smart Conversation Space

Abstract The spatial environment surrounding participants in conversation has a critical, if not conclusive, influence on the qualitative and quantitative aspects of conversation conducted therein. Imagine how awful the conversation would be if one had to converse with a stranger in a vacuum with only the two of you present. A contrastive situation would be one in which the conversation takes place in a cheerful cafe surrounded by other people with a soft wind blowing off the distant mountains and a relaxing atmosphere on the nearby beach. The role of a smart conversation space is twofold: to improve a given space with augmented reality or build a virtual space to support conversation and second, to record and measure conversation not only for better user services, but also for better scientific investigation of conversation in both ordinary and smart spaces.

Keywords Communication environment · Immersive environment · Human-agent interaction · Interaction management · Experimental settings for HAI

6.1 The Architecture of Smart Conversation Space

The smart conversation space implements the theory of conversation enhancement introduced in Sect. 5.7, which is a formalization of the idea of a primordial soup of conversation introduced in Sect. 1.3.

Two items of equipment are included in the virtual space accommodating the shared conversation space as illustrated in Fig. 6.1. One is an open conversation space enhanced by augmented reality (AR) where people can move around talking to each other with reference to objects and events in the space. AR technology provides the participants with devices such as wearable glasses to overlay auxiliary information on the physical environment to encourage active conversation. The other is an immersive conversation space where each participant is completely encircled by large

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_6

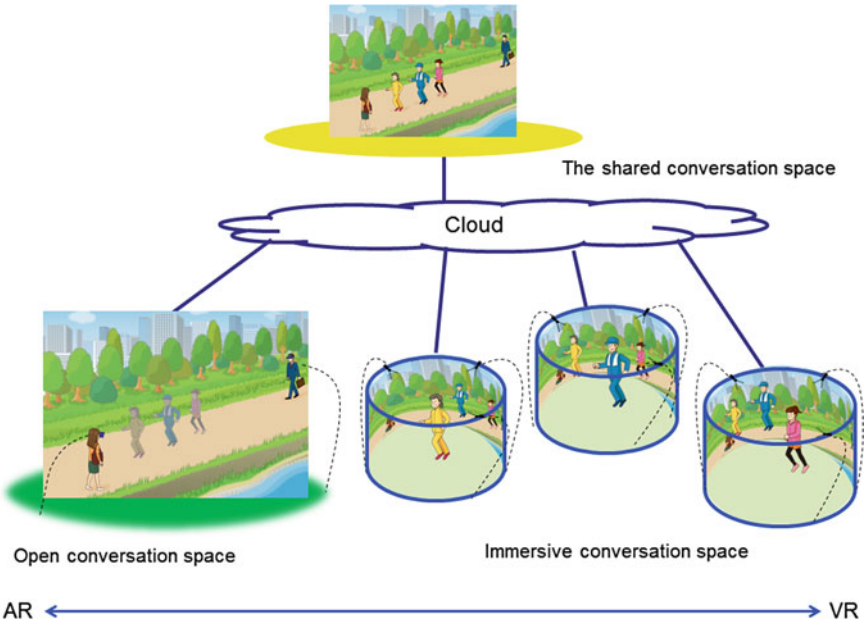


Fig. 6.1 Smart conversation space. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

screens and the behavior of the participant is sensed and reflected on his/her image in the shared conversation space. The immersive interactive environment allows participants from spatially distributed locations to participate in the interaction in the shared conversation space, with the background taken from a global service such as Google Street View or from an omnidirectional camera attached to a mobile robot. By integrating both items of equipment, the smart conversation space can be configured as an arbitrary cyber-physical conversation space not only for entertaining users but also for investigating how they behave.

There are many technical challenges in making this happen. Regarding the open conversation space, the challenges center on measurement and recognition of a physical world environment in real-time. The system needs to know how the physical world is configured and how the users are located therein to provide the correct information on the fly. Regarding the immersive conversation space, it had to be designed and implemented from scratch, owing to the lack of similar existing systems. In what follows, we present a method for associating conversation quanta with the physical environment, a method for capturing human behavior in an open conversation space, and techniques for building an immersive conversation space. We also explain how the immersive conversation space is used for applied research in conversational informatics.

6.2 Situated Knowledge Media

The smart conversation space extends our early attempts at building a communication medium to facilitate the sharing of knowledge about scientific instruments between expert engineers and end-users distributed across the globe. The situated knowledge media project aimed to empower online customer service not only to benefit clients who need help, but also to help engineers understand their products and learn from customers. Neither the engineers nor end-users were assumed to be computer-savvy even though as it turned out, they had plenty of expertise in manufacturing and using scientific instruments. Nor did we assume that they had worked closely together. Instead, we assumed that their work places would more than likely be distributed across the globe and that they would have to rely on the Internet as their main communication means. We also assumed that they would need to communicate frequently with one another, since the development of advanced scientific instruments often creates unprecedented situations that preclude the preparation of concrete user manuals before shipping of the instruments and the lack of a complete set of manuals often gives rise to questions and discussion.

Our solution involved providing a situated knowledge management system, which as shown in Fig. 6.2, is a mixed reality system that uses virtual reality (VR) and AR techniques to allow the engineers and end-users to share a simple implementation of conversation quanta as HyperCard-like objects associated with places on the

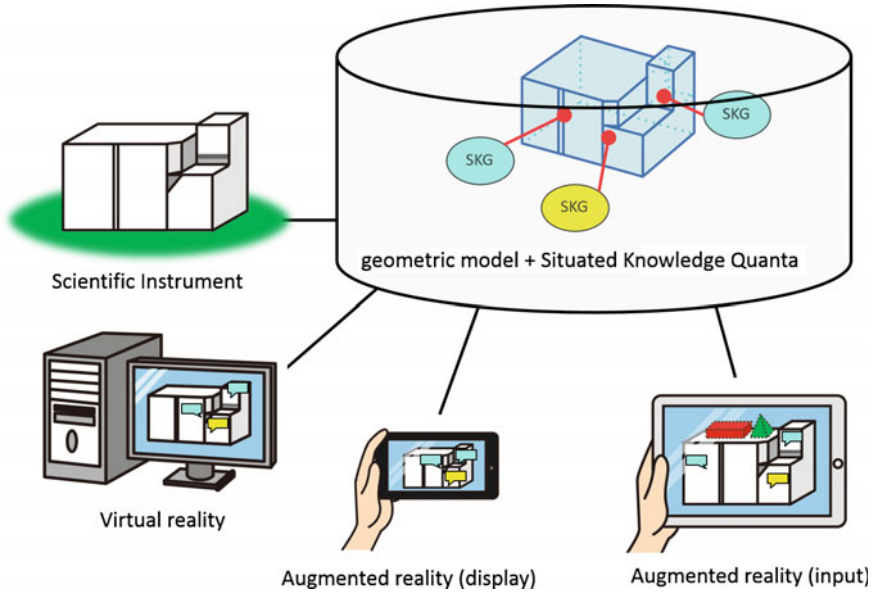


Fig. 6.2 Overview of the situated knowledge management system. Adapted from (Merckel and Nishida 2010). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

surface of the scientific instruments. It permits engineers to associate their notes by annotation based on a "point-and-tell" or "do-this-way" metaphor, and the user to retrieve information by "point-and-ask" or to express a request by demonstrating "something-like-this".

The real challenge was the real-time, light-weight object pose recognition algorithm allowing users to associate conversation quanta with various locations on the scientific instruments as a form of memory for possible retrieval in the future. This algorithm employed a CAD model (piecewise linear complex) and the camera-image of the target object to estimate the pose of the object with respect to the camera (Merckel and Nishida 2009a, 2010); an interface to correct the estimated pose; and a low-overhead three-dimensional (3D) item drawing engine (Merckel and Nishida 2009b). The resulting suite works both in augmented reality and augmented virtuality environments. An augmented reality environment allows annotations to be overlaid on the camera-image of the target object, while an augmented virtuality environment allows for the creation of a 3D virtualized target object by automatically pasting surface textures. The 3D item drawing engine consists of a handheld AR system, which allows the user to directly draw freehand 3D lines in the context of the subject instruments.

The situated knowledge management system comprises key components of a smart conversation space, including a mechanism to provide users with a means for recording and reproducing situated messages.

6.3 Capturing Human Behavior in Open Conversation Space

Multi-party conversation including surrounding objects in the environment is one of the most complex situations. We would like to analyze the interactive behavior among the participants and the surroundings from the perspectives of an outsider and a participant. The challenge is that the complexity of the analyzed interactive situation increases with the importance of the subjective analysis to interpret the measured data. We therefore, need a system to record and display the conversation scenes in the form of 3D model data. The system should be able to record the surrounding environment including movable objects and the time series data of each participant, such as the positions, surfaces, and shapes of the head, arms, torso, and legs.

We need to develop a new algorithm for analyzing multi-party interaction, since although numerous systems have been developed to record and reconstruct 3D model data of human body surfaces and motions, e.g., (Moeslund and Granum 2001), these are based on the assumption that there are no objects around the person, which does not hold for multi-party conversation, which is the focus of this study.

Three-dimensional conversation capture by multiple Kinects (3DCCbyMK) is a system that can measure and record the entire conversation space including multiple persons and the surroundings. The system can capture a 3D model and skeleton data in multi-party interaction using RGB cameras and range sensors. In addition, a 3D model of the surrounding environment can be reconstructed. 3DCCbyMK consists

of two subsystems: one for capturing 3D models and the motion of persons in the conversation, and another for reconstructing a static 3D model of the surrounding environment. The system can record and display the entire conversation as time series data.

The current version of 3DCCbyMK takes as input a conversation with up to four persons talking freely, moving, and making gestures such as pointing, in a space with a 3-m diameter, which includes objects on a table and the walls as in an exhibition. The system reconstructs the objects and the persons' motions and movements. The system can also display the conversation from an arbitrary perspective. Typical behaviors expected to be captured by the system during the conversation include moving in the space, pointing gestures, hand gestures, and body motions that do not make contact with other persons or objects.

The 3DCCbyMK system records and reconstructs in 3D an entire indoor multi-party conversation to analyze the interaction among participants in the conversation and the surrounding objects. We use Kinect sensors as RGB cameras and range sensors. The middleware used to obtain the Kinect data is OpenNI,[1] by which we can capture an RGB image, depth map, skeleton data, user masks, and tracking ID. An overview of the system architecture is depicted in Fig. 6.3, where the components on the left constitute the data capturing subsystem, those in the center the data processing subsystem, and those on the right the display subsystem.

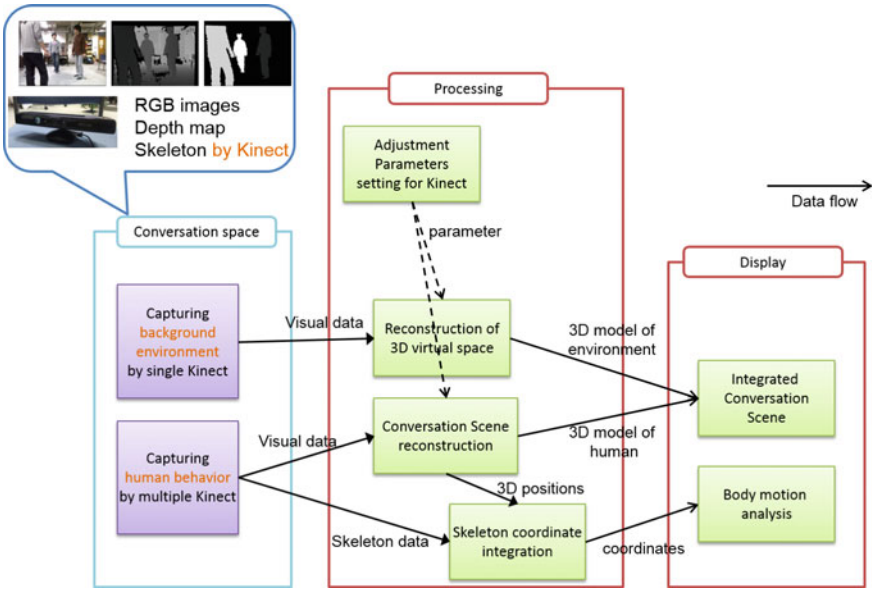


Fig. 6.3 The architecture of the 3DCCbyMK system

[1] http://en.wikipedia.org/wiki/OpenNI.

The data capturing subsystem includes two parts for capturing the environment and participants, respectively. When capturing the environment, the user scans the surroundings using a Kinect sensor. In contrast, when capturing human behavior, multiple Kinect sensors (about five) are placed around the conversation field. The subsystem can easily calibrate the sensors so that the user has flexibility in placing the sensors according to the capturing environment and conversation.

The data processing subsystem consists of four subsystems, namely, lens distortion correction, 3D environment model building, participant motion estimation, and conversation scene reconstruction. The lens-distortion-correction subsystem corrects the distortion of the RGB lens and range sensor. To correct RGB distortion, we use the camera calibration function of OpenCV[2] for parameter fitting. To correct range sensor distortion, a least-squares method is applied to the data obtained from the measurements of calibration boards placed at certain distances. In addition, when multiple Kinect sensors are used, the depth maps may contain some gaps because of interference by the infrared lights of multiple Kinects. The lens-distortion-correction subsystem fills these gaps using the method reported by Maimone and Fuchs (2011).

The 3D-environment-model-building subsystem uses the SLAM (simultaneous localization and mapping) method based on the scanned data of the environment from a single Kinect sensor. The subsystem reconstructs a 3D model of the surrounding environment using image features and the depth map. The subsystem initially estimates the relative 3D location of the scanning Kinect sensor using image features. After the estimation of the Kinect sensor position of each scanned frame, the subsystem can integrate the scanned depth map data and reconstruct the environment.

Figure 6.4 shows how each frame in the captured data is processed. First, the RGB image is converted to a gray scale image. Second, the image features calculated by the SURF (Bay et al. 2006) method are used to calculate the degree of similarity among them. Third, a 3D coordinate transformation is calculated from the correspondence relations of image features. To calculate the transform, a least-squares method is applied to the image feature data, which is weighted according to the distance between the features, the degree of similarity, and the values of the depth map around the image feature. The 3D-environment-model-building subsystem uses the LMedS (Zhang 1997) method to reduce errors in image feature matching. If the transform of the current frame is very different from that of the previous frames, the current transform is not adopted.

After estimating the Kinect sensor positions, colored 3D points are plotted as absolute coordinates. The colored 3D point cloud is estimated by the 3D Kinect positions, the RGB image, and the depth map. The subsystem does not reconstruct smooth surface like polygons; instead it produces a colored 3D point cloud as the reconstructed environment. The subsystem can also integrate the reconstructed environment and 3D models of persons using the data captured from the surroundings when the subsystem records the 3D models and skeletons.

The participant-motion-estimation subsystem uses the skeleton data from multiple Kinect sensors to estimate the motion of the participant. First, personal IDs are

[2] http://opencv.org/.

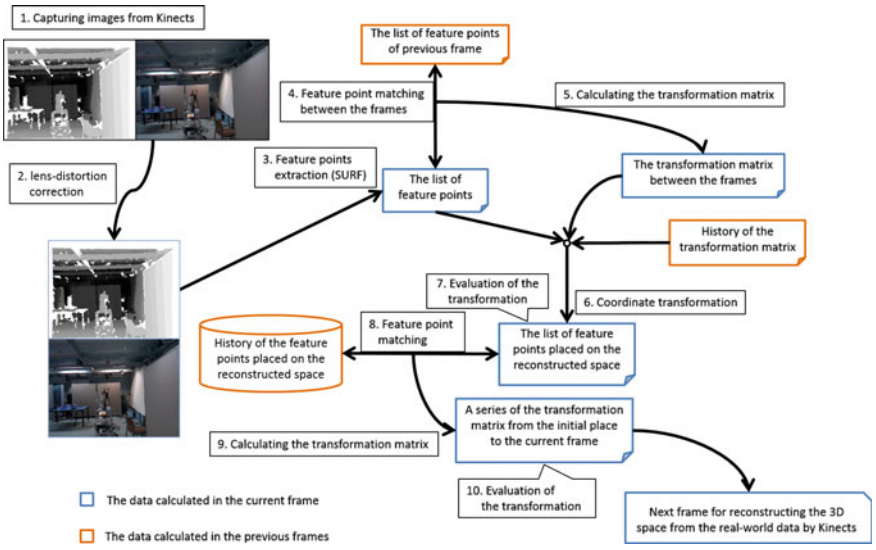


Fig. 6.4 The processing flow to each frame of the captured RGB image and depth map

allocated to each set of skeleton data for each Kinect sensor. Second, each set of skeleton data is projected onto an integrated coordination while the personal IDs are integrated based on the overlap of the skeleton coordination data. Third, each joint coordinate of the skeleton data is integrated after being weighted according to various heuristic conditions, such as how far the joint is from the Kinect sensor, whether the joint is occluded, how many sensors have captured the joint, and whether the captured person is facing the Kinect sensor. To correct any misunderstanding of right and left joints, the subsystem checks the time series data to ensure consistency of joint recognition.

The conversation-scene-reconstruction subsystem builds a 3D model for the conversation scene by integrating the 3D environment model and the skeletons of participants in the conversation. Currently, a simple method is employed to integrate multiple Kinect sensor coordination. To decrease the overhead of calibrating multiple Kinect sensor coordination, the subsystem integrates them using the skeleton data and depth map. First, a person is captured by multiple Kinect sensors. Second, the skeleton data are used to calculate the relative positions of the Kinect sensors. Third, measured 3D points near the matched skeleton points in the depth map are used for fine adjustment using a least-squares method. This method recalculates the sensor coordination after obtaining measurements from the skeleton data in measured scenes if the data have been correctly captured.

How effective is the lens distortion correction method? Figure 6.5 shows the output from the system for a laboratory room $8 \times 6 \times 2.7$ m without lens distortion correction. The reader can easily see that the reconstruction has not been very successful, especially compared with the result shown in Fig. 6.6, which is obtained by processing the same data with lens distortion correction. The result with lens distortion

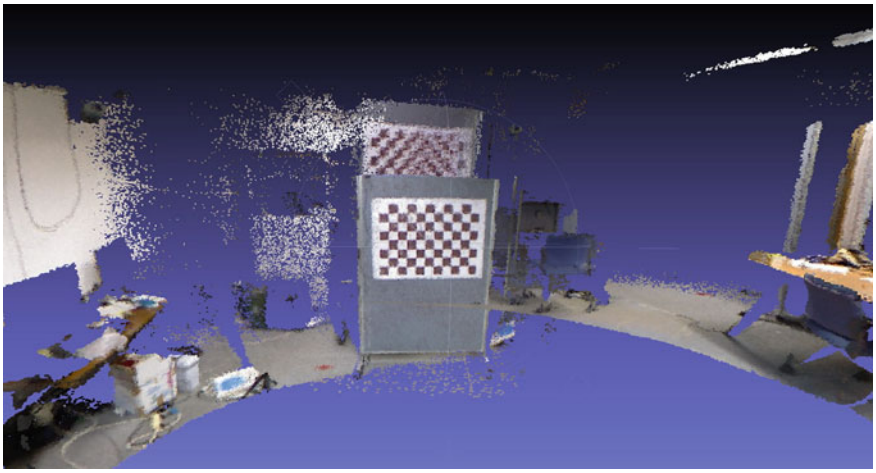


Fig. 6.5 Result of reconstruction without lens distortion correction

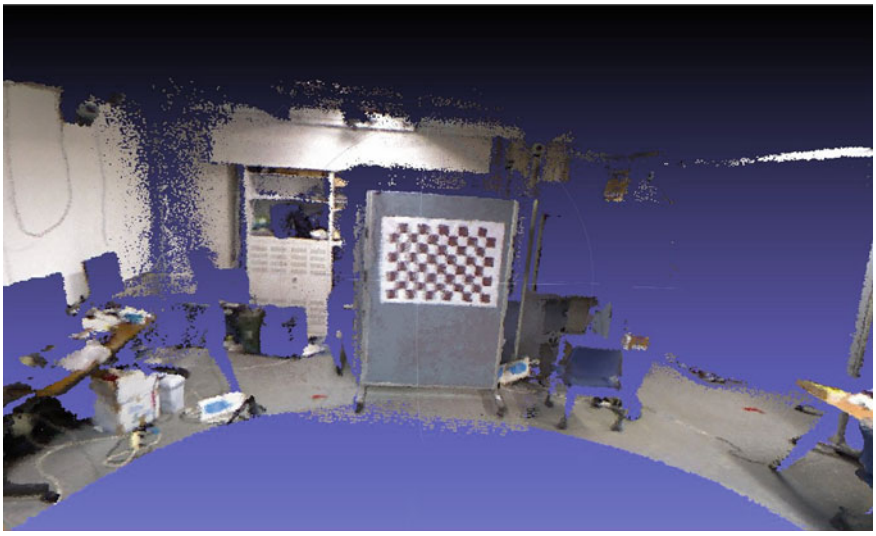


Fig. 6.6 Result of reconstruction with lens distortion correction

correction is clearly better; for example, the ghost image of the checker board has been eliminated and several surfaces are restored correctly. Figure 6.7 shows the result of the reconstruction using scanned data from the center of the room. Some parts have not been filled in, but we can reconstruct the entire environment using denser scanned data.

Now, let us consider how well our 3D model reconstruction algorithm works for persons in conversation using five Kinect sensors. Figure 6.8 shows a recorded conversation scene with a captured area of 3.4×2.6 m. We captured about 80 s of

Fig. 6.7 Result of reconstruction using scanned data (looked down from the center top of the room)

conversation between three persons. To evaluate the integration and reconstruction, we focus on the skeleton data of sensor E. Having extracted good data from sensor E, we used these skeleton data as the correct skeleton data.

Figure 6.9 shows the result of the reconstruction using only the data from the captured conversation and the result of integrating the captured data of the conversation and environment. Although the quality is still lacking, the position of each person in the conversation environment can be identified. By using the reconstructed 3D model data and skeleton data, we can to some extent analyze human communication in a multi-modal conversation. For example, we analyzed the multi-modal interactive behavior of participants learning ballroom dancing with an instructor. When beginners learn techniques involving body motions, they have to understand what is important in carrying out these techniques. We focused on teaching a ballroom dance as an example and proposed a method to extract the important points, which are usually learned implicitly. We analyzed and acquired the important features for smooth and effective learning using the 3DCCbyMK. The system and the reconstructed 3D model are also helpful when presenting the features to beginners. In Chap. 8, we discuss our research study using this system.

6.4 Immersive Collaborative Interaction Environment

An immersive conversation space uses a series of large screens encircling the user to provide rich ambient audio-visual information so that s/he may be involved and feel closely attached to the illustrated scene. The user in return responds with diverse social signals such as voice, body motions, altered gaze, facial expressions, and so on.

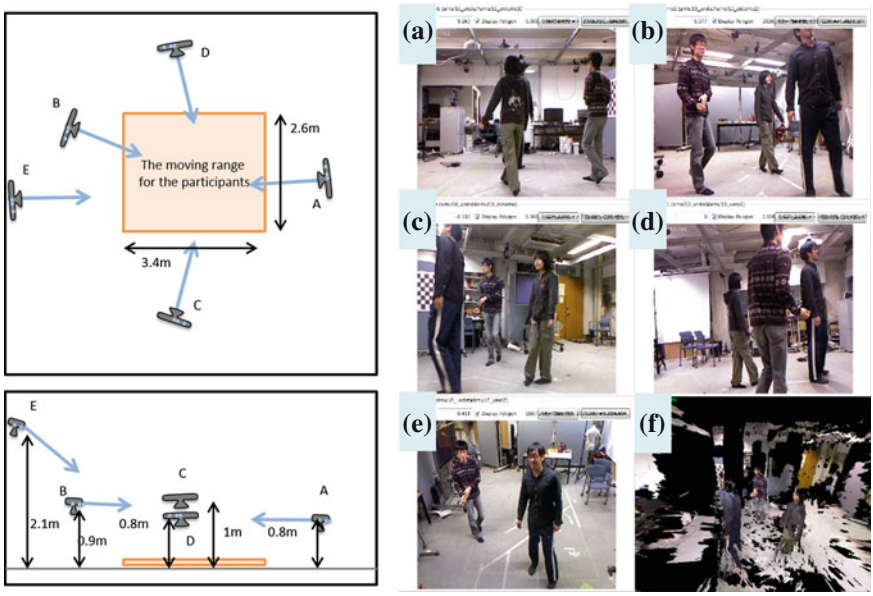


Fig. 6.8 Setup for recording a conversation between three individuals. **a–e** scenes captured from Kinect sensors (**a**)–(**e**), respectively. **f** a snapshot of the 3D moving picture synthesized by integrating the Kinect sensory data

Fig. 6.9 Result of the reconstruction showing only the data when capturing the conversation (*left*) and integration of the captured data for the conversation and environment (*right*)

Immersive Collaborative Interaction Environment (ICIE) is an immersive conversation space, featuring rich ambient information for the user to enjoy with low cognitive load and the ability to capture human motion with few physical constraints. The ICIE allows a human operator to control an avatar or tele-operated robot in a situated fashion in a human-agent interaction (HAI) environment from a first-person perspective.

The ICIE is not only useful as a powerful information kiosk, but also provides an advanced experimental environment for communication scientists to capture social

interactive behavior from a first-person perspective in which ambient information is controlled to enable the investigation of different social interactions. The ICIE allows investigators to design interactive behavior influenced not only by attributes of each communication member but also implicit and explicit conditions in the ambient information including the situation, objects, and communication partners.

It should be noted that alternative approaches such as a head mounted display (HMD) or Cave Automatic Virtual Environment (CAVE) (Cruz-Neira et al. 1993) are not necessarily better than the ICIE. Although the HMD is a typical device providing immersive images, wearing devices such as the HMD prevents an operator from social actions because of the weight and covering of his/her face, and hence it is hard for the operator to recognize information from the surrounding area and to provide social signals. Although CAVE provides a more immersive view than ICIE, the cubic arrangement of the screens of a CUBE system skews an image at the corners and edges. In contrast, the ICIE adopts a 360° immersive display, composed of eight portrait orientation LCD monitors each with a 65-inch screen size (0.9 m wide, 1.6 m high) in an octagon shape.

The ICIE, the architecture of which is illustrated in Fig. 6.10, uses an immersive display to provide the operator with a first-person perspective without physically and temporally interfering with his/her interactive behavior. Since we expect that the behavioral data obtained in the ICIE will be used for modeling social interaction between human and conversational agents, we need to obtain interactive data corresponding to the information, which the conversational agent uses as information from a human's first-person perspective. For this purpose, we require a digital

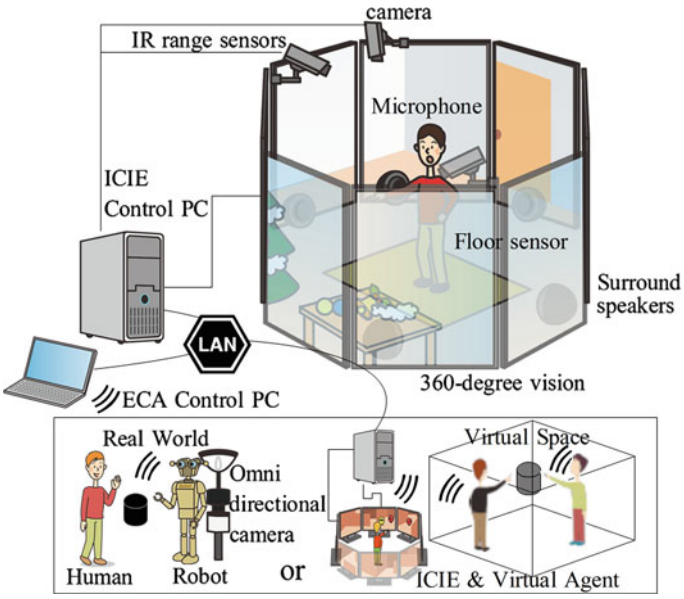


Fig. 6.10 ICIE architecture. © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

data projector that can easily control and change the provided information. Eight surround-sound speakers reproduce sounds, which the operator can hear in the virtual space. The operator in the ICIE can to a certain degree, detect the direction of the sound source in real-time. In the immersive environment, the operator interacts with virtual agents and avatars controlled by other ICIE operators. The ICIE provides an interface to the virtual world. In the rest of this section, we discuss some of the features of the ICIE in more detail.

6.4.1 Providing a First-Person Perspective

ICIE has three subsystems for virtualizing the physical world: a wide-area virtualizer, small-space virtualizer, and human-body virtualizer.

The wide-area virtualizer constructs immersive virtual spaces with high resolution visual information for social interaction (Mori et al. 2011). First, the system roughly reconstructs the 3D geometry from many real world photos using the stereo method. Thereafter, the system constructs panorama images and interpolates between panorama image pairs if the user moves to a position where the system does not have a corresponding image. In parallel, the system creates a depth map for each panorama image from the 3D geometry. Using this system, we can construct a widespread virtual space with high resolution visual information while the ICIE provides a realistic first-person perspective in this area as shown in Fig. 6.11.

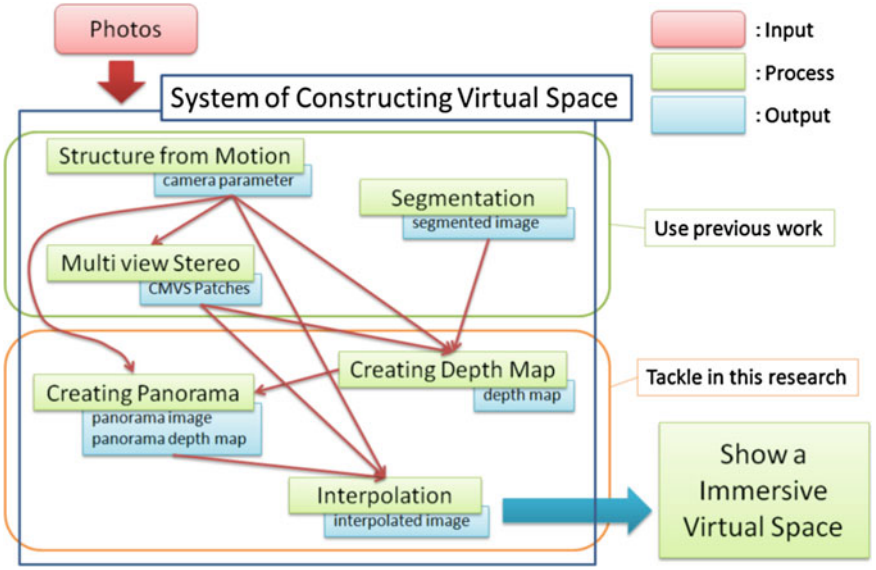


Fig. 6.11 Overview of the wide-range virtualizer (Mori et al. 2011). © 2011, IEEE. Reproduced with permission

The small-space virtualizer captures the visual scenes for small spaces, such as indoors, including human body motions in the space. It consists of two components for capturing static objects and dynamic objects, respectively. The static object capturing component reconstructs static virtual objects in detail from the time-series data, which contain depth data and image data captured by a range sensor and a RGB camera, respectively. The dynamic object capturing component reconstructs the 3D time-series data of dynamic objects from the 2D time-series data captured by multiple range sensors. Human body motions can be captured by the dynamic object capturing component. The body motion data comprise the 3D time-series skeleton data. This system allows for the observation of social interaction in the real world from an arbitrary perspective. We can also simulate a particular social interaction. This sub-system forms part of the 3DCCbyMK system.

The human-body virtualizer creates human-like virtual avatars based on the captured human surface and skeleton data. Using this system, we can change the information depending on the individual in the virtual space. It is important to be able to control and investigate the influence of human appearances in social interaction.

6.4.2 Obtaining Social Interaction Behavior

To capture the omnidirectional motions of the operator with low cognitive and physical loads to derive social interaction based on detailed behavioral data, we need to address several problems. First, the ICIE should be able to capture the operator's interactive behavior using a non-contact motion capture system with low physical load. Second, the immersive conversation space is often narrow and closed. Third, the operator's background image changes dynamically owing to the 360° vision projected in the environment. Under these conditions, social interactive behavior must be captured.

Our motion capture system for narrow and closed spaces uses multiple range sensors to obtain a 3D point cloud, RGB images, and human skeleton data without contact sensors. The system uses the captured omnidirectional motion data obtained from four range sensors placed in complementary positions around the immersive display, as shown in Fig. 6.12 (Ohmoto et al. 2013). The voice of the operator is captured by a light-weight headset microphone or zoom microphone.

6.4.3 DEAL: A Platform for Constructing the ICIE

Several types of systems are needed to observe human-agent interaction, obtain the data, and analyze social interaction based on objective information. For this purpose, Ohmoto et al. (2013) developed a system design platform which is named "Distributed Elemental Application Linker (DEAL)", as depicted in Fig. 6.13. DEAL, which was developed to extend the GECA framework (Huang et al. 2008), can

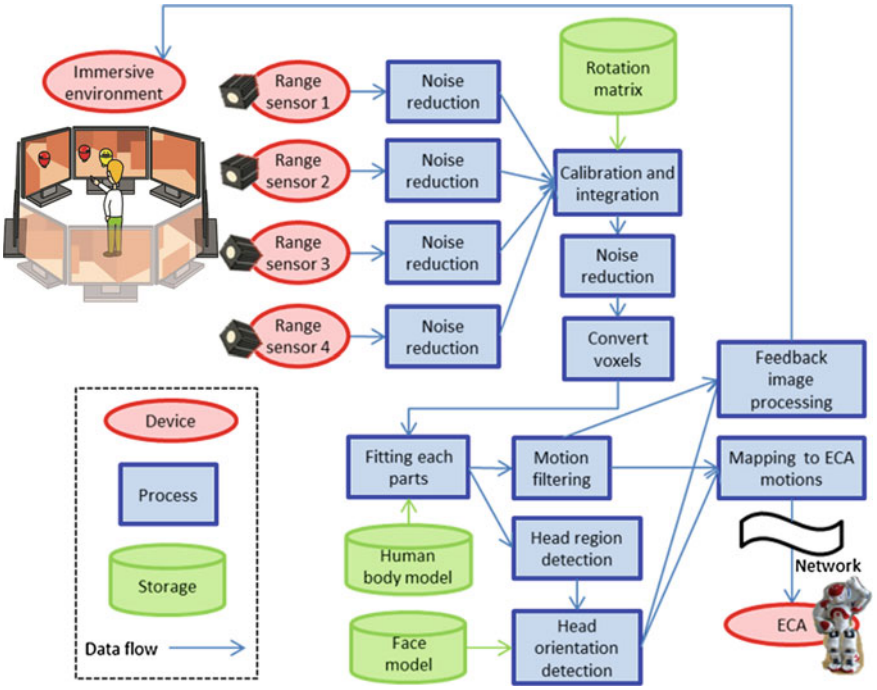


Fig. 6.12 Overview of the system which obtains social interaction behavior in ICIE. Adapted from (Ohmoto et al. 2013). © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

flexibly integrate modules of functions, such as capturing social interactive behavior and providing a first-person perspective, in different network places and different configurations.

DEAL implements functions through two types of plug-in modules: a "function plug (FP)" and "control jack (CJ)". An FP is implemented in a single general function of the ICIE such as obtaining sensor data, integrating different sensors, and displaying the results. It is similar to an encapsulated class in object-oriented programming and can work as an independent application. A CJ is implemented in an interfering network, in which various FP modules are constructed. To share plug-in module data, we use a blackboard model, a methodology widely used in distributed and large-scale expert systems. The basic idea is to use public shared memory from/to which all knowledge sources can read and write information. The FP sends provided data to the blackboard and obtains the required data from the blackboard, while a CJ can change the data on the blackboard. The blackboard can be accessed through DEAL's blackboard manager.

DEAL systems can connect with each other using the blackboard as a network interface. Each blackboard in DEAL can be accessed through the DEAL blackboard manager. The interconnection serves two purposes: The first is as a remote-controlled interface so that a DEAL system can directly execute FPs included in

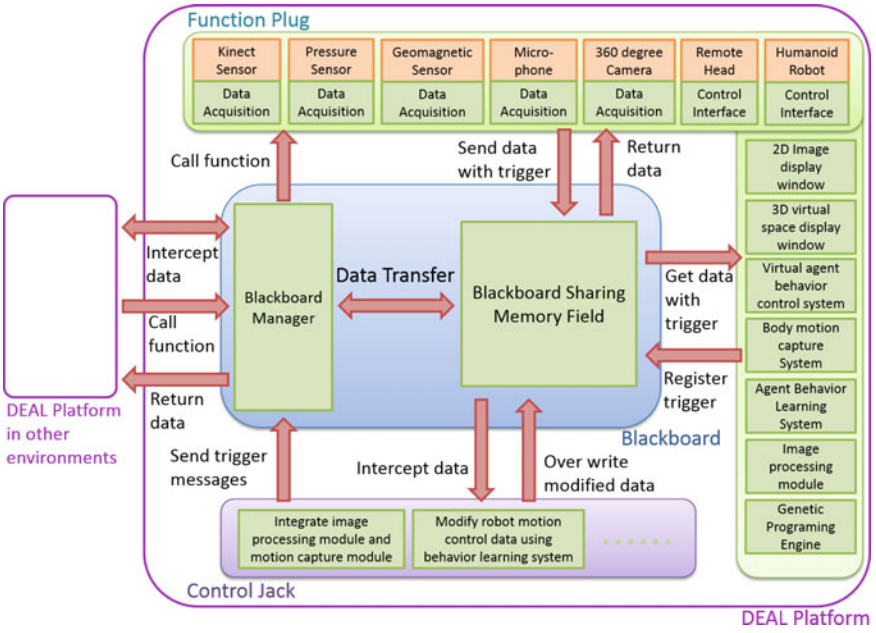


Fig. 6.13 Overview of the DEAL architecture. Adapted from Ohmoto et al. (2013)

an interconnected DEAL system, through the network. The other is as an intercept interface; the blackboard manager provides two interfaces through CJs for reading and writing data on the blackboard. These functions of the blackboard manager are easily realized to cooperate and collaborate with each of the completed applications and functions in DEAL.

DEAL systems can be used not only in the ICIE, but also in other environments, such as for measurement in an open space and data analysis. In these cases, FPs serve as interfaces to connect physical devices and independent systems while CJs are the modules for processing the data provided by the devices and systems. For example, we introduce an environment for measuring human behavior in an open space in Chap. 8. Systems using the ICIE and DEAL systems can handle the measured data without changing the components of the systems if the data format is the same. In another situation, if diverse machine learning methods have been implemented as FPs, a system using machine learning could simultaneously evaluate the performance of learning methods using DEAL systems. DEAL systems can also be used as a platform for integrating distributed resources.

6.5 Application of the ICIE

The immersive conversation space realized by the ICIE plays a critical role in human-agent interaction tasks where a first-person view is required by the user to provide a better service or to investigate how s/he behaves under varying situations. In the filming agent project, where the goal is to build a film shooting robot that can interact with a human expert to record key features of his/her work, it is crucial to build a corpus embodying how an experienced cameraman would interact with a human expert at each stage of the given task. In the cooperative multi-agent interaction project, where the goal is to design a communication framework for human agent collaboration, it is critical to show the participant a first-person view of the collaborative chasing game to obtain detailed data encompassing how the participant may interpret and produce social signals for interacting with agents under varying circumstances. In the tele-presence system project, we exploit an immersive communication space to build a novel communication environment in which human operators can interact directly with individuals through physical avatars.

6.5.1 Filming Agent

In the filming task, knowledge of the specialist work is needed to adequately record the work. In addition, it is not obvious how to use the knowledge in the filming task. To obtain the knowledge for this work and ways of using the knowledge, we need to analyze actual interaction data during a filming task. We therefore used the ICIE as a wizard of OZ (WOZ) system, where the filming robot was controlled using the ICIE. In this way, we obtained ways to shoot and use the knowledge at the same time through the interaction between the specialist and the controlled robot whose operator had knowledge about the specialist work. By using the interaction data, the filming agent learned ways to shoot depending on the performance of the specialist, based on the knowledge of the work and multi-modal data such as the instructor's gestures, gaze directions, and physical relationships between the instructor and specialist.

To implement a filming robot, we first constructed a learning model to acquire action rules and task specific knowledge over time. This model contains three phases: the task analysis phase, WOZ phase, and on-the-job training (OJT) phase. In the task analysis phase, human (specialist) behavior in the task is analyzed to find methods for encoding the measured behavioral data for machine learning. In the WOZ phase, a robot learns the basic action rules and typical task specific knowledge based on the measuring and encoding data obtained through the WOZ controlled robot-human interaction. In the OJT phase, the robot automatically interacts with the human. When encountering a novel situation in the OJT phase, the robot speculates appropriate action rules based on the task specific knowledge obtained in the WOZ and OJT phases. If the robot makes a mistake, the instructor controls the robot remotely and modifies the robot behavior. Based on the controlled data, the robot learns corrected

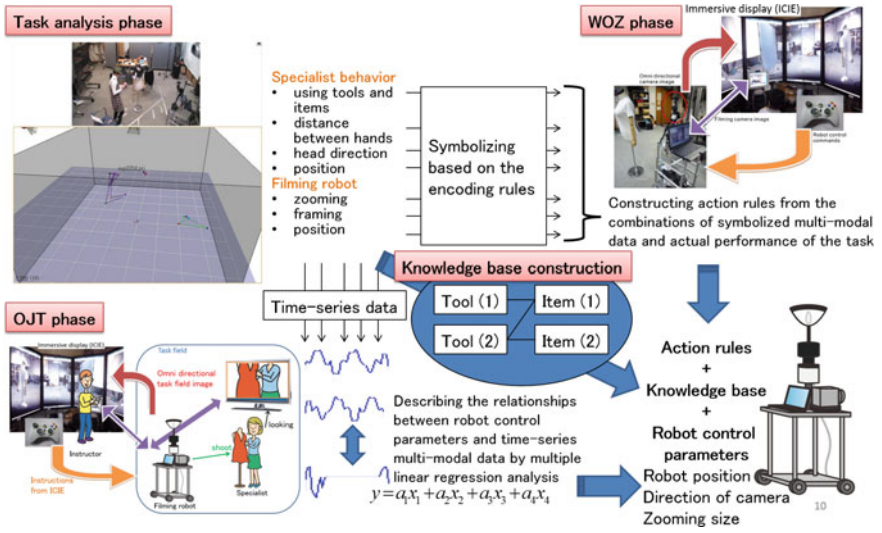


Fig. 6.14 Overview of the phases to acquire action rules and task specific knowledge over time. © 2014, Ohmoto and At, Inc. Reproduced with permission

action rules. After the modification, the action rules and knowledge database are sequentially updated.

The ICIE is used in the WOZ and OJT phases. Since it can capture the user's behavior from a first-person perspective, we can use the ICIE to analyze the operator's behavior in the WOZ phase and to simulate the filming robot's first-person perspective in the OJT phase (Fig. 6.14). In this task, the operator needs to control at least four devices, namely, a camera for filming, a controller for moving the robot, cameras to survey the area around the robot, and a camera providing a whole space image including real-time position data of the working specialist and the filming robot. It is thus hard for the operator to concentrate on the interaction without an immersive environment. We placed an omnidirectional camera on the filming robot, which projected onto the immersive display in the ICIE, and substituted them for cameras surveying the area around the robot and a camera providing a whole space image. The filming camera was placed on top of the filming robot and the camera image was projected onto another window on the immersive display. Since the projected window follows the operator's head direction, the operator needs to consciously control only the filming camera and the controller for moving the robot. Therefore, the operator can control the filming robot through the ICIE as if the robot were part of his/her own body. In other words, since the operator needs to control the devices based on his/her first-person perspective, the ICIE can decrease the physical and cognitive load in controlling the devices, allowing smooth interaction with the other participants.

We implemented the learning model in a robot and evaluated it through an experiment to record handicraft decoration. Figure 6.15 shows the experimental setting including the locations of the researcher, the specialist, and the robot. In this exper-

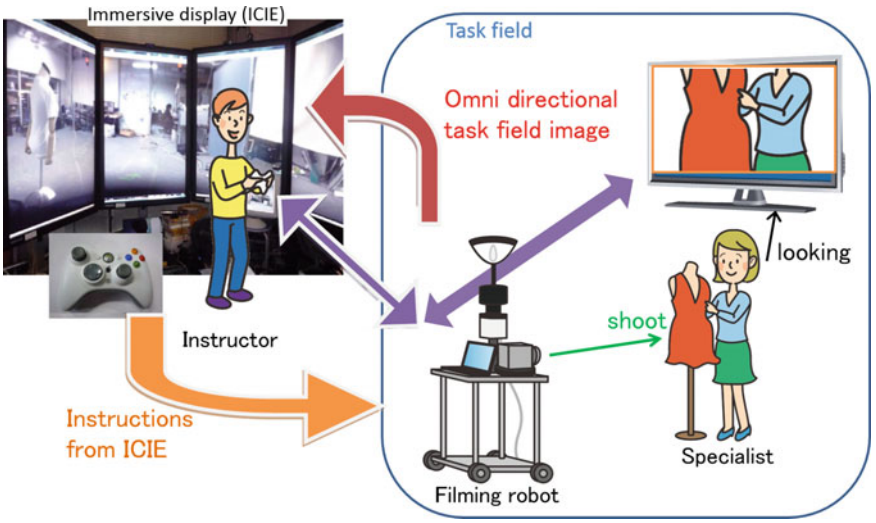


Fig. 6.15 Experimental settings to evaluate the learning model. © 2014, Ohmoto and At, Inc. Reproduced with permission

iment, task-specific knowledge refers to the relationships between the tools used in making the handicrafts and decorating objects and associations between the relationships and multi-modal data of a human involved in the task. We used ICIE in the WOZ and OJT phases. From the task analysis, we determined focused multi-modal data of the human to understand the task conditions, and we classified filming behaviors into three filming modes, namely, tracking hand(s), shooting a certain region, and recording the entire object. In the WOZ phase, we associated filming modes with a combination of modalities, whereas in the OJT phase, we clarified highly correlative modalities with robot controlling parameters. Finally, we compared the knowledge learned by the robot during the experiment and that of the handicraft workers. In this task, we used five tools and five decorating objects. Once all the action rules had been learned through WOZ interactions, we carried out 25 interactions. However, we only conducted five WOZ interactions and five OJT interactions to acquire appropriate filming behavior in the experiment. Furthermore, the knowledge the robot learned differed from that of the handicraft workers. We showed the usefulness of this learning model through an evaluation.

6.5.2 Cooperative Multi-agent Interaction

The goal of this study was to observe and analyze the synthetic use of verbal and nonverbal information in a chasing task involving multiple agents in the ICIE. The synthetic behavior can be observed in an environment in which humans can use verbal

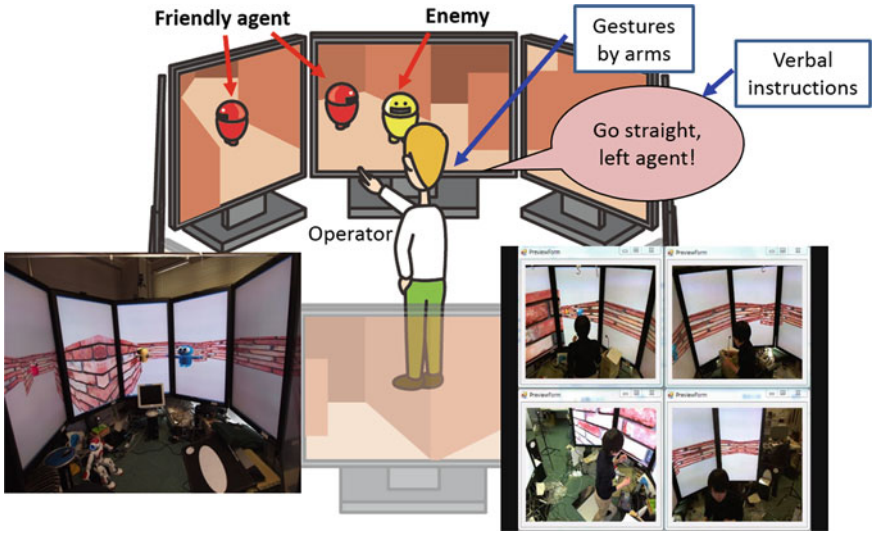


Fig. 6.16 A scene in the chasing task of the experiment. © 2014, Ohmoto and At, Inc. Reproduced with permission

and nonverbal information with low physical and cognitive loads. An experimental environment was built on top of the ICIE.

To investigate a method to interpret a human's instructions synthetically using verbal and nonverbal information, we conducted an experiment that involved a chasing game in which multiple agents interacted with participants. The task was conducted in an immersive display using the ICIE. In the task, a participant instructs two friendly agents to chase an enemy agent using keywords and body motions. Figure 6.16 shows a scene in the experiment. A participant joins the task as a player with the same body as the agents and the same first-person perspective and provides instructions in a similar way to that in the real world. The recognition system with the ICIE identifies the agent to which the participant has provided instructions based on the head and body directions, objects in his/her perspective, and body motions. Since the enemy can move faster than friendly agents, the participant needs to formulate a plan to capture the enemy. We analyzed the verbal and nonverbal information provided by the participants while instructing the friendly agents in an efficient manner.

Consequently, we confirmed that the role of synthetic use of verbal and nonverbal cues strengthened the presented information that the participant wished to emphasize. In addition, we classified the synthetic instructions into three specific categories ("avoiding ambiguities," "adding new meaning," and "emphasizing verbal or nonverbal expressions") and others (such as continuous instruction). Moreover, we constructed a decision tree to classify the synthetic cues into the categories using keywords, gestures, positions of the enemy and friendly agents, and the sequence of instructions. We are currently expanding the algorithm and developing an agent

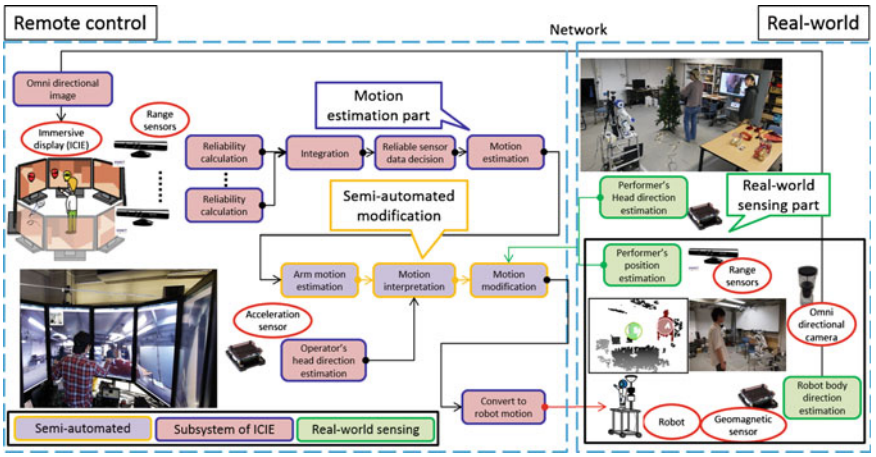


Fig. 6.17 Overview of the tele-presence system using the ICIE. © 2014, Ohmoto and At, Inc. Reproduced with permission

that can interpret the instructions depending on the situation based on estimating a participant's planning model to perform the task.

6.5.3 Tele-presence

Tele-presence can be distinguished from a video chat system by using a physical body to express nonverbal emotion and to interact with real-world objects. The physical avatar is not only a physical body that can be controlled by the operator, but also an "agent" that can mediate interaction between the operator and real-world people and objects. For this purpose, the tele-presence system needs to present a first-person perspective of the controlled robot body to the operator and to express the operator's social expressions using body motions through the physical avatar. The ICIE can satisfy these conditions with low cognitive and physical loads.

We used Nao robots as physical avatars in this study where an operator shares the first-person perspective of the Nao robot using an omni-directional camera placed on the Nao's head position, as shown in Fig. 6.17. The operator's body motions (body direction and arm movements), captured by four range sensors, reflect the Nao, but he/she cannot feel any feedback of the motions. Alternatively, the Nao's full body image is taken by the ambient camera placed at the remote location and provides the operator with a window following his/her head direction. The operator's body direction controls the movable carriage on which the Nao stands. The operator's arm movements are used to control the Nao's arm movements. Since the range sensors cannot capture the operator's head direction, we use a geomagnetic sensor placed on the operator's head, the measured data of which reflects the Nao's head direction.

Although the Nao's appearance is close to that of a human, the operator cannot fully express his/her intentions by using the avatar's body in real-time tele-presence communication. To solve this problem, we designed the tele-presence avatar not as a faithful reproduction of the operator, but as an "agent" that has autonomous abilities to communicate to some extent with real-world objects. This means that the tele-presence is real-time collaborative interaction between the operator and the "agent." In this approach, the operator must become proficient in collaborating with the "agent" avatar. On the other hand, the operator can avoid miscommunication caused by faithful reproduction of the operator's motions, such as a gap in a pointing target and expressions that contain large semantic differences.

We developed a new telepresence system that semi-automatically generates interactive behavior based on the operator's behavior recognition and the remote location sensing data to avoid miscommunication. We conducted an experimental observation using the ICIE to detect what impairs communication in collaborative work. Based on the results of the observation, we focused on three situations: correcting and converting an explicit pointing gesture, expressing backchannel by turning the robot's head, and filtering motions of the robot depending on the task conditions. We implemented a remote sensing component to estimate position and head direction of the collaborative workers and a semi-automatic controlling component generating motions by integrating the operator's behavior with the remote sensing data.

To evaluate the implemented component, we conducted an experiment involving the task of decorating a Christmas tree. In this task, the Nao, which is controlled

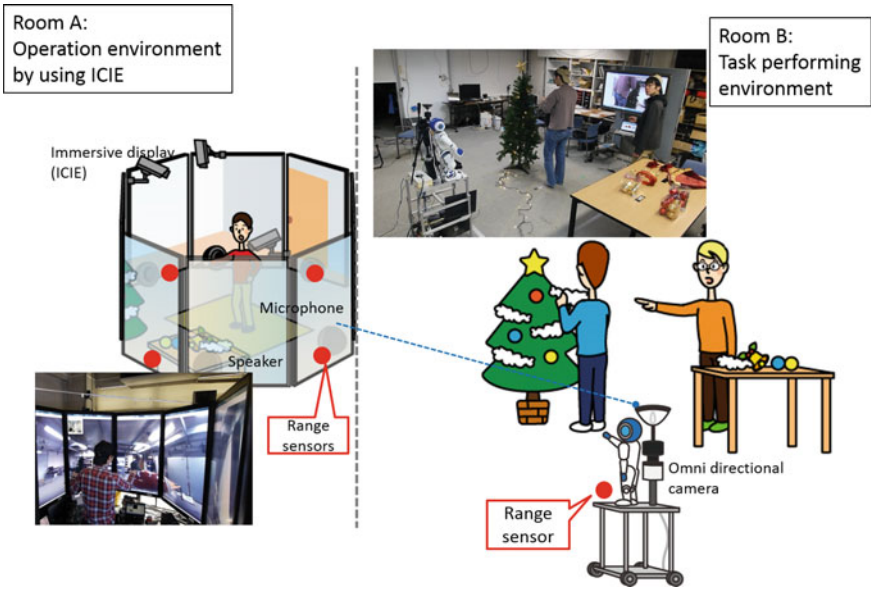


Fig. 6.18 Experimental setting for Christmas tree decoration task. © 2014, Ohmoto and At, Inc. Reproduced with permission

from the ICIE, acts as the instructor to decorate the Christmas tree and two human participants actually decorate the tree in a remote place, as shown in Fig. 6.18. The telepresence system modifies the robot head direction and pointing gestures based on the recognition of the participants, placement of the decorative objects, and the state of the tree. We evaluated whether the semi-automatic telepresence system could reduce miscommunication by video analysis and questionnaires. According to the video analysis, when participants interacted with the robot without the semi-automatic controlling component, they expressed unnatural nonverbal behavior. For example, some of them used both hands when pointing to a particular place, while all of them used a different hand depending on the pointing direction, for example, using the left hand when pointing towards the left and vice versa. On a number of unnatural nonverbal behaviors there is a significant difference between the results of the semi-automated system and copying the user's motion. The results of the questionnaires show that participants understood the intended behavior of the operator better when interacting with the robot with the semi-automatic controlling component than without the component. A significant difference was found. These results show that the semi-automatic component is effective in reducing miscommunication in tele-communication.

6.6 Summary

The smart conversation space implements the theory of conversation enhancement, which is a formalization of the idea of a primordial soup of conversation. We presented the architecture of smart conversation space. We have introduced two types of equipment, i.e., an open conversation space enhanced by augmented reality and an immersive conversation space, to project situations and activities in the shared virtual space. We introduced the situated knowledge media and 3DCCbyMK to implement the idea of an open conversation space. The situated knowledge media is an early implementation of the smart conversation space to enhance online customer service not only to benefit information consumes but also empower information producers. 3DCCbyMK is a three-dimensional conversation capture that uses the Kinect technology to measure and record the entire conversation space including multiple persons and the surroundings. Then, we have introduced ICIE to implement the idea of an immersive conversation space that can augment, record, and measure conversation in a fully controlled immersive interaction environment. A fully-immersive first-person perspective of the physical world is captured and reproduced by three subsystems: a wide-area virtualizer, small-space virtualizer, and human-body virtualizer. The audio-visual behaviors of the user in the ICIE operating cell are captured the omni-directional motion capture and a headset or zoom microphone. DEAL is a software platform that allows for extension with function plugins and control jacks. Finally, we presented three implemented scenarios using ICIE: the filming robotic agent scenario, the cooperative multiagent interaction scenario, and the tele-presence scenario.

Chapter 7
Computer Vision Techniques for Conversational Interaction

Abstract In this chapter, we focus on the computer vision, image understanding and image synthesis approaches related to develop the conversation systems, namely, finding human faces, recognizing facial expressions and gestures, and synthesizing facial expressions and body gestures. In Chaps. 2–4, we introduce the theory, history and techniques for organizing conversation systems, then show the concept of the conversational quantization in the Chap. 5. In developing these systems in the real world, visual information is one of the most important input because it can be applied for varieties of tasks including the recognition of conversational interactions, and the capturing and synthesis of conversational contents.

Keywords Visual recognition · Visual synthesis · Face detection and analysis · Gesture recognition and synthesis · Character animation

7.1 Human Emotional State Recognition Through Visual Recognition Technology

We can perceive people's emotional states through visual information such as facial expressions, postures, and behaviors. For example, when one is disappointed, his shoulders drop, gaze is lower, and face looks sad; thus, human emotional states can be recognized by combining several types of visual information.

Visual recognition of human emotional states has been extensively studied. This is because visual information can be easily obtained using cameras; therefore, easy-to-use and cost effective. Other human emotional state recognition techniques, such as skin conductivity and heart rate, require relatively expensive sensors attached to the body. In this chapter, we introduce the foundations of the following topics in visual recognition/synthesis techniques of human behaviors:

- face detection techniques
- facial expression, recognition, and facial expression descriptors
- facial expression synthesis
- gesture recognition, gesture descriptors, and their synthesis.

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_7

7.2 Face Detection Techniques

Face detection involves detecting human faces in a scene image, and much research has been done on this. This technique has recently become one of the most widely used techniques in many image recognition applications, such as in consumer electronics, security, human computer interaction, and human robot interaction. In face detection studies, a variety of techniques has been developed to (1) improve detection accuracy and (2) reduce computational cost (fast detection). One of the first studies was reported by Kanade (1973). After this work, a lot of considerable studies have been conducted (Yang et al. 2002; Hjelmås and Low 2001).

The traditional approaches are based on the template matching. In these techniques, face templates or facial-part templates are first prepared and find them from the input image through searching throughout the entire image (Fig. 7.1). Image similarity between face templates and image pixels are evaluated using normalized cross

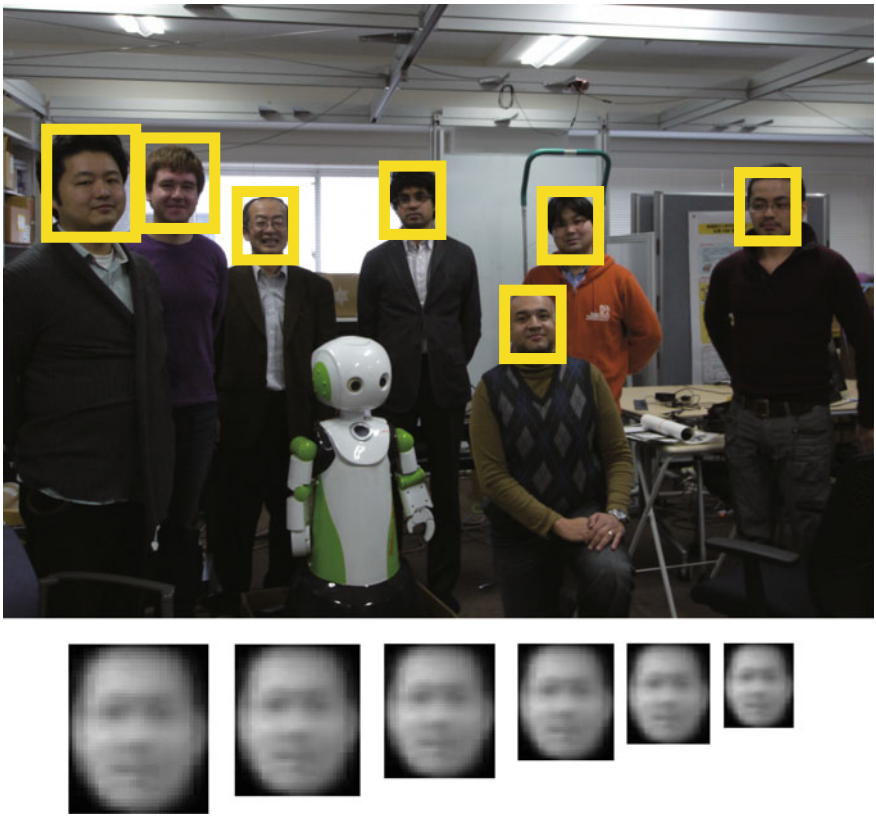


Fig. 7.1 Face detection using template matching method. Face templates (*bottom*) are searched throughout entire image and find facial regions (*top*). Multiple scale of templates are used for adapting for different sizes of faces

correlation (*NCorr*), sum of squared differences (*SSD*), sum of absolute differences (*SAD*).

$$NCorr(X, Y) = \frac{\sum_{x=0}^{X_t} \sum_{y=0}^{Y_t} \mathbf{I}(X + x, Y + y)\mathbf{I}_t(x, y)}{\sqrt{\sum_{x=0}^{X_t} \sum_{y=0}^{Y_t} \mathbf{I}(x, y)^2 \sum_{x=0}^{X_t} \sum_{y=0}^{Y_t} \mathbf{I}_t(x, y)^2}} \tag{7.1}$$

$$SSD(X, Y) = \sum_{x=0}^{X_t} \sum_{y=0}^{Y_t} |\mathbf{I}(X + x, Y + y) - \mathbf{I}_t(x, y)|^2 \tag{7.2}$$

$$SAD(X, Y) = \sum_{x=0}^{X_t} \sum_{y=0}^{Y_t} |\mathbf{I}(X + x, Y + y) - \mathbf{I}_t(x, y)|, \tag{7.3}$$

where $\mathbf{I}$ and $\mathbf{I}_t$ are the intensity of a target image and the template image, and X_t and Y_t are width and height of a template. The term *NCorr* corresponds to higher value and *SSD* and *SAD* produce smaller responses if the target image and template are similar. Therefore, face detection can be carried out by determining the position (X, Y) where the responses of these evaluation function are higher (lower) than the predefined threshold. This approach is simple to understand and implement; however, there are many application issues. First, it requires many templates to adapt to the scale and rotational variations. Because this template matching requires sweeping the template over the entire image region, it requires a large amount of computational cost. Though templates should represent the 'general' features of human faces, it is not easy to obtain such templates.

Extending this technique, in particular, addressing the difficulty in obtaining an optimal template, machine-learning-based approaches have been proposed. Rowley et al. (1998) developed a neural-network-based approach for this task illustrated in Fig. 7.2. First, many facial regions are collected and normalized into 20×20 pixels, which become the normalized templates of the facial images. From each template, an input feature vector is obtained, namely the pixel intensities in the image are assumed as one dimension of the feature vector. In a template of 20×20 pixels, the number of dimensions of the vector becomes 400. Then these feature vectors are used for learning a neural network. Each element of the vector corresponds to a leaf node of a neural network and performs back-propagation learning. In the recognition step, the input image is transformed to several scales of images for the purpose of adapting to the scale difference of the facial regions. Then, sub-regions from the image are taken and input to the learned neural network to verify whether the template has facial features. With the machine learning technique for obtaining optimal facial templates, this neural network-based approach can solve the issue of finding good templates for recognition; however, it still requires a large amount of computational cost to adapt to scale/rotational change; therefore, real-time processing is not possible.

To address the above issues, approaches using Haar-like features and weak classifiers have become widely used recently (Viola and Jones 2001). The Haar-like features are discrete representations of Haar wavelets that obtain responses of spatially

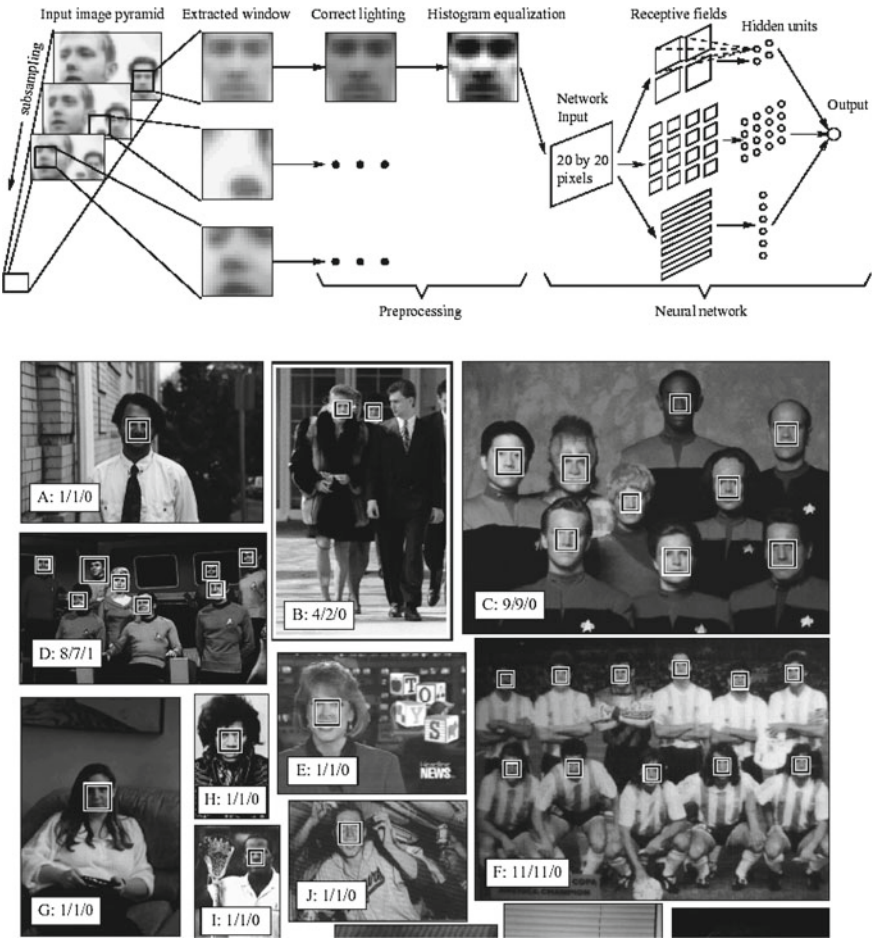


Fig. 7.2 Face detection using neural network-based method (Rowley et al. 1998). © 1998, IEEE. Reproduced with permission

local derivatives in variable scales, rotations and spatial patterns (Fig. 7.3). This approach is based on the idea that people can recognize faces if the facial parts form particular spatial relationships, as shown in Fig. 7.3 (center). By combining these filter responses using *weak classifiers*, they can achieve real-time recognition because (a) the Haar-like feature can be easily and efficiently computed through a fast computation method using *integral image* techniques, and (b) the weak classifier can be also computed efficiently because it is a simple linear sum of the feature responses.

This approach is constructed as follows. In the learning step, many Haar-like features are calculated for face images. Figure 7.3-(left) shows several examples of Haar-like features that output the sum of (a) the pixel intensities within the white rectangles and (b) negative pixel intensities within the black rectangles. This value can

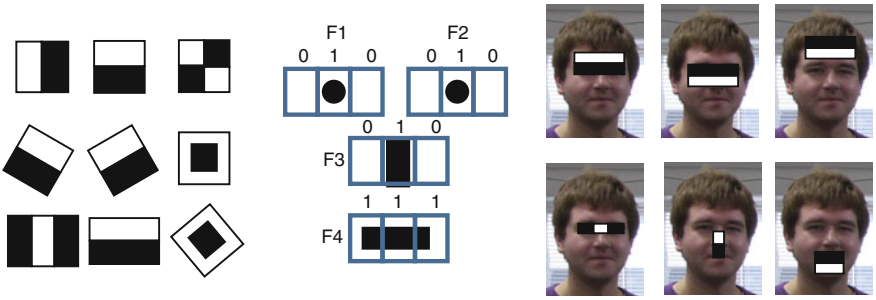


Fig. 7.3 Haar-like features produce local delivatives of the pixel intensities in image template regions. If we apply the haar-like feature of (h) for the *right-hand* image, {F1, F2, F3, F4} becomes $\{+, +, +, -\}$. Multiple Haar-like features are combined for the actual recognition tasks, such as human faces

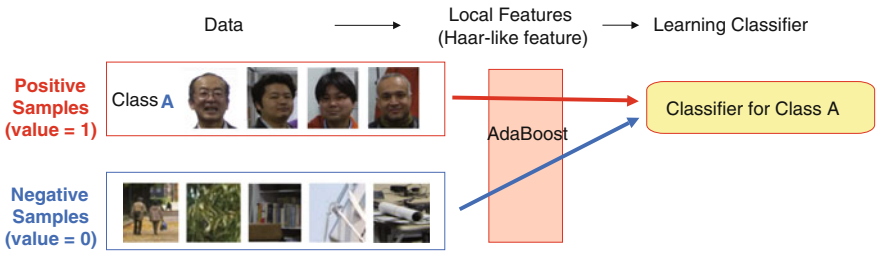


Fig. 7.4 AdaBoost technique for learning facial detectors using Haar-like features. Features obtained from positive samples (face data) and from negative samples (other objects) are used for learning the weak classifiers through AdaBoost algorithm

be quickly computed by using integral image techniques. As a result, we can obtain a feature vector whose dimensions are the same as the number of Haar-like features. Then, these features are learned using weak classifiers, which are the combinations of many simple classifiers, such as thresholding in one dimensional space, by using the AdaBoost technique (Freund and Schapire 1997). Due to the high demand of this technique for consumer devices and security systems, many extensions have been proposed such as combining several Haar-like features as a one feature (Mita et al. 2005) or more complicated Haar-like feature windows (Fig. 7.4).

7.3 Recognition of Facial Expressions

The human face is the most important body area for expressing human emotions. We can detect many signs reflecting human emotions such as facial expressions, gaze, or skin color, in particular, facial expressions play the most important role (Zoric et al. 2007). Facial expressions can be described by the movements of facial parts related

to the movement of mimic muscles in the face. Therefore, the recognition of facial expressions requires *facial part recognition and tracking* methods. On top of the facial part tracking, their motions are parameterized (*facial parameterization*) and related to human emotions. In the following subsections, we explain vision-based facial part tracking and facial parameterization approaches.

In facial part recognition and tracking using an image recognition technique, the most successful approach is the active appearance model (AAM) proposed by Edwards et al. (1998) (Fig. 7.5). The AAM approach assumes a human face as a three-dimensional deformable mesh model whose vertices correspond to the facial feature points. Then, a vector containing the location of the vertices $\mathbf{S}$ can be formulated as a weighted sum of the *basic shape* $\mathbf{s}_0$ and the *motion vectors* $\mathbf{s}_1$ to $\mathbf{s}_N$ which represents the deformation of the vertices.

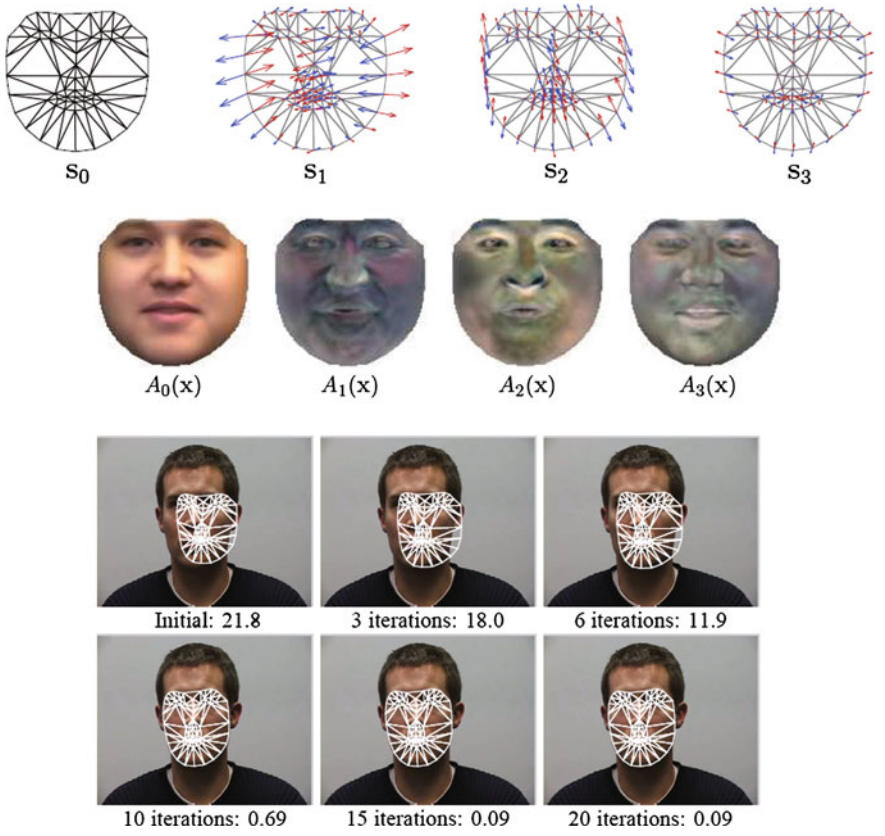


Fig. 7.5 Active appearance model (AAM). *Top* Facial mesh model (basic shape and motion vectors). *Middle* Appearance model. *Bottom* Facial parts tracking using AAM (Edwards et al. 1998). © 1998, IEEE. Reproduced with permission

$$\mathbf{S} = \mathbf{s}_0 + \sum_i p_i \mathbf{s}_i . \tag{7.4}$$

Similarly, the AAM approach uses a face appearance model which consists of a face image in a natural state A_0 and the difference in image intensities A_j while changing the variable facial expressions, producing a similar model to the mesh model.

$$A = A_0 + \sum_j q_j A_j . \tag{7.5}$$

Finally, the warping function that transforms the weight parameters in the mesh model p_i to the parameters in the appearance model q_j is obtained.

In the tracking step, the system iteratively applies parameter estimation to an input image and detects the optimal weight vector $\hat{\mathbf{q}} = (\hat{q}_i, \ldots, \hat{q}_N)$ by using the appearance model, then $\hat{\mathbf{q}}$ is warped to the mesh model and the corresponding mesh model vector $\hat{\mathbf{p}} = (\hat{p}_i, \ldots, \hat{p}_N)$ is obtained. As a result, the AAM approach can estimate the positions of facial parts. The resulting motion vectors and weights $(\hat{p}_i, \mathbf{s}_i)$, with respect to the natural face, are used for the recognition of facial expressions.

7.4 Facial Parameterization

Once we obtain the vector of the facial parts, we can recognize the facial expressions using the input image sequence. Thus, we need a method to recognize human emotions from the movements of the facial parts. For this purpose, several methods has been proposed for the parameterization of facial-part movements and the relationship between the parameters and emotions (Essa and Pentland 1997). In this section, we introduce two major approaches: Facial Action Coding System (FACS) and Facial Animation Parameter (FAP).

7.4.1 Facial Action Coding System

The Facial Action Coding System (FACS) approach was developed by Ekman and Friesen (1978). This study has greatly influenced later studies in facial parameterizations (e.g., Donato et al. 1999). The idea of this approach is encoding the movements of facial muscles as a facial action unit (AU) (Fig. 7.6). For example, AU 1 corresponds to raising the inner brow, AU 2 corresponds to raising the outer Brow, and AU 26: corresponds to dropping the Jaw. However, several AUs do not directly correspond to a facial muscle movement, such as AU 19 (sticking out the tongue), AU 33 and AU 66 (crossing the eyes). By combining these AUs, we can parameterize facial movements obtained from image sequences for the purpose of psychological experiments.

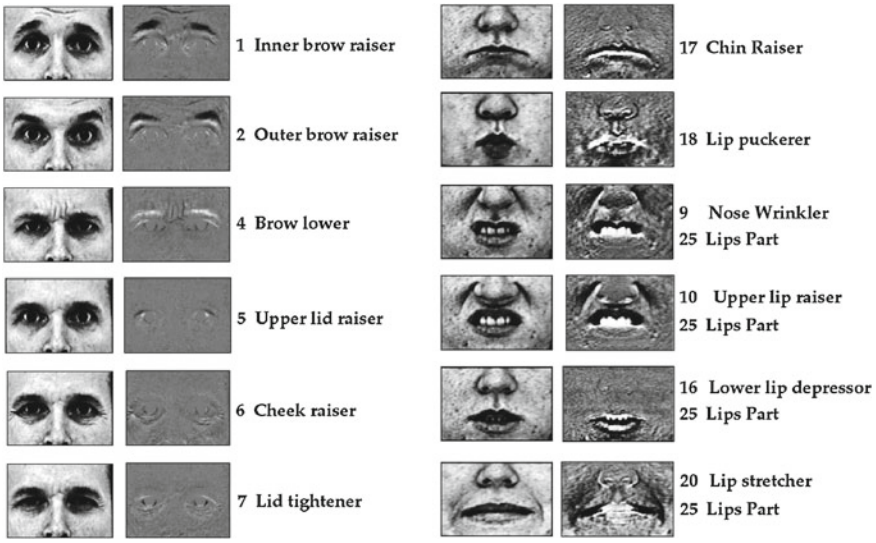


Fig. 7.6 Several examples of the facial action unit (AU) in facial action coding system (FACS) approach (Donato et al. 1999). © 1999, IEEE. Reproduced with permission

7.4.2 Face Animation Parameter

Face Animation Parameter (FAP) was developed by the Moving Pictures Experts Group (MPEG) as one of the MPEG-4 components (Pandzic and Forchheimer 2003). The objective of FAP is generating agents' character animations from the descriptions, including facial animations, gestures, and eye movements of virtual humans and humanoids. The differences between the FACS and FAP approaches are as follows:

- The FAP approach includes not only facial animations, but gestures and head and eye movements.
- The FAP approach is designed for synthesizing the animations of human-like avatars though FACS is developed mainly for just parametrizing facial movements.

With the FAP approach, the neutral state of a face is defined when (a) all face muscles are relaxed, (b) eyelids are tangent to the iris, and (c) pupils are 1/3rd the diameter of the irises. Then, it defines 84 feature points (FPs) on the face and the facial expression parameters are the movements of the FPs (Fig. 7.7). Displacement from the natural state denotes the magnitude of the parameter, which expresses the strength of the facial expression.

The MPEG-4 standard defines 6 primary facial expressions: *joy, anger, sadness, fear, disgust*, and *surprise*. Each expression can be described by the combinations of FAP parameters. For *Sadness*, for example, the corresponding FAP is *close_t_l_eyelid (FAP 19), close_t_r_eyelid (FAP 20), close_b_l_eyelid (FAP 21),*

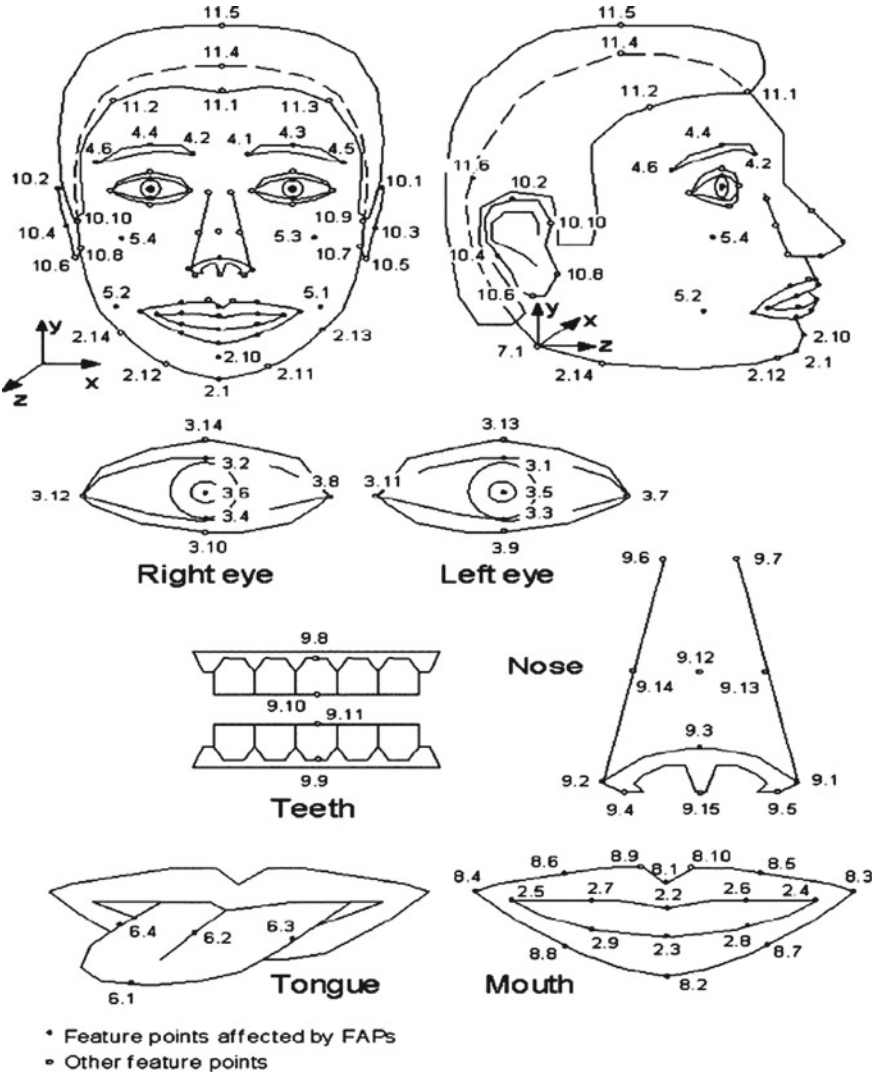


Fig. 7.7 Facial animation parameter (FAP) approach. 84 feature points are created over face and their movements are parameterized (Pandzic and Forchheimer 2003). © 2003, John Wiley and Sons. Reproduced with permission

close_b_r_eyelid (FAP 22), raise_l_i_eyebrow (FAP 31), raise_r_i_eyebrow (FAP 32), raise_l_m_eyebrow (FAP 33), raise_r_m_eyebrow (FAP 34), raise_l_o_eyebrow (FAP 35), raise_r_o_eyebrow (FAP 36).

7.5 Facial Animation Synthesis

There are two types of approaches in synthesizing facial animation. The first approach uses the semantic/symbolic descriptors of the expressions for describing the animation. Namely, a user prepares an AU or FAP sequence as the 'scenario' of the agent's facial behavior, then it is converted to a sequence of FAPs. As a result, the character's three-dimensional facial shape is generated using a deformable mesh model whose displacement of the mesh FPs are synthesized using the FAPs.

The other approach is called expression mapping (expression cloning). With this approach, a user's facial expression is captured using vision or motion capture systems and mapped to a face of a character (Zhang et al. 2003). In the preprocessing step, the facial image is first segmented into several portions according to the similarity in movement. In the synthesis step, the facial FPs are automatically detected and their motions are mapped to the character's face. Surface smoothness is also considered. Similar expression cloning is also conducted by using an RGB-D sensor (Weise et al. 2011). In this previous study, the current facial state was mapped to 39 types of facial expressions by using a real-time RGB-D sensor and animation was synthesized as the smooth transition of these states.

Compared to these two approaches, the symbolic-model-based approach is simple and it is easy to synthesize the facial expressions only from a sequence of facial expression data. The size of face model data and computational cost required to produce the animation is relatively smaller than with the facial cloning approach. Also, this approach does not require users to control the avatars face; therefore, we can directly map the linguistic emotion to the facial expressions. However, the resulting animation is not realistic. The expression mapping approach can produce very realistic facial expressions of arbitrary characters. However, this approach just transforms the users facial expression to the characters. Therefore, it requires a user's real-time control or prerecorded data. Also, the dataset and computational cost is higher than the symbolic-model-based approach. There exist intermediate solutions of these approaches. Zhang et al. (2010) proposed a facial expression synthesis method using a three dimensional PAD (pleasure-displeasure, arousal-nonarousal and dominance-submissiveness) model. Their method can be controlled by the FAP but the parameters of the facial expressions are obtained from the actual human faces and their learning processes. As the result, their method successfully generate the facial animations of talking avatar with varieties of emotional states.

7.6 Gesture Recognition and Synthesis

Gestures are commonly used in human communication for clear and seamless conversation. This is not only for human-human communications but for human-agent communications. In everyday scenes, we use the following gestures using different body parts (Mitra and Acharya 2007):

- hand and arm gestures
- hand poses, sign languages, entertainment applications
- interaction in virtual environments
- head and face gestures
- nodding or shaking of head
- facial expressions
- looks of surprise, happiness, disgust, fear, anger, sadness (facial + body expression)
- body gestures
- involvement of full body motion such as (a) two people interacting outdoors, (b) analyzing movements of a dance or (c) recognizing human gaits.

A gesture recognition method can be created by a variety of sensors. In computer vision research, they can be categorized into several approaches such as types of image sensors, model representations, and recognition methods (Fig. 7.8).

Vision-based gesture recognition has been studied for many years, particularly for automatic recognition of American Sign Languages (Ong and Ranganath 2005). For this task, a variety of sensors have been used including a single image, multiple images, hand and finger sensors (Cyberglove), and range sensors.

Usually, gesture recognition approaches consist of gesture spotting and gesture recognition. Gesture spotting detects the start and end frames from the input sequence and segments out the frames. Gesture recognition retrieves a feature vector from an input motion segment then inputs them to the recognition algorithms.

(a) *Feature vectors*

There are a variety of feature vectors for recognition. For example, in accelometer or gyro sensors, input data are applied for frequency decomposition and separated into meaningful frequency and noise. In an image sequence (video), because it is composed of very high-dimensional signals, feature selection is key for recognition. The most common approach is using edge information. Freeman proposed using

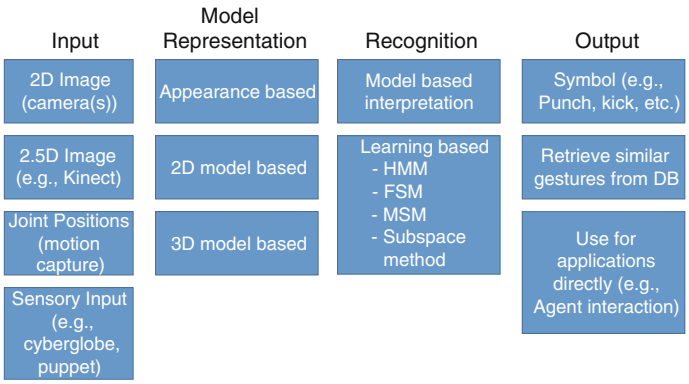


Fig. 7.8 The categories of the gesture recognition methods

orientation histogram features for hand-pose and configuration recognition (Freeman et al. 1994).

Similar ideas are also applied to human body recognition. The histogram of gradient feature (Alal and Triggs 2005) is commonly used in human detection. This approach obtains features frame by frame. Another approach assumes the video as a 3D spatio-temporal volume and finds features from it.

(b) *Gesture recognition methods*

There are two types of recognition methods: time-sequence analysis and non-temporal analysis. The time-sequence analysis uses the temporal relationship of feature vectors. In this type of method, the most important issue is *spotting* and *time-warping*. *Spotting* involves finding the first (and end) frames of gestures. *time-warping* involves allowing temporal stretching and shrinking of the time-series of the input signals.

Dynamic programming (*dynamic time warping*) is a method for evaluating the similarities of two feature sequences (Fig. 7.9). Given two signal sequences $A(i)$ $(i = 1, \ldots, N)$ and $B(j)$ $(j = 1, \ldots, M)$, the distance of these data $D_{A,B}$ can be computed as follows.

$$D_{A,B} = D(N, M) \tag{7.6}$$

$$D(n, m) = \min \begin{cases} D(n-1, m-1) + d(A(n), B(m)) \\ D(n-1, m) + d(A(n-1), B(m)) \\ D(n, m-1) + d(A(n), B(m-1)) \end{cases} \tag{7.7}$$

$$D(1, m) = d(A(1), B(m)) \tag{7.8}$$

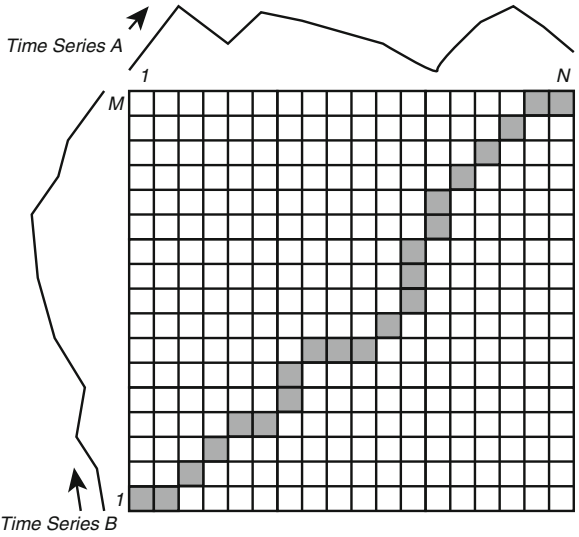


Fig. 7.9 Dynamic time warping method for evaluating time-series of sequence

$$D(n, 1) = d(A(n), B(1)), \tag{7.9}$$

where $d(A(s), B(t))$ is the distance between $A(s)$ and $B(t)$. This method is common in evaluating the time-series of information because it adapts to temporal stretching and shrinking. However, this technique has the following issues.

- It only produces the distance of two time-series of data; therefore, it does not naturally adapt to the individual differences in motions.
- To enable the recognition of a variety of motions in one category, we need to either (a) prepare many samples and find the most similar one to the input data, or (b) the representative sequence to compare, i.e., the average of the samples. However, the former approach is time consuming and the latter is difficult to find nice example motions.

Hidden Markov models (*HMM*) is also commonly used in speech recognition, and applied to gesture recognitions as well. It consists of the state-based graph structure shown in Fig. 7.10. At each state, the state transitions to another state by a predefined probability and outputs the signals. The action class is determined to find the graph that output the highest probability for the input signal. The advantage of HMM compared to dynamic programming is it can adapt to the individual differences. However, the disadvantage is that it requires many examples to learn the motion.

In contrast, the *subspace method* does *NOT* use the temporal information; it uses the relation of the multi-dimensional feature vectors for recognizing the time-series of data and finds similarities of groups of data. In gesture recognition literature, a particular gesture has unique relationships between body portions (joint position or angles); therefore, this relation can be used for identifying gestures. For example, when *punching* behavior, some people mainly move one hand but others do not move. For *gait*, joint movement can be assumed to be a linear system. Therefore, we can create a subspace of human motions that represent the relationship between different body movements to recognize motions. This method is called *subspace-based recognition*, and a variety of such methods have been proposed in *which subspace is used* and *how to use subspace*.

The simplest approach is using principle component analysis for subspace representation. In this approach, the target motion's features, such as joint angles, or image

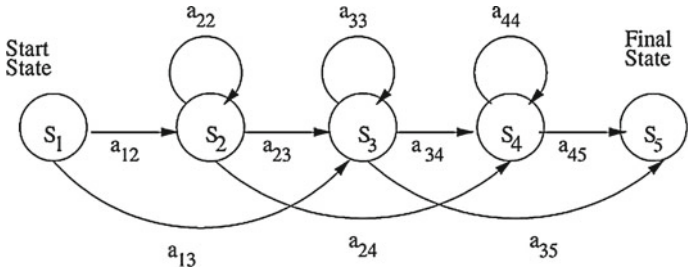


Fig. 7.10 Hidden Markov model (HMM)

features are mapped to a linear space. Assume that the time sequence of the motion features is given as a matrix form $\mathbf{X} = [x_1 \ \cdots \ x_N]$. Now we want to compare the similarity of the input motion $\mathbf{X}$ and a motion in a database $\mathbf{Y} = [y_1 \ \cdots \ y_M]$. First we apply the principle component analysis (PCA) to the database motion $\mathbf{Y}$, and obtain eigenvectors $\mathbf{P} = \left[\mathbf{p}^1 \ \mathbf{p}^2 \ \cdots \ \mathbf{p}^M\right]$. Note that these eigenvectors are arranged in the desending order of their eivenvalues. The subspace of this motion is defined by choosing $M'(< M)$ eigenvectors in the order of largest eigenvalues, namely $\mathbf{P}' = \left[\mathbf{p}_1 \ \mathbf{p}_2 \ \cdots \ \mathbf{p}_{M'}\right]$.

This space describes the major motion of the gesture $\mathbf{Y}$; therefore, gesture recognition is done based on whether the input motion sequences $\mathbf{X}$ lie in this space. The most simple approach is *subspace method* (Fig. 7.11). In the subspace method, the input motion sequence $\mathbf{x}_t$ is mapped to this subspace and distance between the input and reconstructed vectors is evaluated, namely it evaluates the following errors,

$$Err(\mathbf{x}_t, \mathbf{P}') = \left|\left|\mathbf{x}_t - \left(\mathbf{P}'\,\mathbf{P}'^{\mathsf{T}}\left(\mathbf{x}_t - \overline{\mathbf{x}}^{\mathcal{P}'}\right) + \overline{\mathbf{x}}^{\mathcal{P}}\right)\right|\right|, \tag{7.10}$$

where $\overline{\mathbf{x}}^{\mathcal{P}'}$ is the average vector of the eivenspace $\mathbf{P}'$. If the input feature $\mathbf{x}_t$ follows the subspace of the learned motion sequence $\mathbf{P}'$, the error is smaller, if not, the error become larger. Therefore, we can recognize the gesture category of the input motion sequence by finding the gesture in a database whose subspace produces the largest error in the input sequence. Namely,

$$Err_{\mathrm{SM}}(\mathbf{X}, \mathbf{Y}) = \max_t Err(\mathbf{x}_t, \mathbf{P}'). \tag{7.11}$$

However, the subspace method has a major issue comparing two motions, namely, if one motion is a sub-part of the other motion, the distance becomes smaller. This

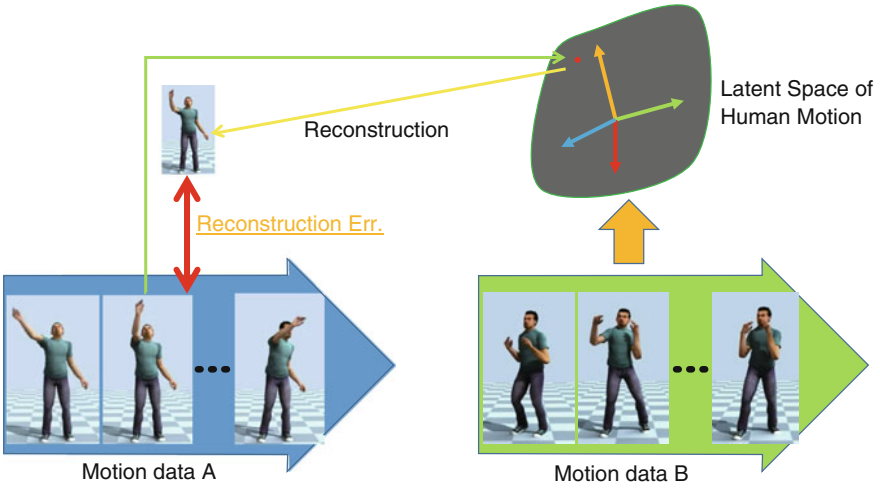


Fig. 7.11 Subspace method for gesture recognition

happens because the subspace construction and projection are performed in one side of the data. To address this issue, Numaguchi et al. (2011) proposed to use the reconstruction error from the samples obtained from database (learned) motion to the subspace of the input motion (dual subspace projection method (DSPM)) (Fig. 7.12). Namely, DSPM uses the following error function:

$$Err_{\mathrm{DSPM}}(\mathbf{X}, \mathbf{Y}) = \max\left(\max_t Err(\mathbf{x}_t, \mathbf{P}'), \max_t Err(\mathbf{y}_t, \mathbf{Q}')\right), \tag{7.12}$$

where $\mathbf{Q}'$ is the subspace constructed from the sequence $\mathbf{X}$.

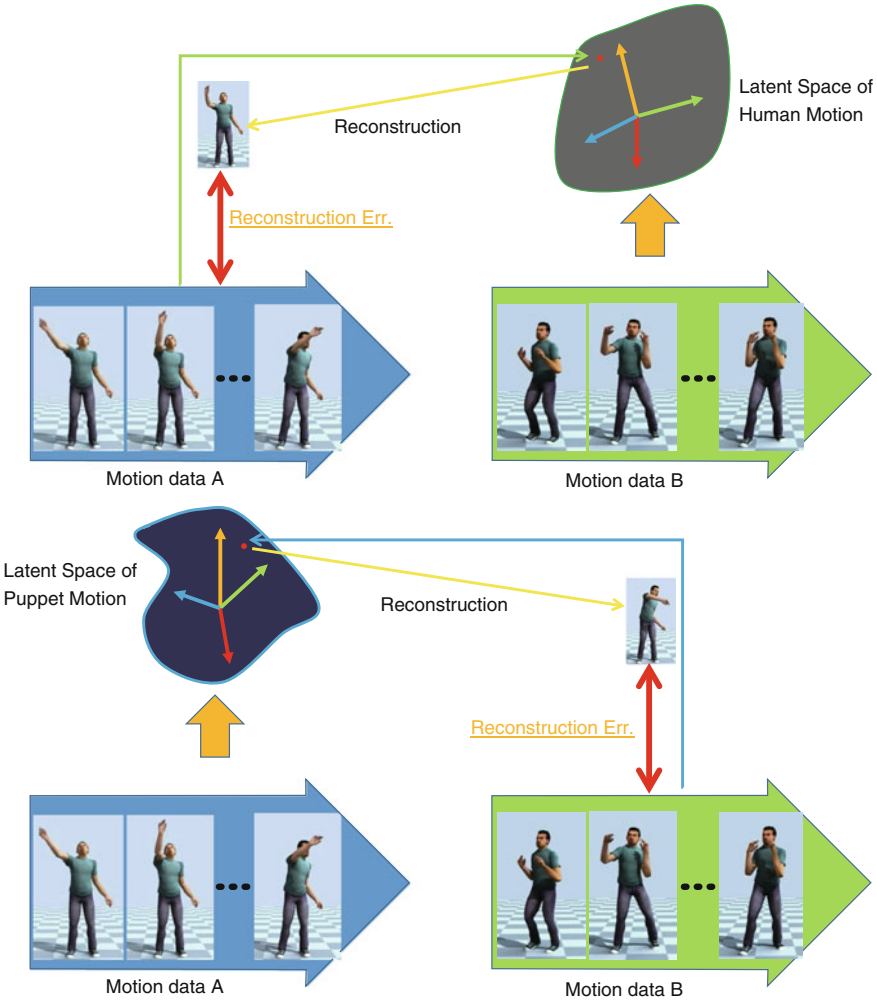


Fig. 7.12 Dual subspace projection method for gesture recognition

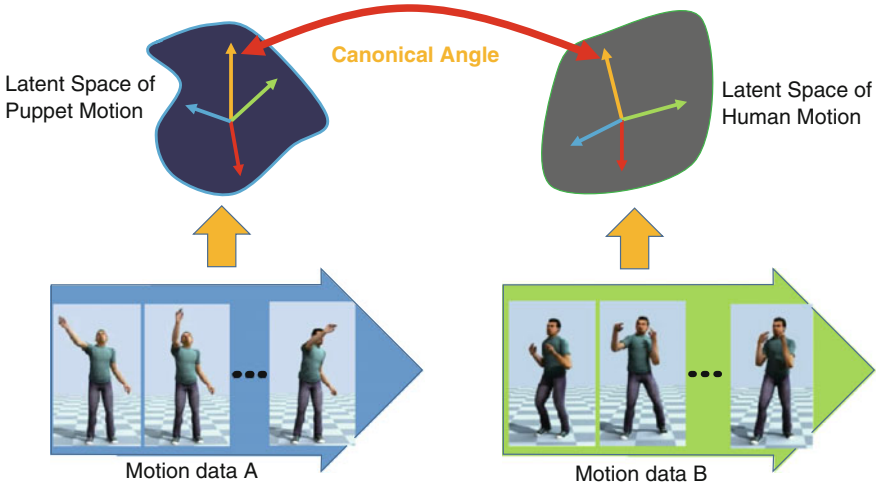


Fig. 7.13 Mutual subspace method for gesture recognition

Another interesting method is directly comparing the obtained subspaces. The mutual subspace method uses the canonical angle of two learned subspaces to compare the time-series of information (Fig. 7.13) (Yamaguchi et al. 1998). The original idea was proposed to compare face image sequences under varying lighting conditions; however, it can be also applied to motion data. The canonical angle is the angle between the engenvectors of the subspaces of the two motions ($\mathbf{P}'$ and $\mathbf{Q}'$). The cosine squared of the first canonical angle ($\cos^2(\theta_1)$) of these subspaces is given by the largest engenvalues of a matrix **PQP**, and this value is used as a similarity metric.

Another method is *bag of motion features*, which extends the idea of the 'bag of features' method in the image-based object recognition method (Laptev et al. 2008). In this previous study, video (image sequence) was assumed as a three-dimensional volume space and space-time interest points (STIPs) were found, which are the high gradient point in spatially and temporally. Using STIPs, the original video is segmented into single actions such as hand-shaking and standing up. From each video segment, spatio-temporal SIFT descriptors (Lowe 1999) are detected from the video volume and used for action recognition. This algorithm is used for movie scenes to find six kinds of gestures including "answer phone", "get out of car", "hand shake", "hug person", "kiss", "sit down", "sit up" and "stand up."

7.7 Gesture Descriptor and Synthesis

Similar to facial expression analysis and synthesis, gesture synthesis can be categorized into two types of approaches: symbol-based and data-driven. The former approach consists of two stages; gesture symbolization and gesture synthesis. There

are several approaches to symbolize human gestures; however, it is still difficult to do this accurately because human gestures have more than 60 DOFs.

7.7.1 Labanotation

Laban and Ullmann (1960) proposed a description method of human body movements called *Labanotation* for describing dance motions. In Labanotation, configurations and movements of the four limbs and body configurations are described using a specialized descriptor, like a musical score, a gesture can be described as a time series of descriptors (Fig. 7.14). Several studies have recently been conducted to (1) automate translation from human body motions captured with a camera or a motion capture system by using Labanotation, and (2) re-produce original motions from Labanotation and synthesized computer animations (Hachimura and Nakamura 2001; Shen et al. 2005; Loke et al. 2005). These efforts are similar to studies on facial expression descriptors, such as FACS and FAP. However, in reality, this is more difficult than facial expression recognition because of (1) the high-dimensionality of human-body motions and (2) limitation of spatial-temporal granularity of Labanotation. As a result, the resulting animation generated from Labanotation is not yet satisfactory.

7.7.2 Data-Driven Approach for Gesture Synthesis

The data-driven approach uses several human motion samples obtained from motion capture data for synthesizing desired human behaviors. Compared to symbolic-based methods, such as Lavanoation, this type of approach does not have foundations of motion descriptors such as symbolic representations. Therefore, motion

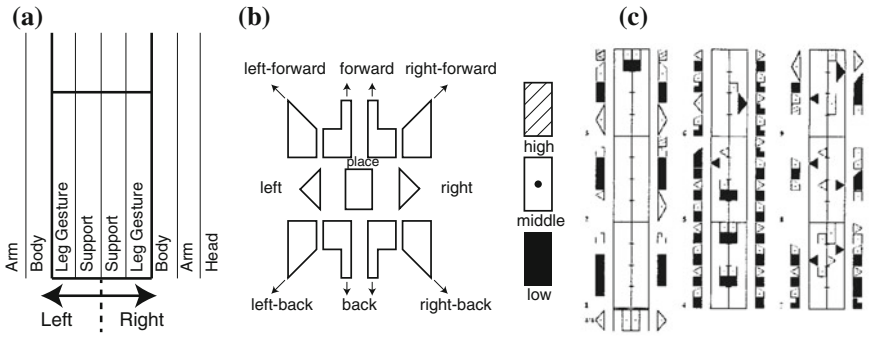


Fig. 7.14 The system of Labanotation. **a** Staff of labanotation. Each column corresponds to the different body portions. **b** Symbols of labanotation, which describes the position of body portions. **c** An example (Hachimura and Nakamura 2001). © 2001, IEEE. Reproduced with permission

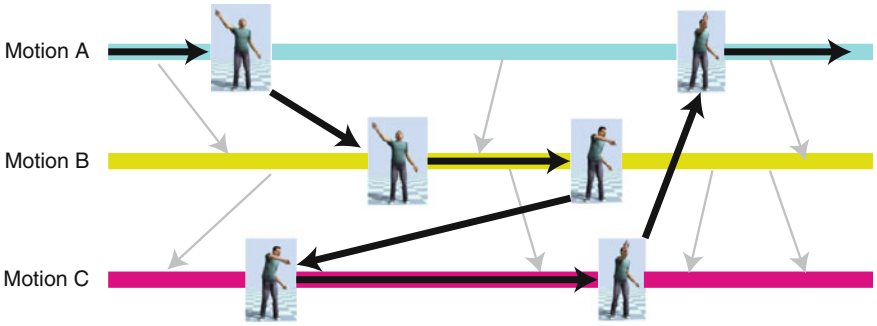


Fig. 7.15 Motion graph approach for arbitrary motion synthesis. First, a graph structure is constructed by finding and connecting similar postures between different motions. Then, desired smooth motion is constructed by traversing the graph according to the user desired criteria such as a path traversal or a music choreography

is synthesized according to user inputs or other signals such as user-defined trajectory (Kovar et al. 2002), speech signals (Stone et al. 2004), sound signals (Shiratori et al. 2006) and real-time user input (McCann and Pollard 2007). Most of these approaches are based on the graph-traversal of motion capture data, which is called *motion graph*. In this approach, a graph structure is first constructed from a a large amount of motion capture data by finding similar posture frames and generating transition frames between them (Fig. 7.15). Motion generation is carried out by traversing the graph. By implementing several rules to the graph traversal, such as following the designed trajectories or temporal coherence to musical beat signals, we can produce motions that satisfy the desired condition. Though this approach cannot produce arbitrary motions from symbolic signals, it can produce realistic motion for desired applications.

7.8 Summary

In this chapter, we introduce visual techniques about human state recognitions and interactions. First, we show several techniques of face detections. In these days, the combination of the Haar-like features and the weak-classifier is popularly used for this task. Then, we introduced the two types of facial expression/synthesis techniques: the semantic/symbolic descriptor-based and the data-driven approaches. For former technique is based on the descriptors of facial parts movements such as FACS and FAP, and the latter techniques use the 2D/3D database of facial movements. We can categorize similarly in gesture recognition/synthesis techniques. Both approaches have advantages and disadvantages. The semantic/symbolic descriptor-based approach is simple therefore good for real-time/mobile systems, however, the resulting picture is not realistic. On the contrary, data-driven approach can generate photo-realistic pictures but requires more computational time and large size of data.

Chapter 8
Measurement, Analysis and Modeling

Abstract Better understanding of conversation paves the way towards better conversational systems. In this chapter we shed light on the practical aspects of multi-modal interaction analysis towards a better understanding of conversation as a phenomenon. On the one hand, investigators need to take great care of methodological issues, since conversation involves plenty of subtleties. Incorporating physiological signals allows us to base our understanding on a more solid foundation than merely depending on audio-visual data, which is extrinsic from the viewpoint of mental processes. Collaborative support tools help annotators share their experiences thereby improving the efficiency and quality of the annotation process. On the other hand, investigators are encouraged to learn from past experiences in terms of how experiments were conducted to derive useful insights. For this reason we report in this chapter three case studies that showcase different issues in measurement, analysis and modeling of conversation and interaction in general.

Keywords Inner state estimation of humans · Experimental design · Methodology for interaction analysis · Physiological signal analysis · Naturalness · Human-robot-interaction

8.1 Methodological Issues in Multi-modal Interaction Analysis

Multi-modal interaction analysis normally consists of multiple phases: experimental planning and design, the preliminary experiment, and full-scale experiment.

8.1.1 Experimental Planning

During the experimental planning stage, investigators need to determine whether they should adopt a categorical or structural approach (Bono et al. 2007) depending on the goals of the investigation. In a categorical approach, the obtained data are

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_8

segmented and classified into categories based on the function of a given communication behavior. In McNeill's categorical approach for gesture analysis, for example, the data are categorized into four groups: iconic, metaphoric, beat, and deictic (McNeill 1992). The categorical approach is often employed when conducting a quantitative analysis.

In a structural approach, investigators aim to extract relationships among segmented data in the interaction. For example, in Kendon's structural approach for gesture analysis, the data are segmented into gesture units, such as home position, preparation, hold, stroke, and retraction, allowing researchers to analyze the relationship between the gesture units (Kendon 1972). The structural approach is mostly employed when conducting a qualitative analysis.

8.1.2 Building Experimental Environment

Based on the chosen approach, which guides how the interaction data are obtained and interpreted, the experimental environment is built. The raw data, which are interpreted during the analysis, are recorded using diverse methods, the most basic of which is video (capturing both audio and visual data). There is a strong relationship between how to record and what to do, although this is not a one-to-one correspondence. For example, even when recording a video, there are many different ways of carrying out the recording depending on the recording targets, for example, facial expressions, gaze directions, body motions (gestures), and walking trajectories. In the experiment, data related to the aim of the investigation should be recorded as many times as possible. However, depending on the experimental conditions and the measurement methods, there are some types of data that cannot be obtained simultaneously. To obtain as many kinds of data as possible, universal experimental environments can be used, which enable the flexible construction of diverse experimental settings.

As discussed in Sect. 6.4, the ICIE serves as a universal experimental environment on which various experimental settings can be built. The ICIE builds a virtual experimental environment by combining various subsystems. Although the user has to scan real-world objects and create specific programs for each experiment, the high flexibility assists in building the experimental settings for a multi-modal interaction investigation.

With regard to measuring multi-modal interaction in an open space, we need an open space that allows participants to move around while talking to each other. Communicative behaviors of participants, both focused and unfocused interactions, need to be captured during the experimental session. A real-world interaction measurement, analysis, and design environment (IMADE), is a universal experimental environment in an open space Fig. 8.1 (Sumi et al. 2010). It allows for the design of experiments in an open space, about 5×5 m, in which participants can move around thereby changing the nature of the interaction. An optical motion capture system is used to measure the behavior of the users without physical constraints. Real-time sensing is necessary if some part of the environment needs to be reacted

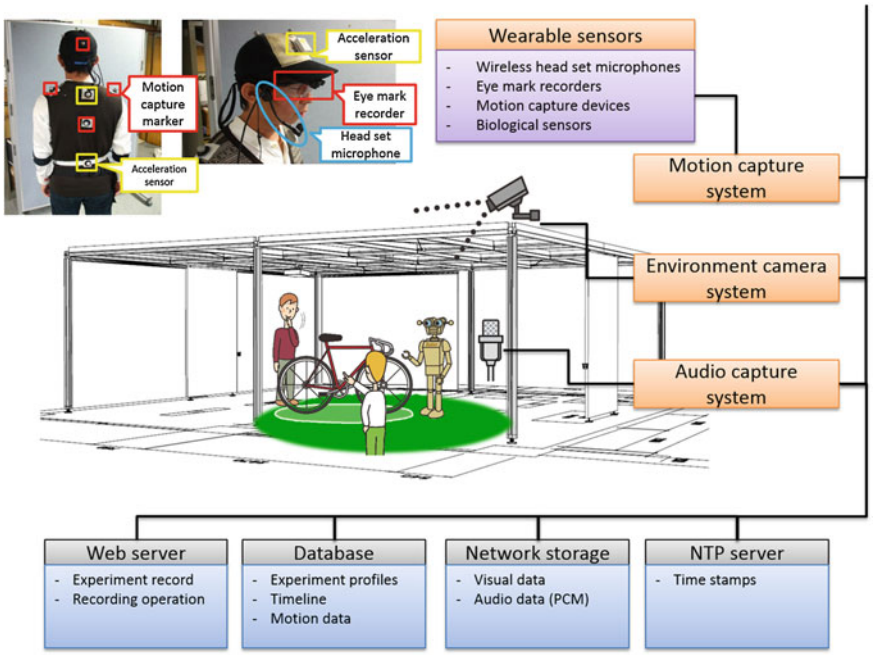


Fig. 8.1 IMADE: a real-world interaction measurement, analysis and design environment (Sumi et al. 2010). © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

on depending on the behavior of one or more participants in a given experiment. Eye mark recorders or throat microphones are used to capture the participants' behavior in a detailed and reliable fashion.

8.1.3 Preliminary Experiment

The aim of the preliminary experiment is to develop and validate the experimental settings and hypothesis under which the data are obtained to achieve the goal of the investigation. This includes confirming whether the target multi-modal interaction can be observed, the expected data can be obtained from the experiment, and the goal of the investigation can be achieved by analyzing the obtained data.

According to the nature of a preliminary experiment, a small number of participants take part in the experiment, with "typical" participants chosen from a pool of available participants. In some cases, the investigator allocates certain roles to the participants to cover appropriate conditions for interaction in the preliminary experiment. Either way, the investigator must carefully consider the choice of participants and their instructions because the participants are not supposed to consciously behave in a particular way to yield a particular result; instead they are expected to behave unconsciously according to the target features in order to investigate the interaction.

There are no clear guidelines as to the number of preliminary experiments. Experience has shown that three or more repetitions under the same experimental setting are appropriate. Before carrying out the preliminary experiment, a working hypothesis should be developed as the provisional goal of the experiment.

In the preliminary experiment, the investigator should strive to obtain not only the data that will be obtained in the full-scale experiment, but also as many kinds of data as possible. Since in many cases, audio and visual data are recorded using video cameras, multi-modal information should be recorded from various angles to different targets. Although the relationships among multi-modal data are important for analyzing multi-modal interaction, these are not that easy to obtain. Therefore, the data in the preliminary experiment should be more carefully analyzed than that in the full-scale experiment in order to develop the hypothesis that will be evaluated in the full-scale experiment.

8.1.4 Full-Scale Experiment

The aim of the full-scale experiment is to evaluate the hypothesis developed in the preliminary experiment under fixed experimental settings. Basically, the experimental settings and hypothesis should not be changed in the full-scale experiment. Thus, many kinds of data can be obtained unless the measurement obstructs the interaction or the process of gathering the main data. Since some measurement methods can obstruct the experiment, data that are not necessary to evaluate the hypothesis are not captured. Since video recording rarely obstructs the experiment, video cameras should be used to record the participants' behavior.

The data analysis in the full-scale experiment is relatively simple because the main purpose of the full-scale experiment is to evaluate the hypothesis by statistically analyzing the obtained data. As there are many statistical tests, the appropriate tests should be carefully selected and applied to the obtained data to evaluate whether there is a significant difference between the tested data. In other words, to evaluate which set of data is relatively better, a control experiment must also be carried out to allow for the comparison of the experimental results. If unexpected results that are important to the interpretation of the multi-modal interaction are observed, in many cases, the preliminary experiment must be repeated.

8.1.5 Data Analysis and Interpretation

Analysis of the data takes place in both the preliminary and full-scale experiments. In the former, the analysis is intuitive and analysts try to obtain helpful insight from the data to solve the research issues. The data in the full-scale experiment are analyzed according to the measured data with the analysts attempting to evaluate the hypothesis to interpret the obtained data.

There are two types of data: qualitative and quantitative data. Qualitative data are described using a nominal or ordinal scale, while quantitative data are described using an interval or ratio scale. The number of data described by the nominal scale is quantitative data. In all cases, data obtained from the experiments are first segmented and classified into units of analysis, where the unit depends on the modality. For example, recorded voice data include different modalities: verbal information and paralinguistic information (pitch, power, tempo, rhythm, and so on). In general, verbal information is segmented into words, phrases, and sentences, while paralinguistic information is segmented independently of the verbal segmentation, such as increases/decreases in pitch, speaking time, timing of breaths, and so on.

The segmented data are interpreted in the analysis. It should be noted that analysts interpret the data more or less subjectively. There is often a small gap between the purpose of the research and the objective of the experiment. Analysts can to some extent, subjectively explain this gap using the results of the analysis even when using statistical tests to analyze the data. This is so since researchers can report on the correctness of their results only relative to the diverse results of the experiment. Facts are provided as evidence in proving or disproving the hypothesis.

For the interpretation of data, qualitative analysis is indispensable to obtain a deep understanding of multi-modal interaction. Qualitative analysis depends on a human's subjective interpretation of the obtained data. Typical data for qualitative analysis of multi-modal interaction include the contents of the speech, meanings of gestures, and impressions during the interaction, amongst others. The data are extracted from videos, audio recordings, time series data of body motions, and interviews after the interaction. Researchers then analyze the data and annotate subjective opinions, evaluations and ratings, categories of the meaning of their interpretation, segmentation based on coherent semantic units, and descriptions of relationships among the segmented data. In this way, an annotation is a way of attaching metadata, such as a comment, explanation, presentational markup, or subjective interpretation, to the obtained data for qualitative analysis.

In the rest of this section, we discuss two of the advanced methods for enhancing multi-modal interaction analysis.

8.1.6 Collaborative Annotation

An "annotation" is not a special kind of data processing. There are many different types of annotations, such as intuitive data segmentations, applying nominal scales to the data, describing the interpretation of relationships among different data, and so on. Attaching annotations is generally performed as follows. First, the annotator intuitively segments the data and adds metadata to the segmented data. The annotator need not enumerate all metadata in advance. In addition, the annotator can add multiple metadata to the segmented data. Second, the annotator integrates the metadata, which have commonalities, by comparing the metadata and the segmented data. The commonalities form part of the rules for segmenting and classifying the data. Third,

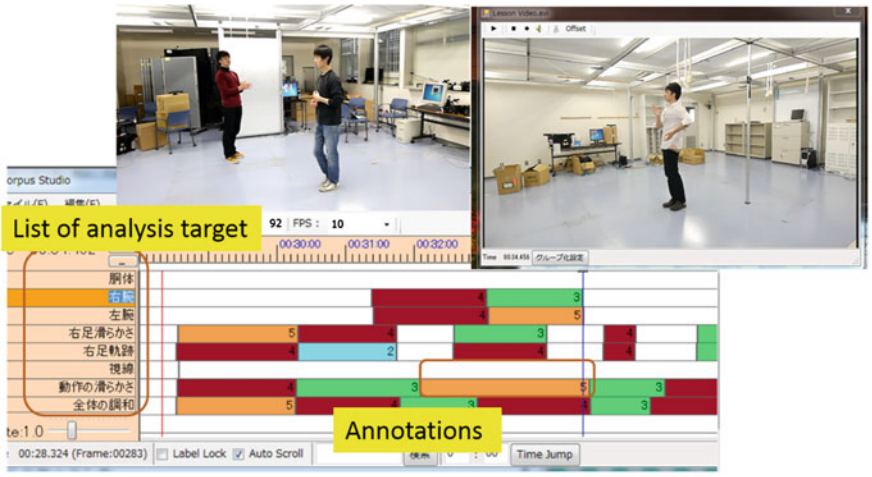


Fig. 8.2 iCorpusStudio (Sumi et al. 2010)

the annotator reduces the kinds of metadata according to the goal of the investigation. In some cases, some of the segmented data are eliminated from consideration. The criteria for reducing and eliminating data are also included in the rules for segmenting and classifying the data. Finally, the annotator creates annotation rules based on the commonalities and criteria for reduction and elimination. Thereafter, the annotation rules are applied to all the data, which have already been annotated, to confirm the validity of the rules. In general, the above procedure is carried out using the data obtained from the preliminary experiment. The procedure is a kind of qualitative analysis since the annotation rules form part of the interpretation of the interaction.

To annotate multi-modal data, data from various perspectives must be considered. Various annotation tools, such as Anvil (Kipp 2001) and Elan,[1] enable researchers to synchronously check multi-modal data, such as multiple videos, voices of interacting members, and measured time series data of body motions, and to annotate the various types of data. iCorpusStudio (Sumi et al. 2010), which is open source software, includes basic functions and tools for annotation, but also accepts plug-in extensions. Annotators can graphically add multiple lines of metadata in annotating multiple videos, voice data, and time series data, as shown in Fig. 8.2.

Consistency is vital to ensure good annotation quality. When an investigator annotates data, there may be consistency at least at the subjective level. However, annotations are not often objective because there are many implicit rules when annotating alone. To ensure objectivity, in many cases, metadata are added by independent persons who have been instructed in the correct way to annotate. To obtain high-quality annotation, we need to reduce both inter- and intra-annotator discrepancies.

[1] http://tla.mpi.nl/tools/tla-tools/elan/.

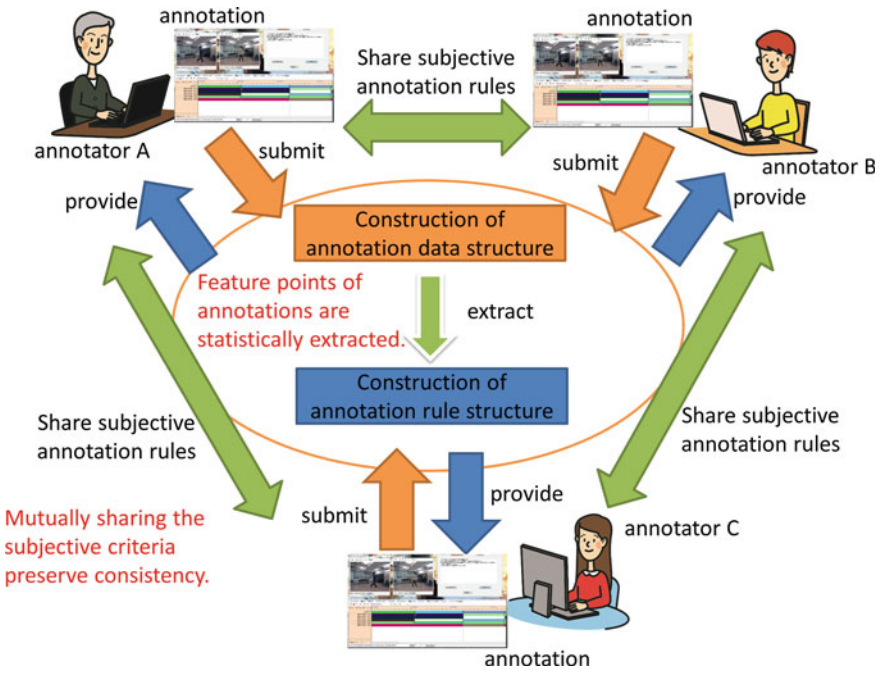


Fig. 8.3 Collaborative annotation system. © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

The conventional approach to constructing clear criteria for annotation is to provide annotation guidelines to which annotators can refer. These are descriptions of annotation rules in which concrete procedures are described, such as how to segment the data, how to symbolize the segmented data, how to categorize symbolic data, and so on. The annotation guidelines are written by a person who has detailed knowledge of the analyzed data and interaction behavior.

Unfortunately, the above approach does not help much. First, it is difficult to describe criteria for high-level interpretation. Second, criteria for annotation differ depending on what kind of data and situation are being analyzed and hence, reusability of annotation guidelines is rather low. Third, annotation guidelines are often incomplete and hence need to be modified repeatedly during the annotation task, which may introduce additional overhead in maintaining consistency of the annotation guidelines.

Collaborative annotation helps multiple annotators share case-based annotation rules including concrete examples and extracted features from the data. Figure 8.3 outlines a collaborative annotation system. Each annotator submits his/her subjective annotations including predefined labels for the annotations and annotated data. The system identifies similarities among the annotated data with the same labels and statistically extracts feature points of the extracted data. Thereafter, the system defines the annotation rule structure, which incorporates the extracted features and summarized

annotated data. When an annotator annotates new data, the system presents certain candidate annotation rules based on the features of the data being annotated. By referring to the extracted features and example annotated data provided, the annotator can annotate the new data. Since the annotation rules are shared among annotators who are annotating the same type of data, they can verify their subjective annotation rules and at the same time reconstruct these based on other annotators' rules. This allows the annotators to construct consistent criteria for the annotation with implicit features and reduce both the inter- and intra-annotator discrepancies.

8.1.7 Physiological Signal Analysis

Humans can infer the psychological state (more or less accurately) from behavioral cues of other people which is a valid path for conversational agents (see Chap. 9 of this book) but,—given the current state of technology—artificial conversational agents may not be as skillful as humans in this task. For this reason, these agents will appreciate the help that physiological signal analysis can provide.

Physiological indices are biological reactions caused by the autonomic nervous system. Many physiological indices (e.g.,Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), Respiration Rate (RR), Skin Temperature (ST), etc.) were shown to correlate with different aspects of human internal states (e.g., Shi et al. 2007; Mandryk and Inkpen 2004; Lin et al. 2005; Lang 1995; Mower et al. 2007; Bradley and Lang 2000). In most cases the analysis was done using extreme controlled conditions in which differences in the internal state is intense enough to be captured by simple statistics of the physiological signal under processing.

For example, Lin et al. (2005) used a 3D computer game environment, similar to the one used by Mandryk and Inkpen (2004), while Shi et al. (2007) used a complex multimodal user interface with 12 different tasks of varying complexity. It is not clear that the correlations found in these extreme conditions can be reliably found in natural conversation situations that constitute the focus of this book.

Nevertheless, Mohammad and Nishida (2010c) showed that physiological indices can be used reliably in normal interaction conditions to infer the psychological state of people interacting with robots. This will be the subject of the case study we report in Sect. 8.2.

Physiological indices have several advantages in the context of conversational informatics. Firstly, reactions to psychological activities can be digitized as physical quantities such as voltages and frequencies. In addition, we can statistically and mathematically analyze the data to a certain degree. Secondly, we can analyze the variability of physiological indices over time because it is recorded as time-series data in real time. On the other hand, questionnaires are usually carried out as surveys on personal psychological activities after the activities have been completed. Thirdly, there are cases where changes of mental states can be found even where there are no visually noticeable changes in continuance or conduct. It is possible to pick up, for example, stimuli that are not perceived consciously. Finally, reactions are

involuntary, and are not susceptible to the effect of logical thinking. Since reflective reactions are produced in response to certain stimuli, it is often possible to make repetitive measurements.

Despite these advantages of physiological indices for conversational informatics, they have some limitations. First, there are cases where measurement results are influenced by individual differences and external stimuli such as illumination and disturbing sounds. The very fact of being subjected to measurement often causes stress, resulting in physiological reactions being affected. Second, noise is usually introduced by the measurement device. Noise generated by alternating current magnetic fields is liable to cause problems. When electrodes or lead wires for measurement vibrate, artifact-caused effects tend to be introduced. Various restrictions are often imposed to reduce noise. Third, since relevant data are time series data and lags are present in reactions, it is difficult to process the data in a uniform manner. There are individual differences in the magnitude of reactions and delay times for reactions. Fourth, collecting physiological data from people requires the attachment of sensors to the body which may affect the *naturalness* of their behavior. Reducing the invasiveness of measurement is one of the major challenges in utilizing physiological signal analysis widely in conversational informatics. Finally, a major problem with physiological sensing of internal state is that in most cases different individuals have different responses to the same stimuli and these individual differences can be much higher than the differences that depend on the stimuli itself. For example, Bradley and Lang (2000) found that only 74 % of the tested subject have statistically significant correlation between GSR level and arousal despite the wide usage of this signal to measure arousal (Lang 1995).

Some problems that are not inherent in the physiological indices themselves may be introduced through the method of analysis. For example, one problem of most available methods for inferring the psychological state of people based on physiological signal analysis is that they put very little emphasize on the effect of the interaction context on the measured physiological indices because most of them are used to measure the response to an inanimate object (e.g., a computer interface). Human-human interaction research shows that normal interactions between humans go through different phases including opening and closing phases. It is expected that the physiological response of every partner to the behavior of other partners will depend on the interaction phase in which this behavior takes place (Argyle 2001).

In conversational informatics we are interested in evaluating the subject's response to the behavior of its partner during a natural close encounter situation. There can be many varieties of such situations and in this book we use the explanation scenario in which the subject is explaining the assembly/disassembly of some machine or device to a listener using verbal and nonverbal modalities. This situation is only used as a convenient tangible case but the discussion can be applied to many other interaction situations as well. The listener can either be another human subject or a robot.

This explanation scenario was selected because of its importance for conversational informatics as well as general human-agent interaction applications (i.e., learning by demonstration, knowledge media robots, companion robots etc.). To simplify the discussion, we will focus mainly on diadic interactions.

There are many types of physiological indices. Although each index is measured in a different way, most of physiological indices are measured by contact sensors on human body. Depending on the experimental settings, some indices cannot be measured. For example, respiration is not appropriate to find mental stress in conversation because the respiration is changed by many factors in the situations, such as breathing for turn taking, loudness of speech and laughing. On the other hand, the respiration can catch the mental stress in attentive listening situation like lecture presentation.

The most important physiological signals that can be used for conversational informatics are as follows.

8.1.7.1 Electromyograms (EMG)

The electrical action potential generated when muscles contract (10 microV–several milliV) can be measured. A pair of electrodes are attached on the measured muscle and tape is used to fix the electrodes and lead wires in place. The electromyograms are relatively strong signals.

These are utilized when attempting to identify when and how muscles are used. In multi-modal interaction analysis, electromyograms are used to check muscle contractions that are not accompanied by preparatory reactions or movements; for example, monitoring when preparation of a gesture starts and whether the muscles became tense.

8.1.7.2 Electrocardiograms (ECG)

Electrocardiograms are electromyograms of the heart. At least two electrodes are attached to trunk areas. It is possible to independently measure fluctuations in all of the atria and ventricles. In our studies, it is general practice to measure pulsations as a single wave.

It is general practice to check for tension and stress based on fluctuations in pulse. Human psychological states are reflected not only in whether the pulse itself is rapid or slow but also in how the pulsations fluctuate. Since there are many methods for analyzing electrocardiograms, an appropriate method must be chosen for analysis of multi-modal interaction.

8.1.7.3 Respiration

There are two major methods to measure respiration. One is using a band with variable resistance that changes based on the expansion and contraction. This band is placed around the thorax of the test subject, and then the expansion and contraction of the thorax are recorded as resistance values. Another type of respiration sensors uses temperature. This type of sensor is attached immediately below the nostrils and

respiration is recorded as temperature changes when exhaled air and inhaled air pass through the nostrils.

The amount of mental activity is often determined by observing changes in respiration. Respiration becomes rapid when there is mental activity, and slows down during relaxation. Respiration is also used to check the so-called "act of bringing respiration into harmony," and the timing at which movement starts.

Respiration is believed to be too slow for reflecting real time change in the internal state of humans but Mohammad et al. (2008) showed that—using appropriate processing—it can be a reliable physiological measure of naturalness.

8.1.7.4 Electro-Dermal Activity (EDA)

EDA contains Skin Potential Activity and Skin Conductance Activity (Skin Conductance Response, SCR and Skin Conductance Level, SCL). There are specific sweat glands that are used to measure skin conductance called the eccrine sweat glands. Located in the palms of the hands and soles of the feet, these sweat glands respond to psychological stimulation rather than simply to temperature changes in the body. As sweat rises in the particular gland, the resistance of that gland decreases even though the sweat may not reach the surface of the skin. Skin conductance measures the electrodermal activity of the autonomic nervous systems.

Skin conductance is correlated with affective arousal (e.g., Lang 1995), cognitive load (e.g., Shi et al. 2007), frustration (e.g., Lin et al. 2005), and engagement (e.g., Mower et al. 2007). Bradley and Lang (2000) showed that 74 % of the subjects exhibited this correlation. This means that emotional perspiration reflects stress and tension. It is also possible to measure anxiety, excitement, surprise, and pain; however, it is not possible to identify what type of mental activity is involved. Since reactivity is high, this should be used in parallel with other physiological indices and behavioral analysis. EDA is a physiological index that is relatively easy to measure.

8.1.7.5 Blood-Volume Pulse (BVP)

Blood Volume Pulse (BVP) measures the cardiac activity of the autonomic nervous systems using a sensor attached to one of the body extremities. The heart rate deducible from BVP has been used to differentiate between positive and negative emotions (e.g., Papillo and Shapiro 1990). Heart rate variability can also be deduced from BVP and is used extensively in human factors literature as an indication of mental effort and stress in high stress environments (Rowe et al. 1998).

8.1.7.6 Electroencephalography (EEG)

Electrodes are fixed to the scalp, and electrical activity that occurs in conjunction with cerebral activity is recorded. Electrical activity that is recorded turns out to be the sum total of significantly wide-ranging activities of the brain. Brain wave recordings are divided into the following two types: EEG (These record the so-called background brain waves. Alpha waves etc. fall in this category.) and ERP (This acronym stands for Event Related Potential. The ERP is observed as a reaction to certain phenomena.). ERP is observed after noise is removed by the averaging method from brain waves that were measured repeatedly during an experimental design (block design) in which the same phenomena occur repeatedly. The spatial resolution is not so good, but the temporal resolution is considerably good.

It is possible to use brain waves to observe many phenomena. However, the brain is almost always active during multi-modal interaction. Since a block design is indispensable, it is impossible to "perform measurements as a temporary measure." Moreover, experimental settings must be determined, including the measuring environment and task design, in which the relation between an interaction event and brain wave reaction can be identified on a hundred millisecond time scale. The design of such an experiment is difficult in multi-modal interaction analysis.

8.2 Natural Interaction Measurement

Protagoras claimed 25 centuries ago that *Man is the measure of all things*. We do not concern ourselves much with the philosophical connotations of this statement in this book but it reveals an important point for conversational agent design: conversation agents are designed to interact with people and as such their success or failure rests upon the *subjective* evaluation of their human partners. This means that the subjective evaluation of people is essential for judging the quality of interaction with conversational agents. Nevertheless, we will occupy ourselves in this case study with devising an *objective* physiological signal analysis scheme for measuring the naturalness of interactions (which we define in terms of minimizing stress levels in human partners). Why would we do that?

An objective measure of interaction naturalness as perceived by people engaged in it can be of value for several reasons. Firstly, it can be used (if online and fast enough) by the conversational agent to better accommodate its behavior to the psychological needs of its human partners. This can move us a step further toward an empathetic technology. We do not expect people hooked up with wires and invasive sensors to feel any natural but the current state of technology allows us to measure many interesting physiological indices (e.g., blood pulse, respiration rate, skin conductance) easily and with minimal contact with the subject. Future technology is expected to be even less opaque and ubiquitous computing is on its way to allow conversational agents access to all kinds of information about the physiological state of people.

Secondly, even though subjective evaluations through questionnaires can cover to some extent the response of the subjects to technology, it is very hard to compare results of such subjective evaluations coming from different research groups due to the difficulties in controlling environmental conditions across experiments. Finally, the results of questionnaires are known to be cognition mediated and disagreement between reported and measured behavior is evident from many psychological studies (Mohammad et al. 2008). In a subject with strong novelty effects, and sometimes high cognitive load like interacting with robots and other conversational agents such limitations of subjective evaluations based on questionnaires become even stronger.

From this brief discussion, we can not only judge the usefulness of having the ability to analyze human physiological data to get some clues about their internal psychological states but we can also elicit some of the desirable features of any system that tries to achieve this goal in the context of conversational informatics. Other than the obvious need for accuracy, it is desirable to have a real-time system for online applications and most importantly, the technology should be least invasive (and if possible technologically transparent) in order to preserve the psychological state of the subject.

There are many possible dimensions of comparison between interactive robots, their appearance, their behavior, the partner's response to them, etc. One obvious measure of interactive robots' behavior is how human-like they are in terms of behavior and appearance, but it is not always the case that human-like behavior and appearance are considered as more *natural* by the partner (Qazi et al. 2006).

As the goal of our metric is to compare interactive robots *as perceived by their human partners*, we focus on the partner's response to the robot.

Definition 8.1 *Natural behavior* of an agent is defined as the behavior that minimizes the elevation in stress level, cognition load, frustration, and anxiety of the partner during the interaction while maximizing his/her engagement.

A robot or agent that has more *natural* behavior according to this definition is more desirable than a robot that has less *natural* behavior in most normal situations. Using this definition, it is possible to measure naturalness of agent's or robot's behavior by measuring the physiological response of its human partner to this behavior.

We will report a case study for measuring interaction naturalness using physiological signal analysis (Mohammad and Nishida 2010c). The dataset used consisted of human-human interactions sessions using *good and normal* nonverbal behavior, human-human interaction sessions using *bad and abnormal* nonverbal behavior as well as human-robot interaction sessions using a simple interaction protocol. Availability of positive and negative examples of behavior is a unique property of this interaction dataset and allows robot designers to utilize learning algorithms that require negative examples for its training and also provides a ground truth to test different interaction evaluation metrics.

The participants (44 subjects) where randomly assigned either the listener or the instructor role. The instructor interacts in three sessions with two different human listeners and one humanoid robot (Robovie II). The instructor always explains the

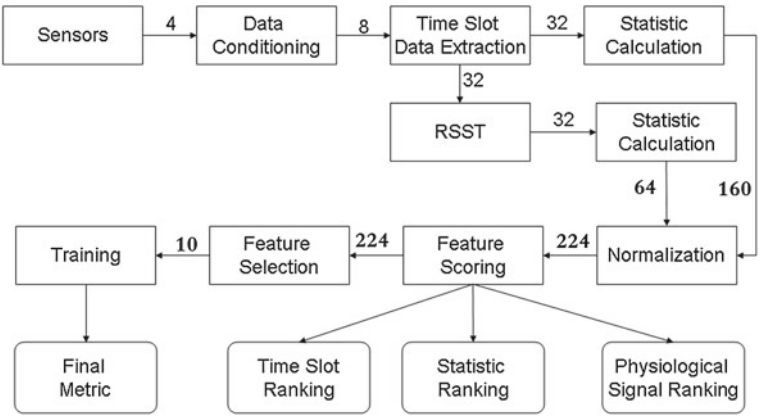


Fig. 8.4 Data processing steps applied by Mohammad and Nishida (2009d) to the physiological indices. Each link shows the dimensionality of the data transferred between processing blocks. © 2009, Springer. Reproduced with permission

same assembly/disassembly task after being familiarized with it before the sessions. To reduce the effect of novelty, the instructor sees the robot and is familiarized with it before the interaction (Kidd and Breazeal 2005).

The listener listened to two different instructors explaining two different assembly/disassembly tasks. In one of these two sessions (s)he plays the *Good* listener role in which (s)he tries to listen as carefully as possible and use normal nonverbal interaction behavior. In the other session (s)he plays the *Bad* listener role in which (s)he tries to appear distracted, and use abnormal nonverbal interaction protocol. The listener is free to decide what is *normal* and what is *abnormal*.

The goals of this work was three fold as summarized by Mohammad and Nishida (2010c):

> Firstly, to assess the usability of various physiological indices and features extracted from these signals in evaluating *naturalness* of behavior as defined earlier (Definition 8.1). Secondly, to support or reject the hypothesis that different time slots of the interaction may not be of the same importance in this evaluation. Thirdly, to use the most important signals and time slots to drive an objective metric of *naturalness*.

Figure 8.4 shows the processing steps used in this experiment. Physiological signals collected from moving people are subject to several artifacts from motion and other distortion and noise sources. This makes it nearly indispensable to remove outliers from the data before processing it. In this study we utilized Rosner's many-outliers test (Rosner 1975) for this purpose and smoothed the resulting signal using a Savitzky-Golay filter of second degree. In their raw form physiological signals may not be appropriate for detection of the internal cognitive/psychological/emotional state of human subjects. We used the following set of features that were calculated from the smoothed signals: Heart Rate (HR), Heart Rate Variability (HRV) as well as raw pulse data (P) were calculated from BVP data. Skin Conductance Level (SCL)

and Galvanic Skin Response (GSR) were calculated from skin-conductance sensor data. Respiration Rate (RR), Respiration Rate Variability (RRV), and raw respiration pattern (R) were calculated from the respiration sensor. This leads to a total of eight physiological signals to be processed.

One of the goals we just quoted from the study was to find effective interaction time slots during which it is easier to distinguish natural and unnatural behavior. To achieve this goal, we extracted 2 min in the beginning and end of each interaction and 4 min from its middle as representatives of the three main stages of the interaction: opening, main part, and closing. The rationale for this 2 min boundary is discussed by Mohammad and Nishida (2010c).

We then extracted the statistics usually used in estimating the psychological state of people from their physiological signals (see Shi et al. 2007; Lang 1995; Liao et al. 2005): mean (MEA), median (MED), standard deviation (STD), minimum (MIN) and maximum (MAX) leading to a 160 features for every session.

Our hypothesis was that changes in physiological signals rather than their specific values are more important in estimating the psychological state. Given that all of the aforementioned features depended directly on the specific values of physiological signals rather than the changes in these values, we employed a change point discovery algorithm called *robust singular spectrum transform* (RSS) that will be discussed later (Sect. 9.5.1) to find the changes in the signals and derived two features from the output of this algorithm: the number of local maxima per second (RSST LMD) and maximum number of local maxima per minute (RSST LMR) were calculated. This led to 64 more features for every session totaling 224 features per session. For details about the RSST change point discovery algorithm, please refer to Sect. 9.5.1.

Analysis of the data revealed two important points: First, the starting and end of the interaction are *less* important for classifying different conditions. Second, not all features and signals are equally important.

The most accurate classifier that was tested on this data was a tree-classifier. The most useful statistics in the first levels of the tree were shown to be RSST based statistics which supported the hypothesis that changes rather than instantaneous values of physiological signals are the most important discriminating factors in judging interaction naturalness.

It is interesting to notice that the behavior of the simple humanoid used in this experiment (randomly change gazing direction) was perceived even worse that unnatural human behavior by the subjects.

8.3 Measuring Social Atmosphere

How deeply an individual is engaged in a conversation can be determined by observing to what extent the individual pays attention to the conversation with other participants and voluntarily participates in the conversation. Involvement, enjoyment, and excitement lead to vigorous body movements. In contrast, people are slow to

respond to stimuli or even remain still when they are not in such a deeply involved mental state.

Ohmoto et al. (2010) assumed that people can also explicitly or implicitly detect the degree of involvement, enjoyment, or excitement of their communication partners based on the vigorous movements reflecting their attitudes, and introduced the *I-measure* as a measure of the involvement, enjoyment, and excitement people experience during conversation.

We focus on the following three questions: (1) whether a person's I-measure can be detected by using visual information, (2) whether the atmosphere of the I-measure affects members who are not directly involved in the setting in which the I-measure was obtained, and (3) whether the affected I-measure can be detected by using visual information. We use both visual information and physiological indices to elicit a reliable interpretation.

8.3.1 Methods for Obtaining the I-Measure

To acquire an I-measure for an individual, we asked two coders to manually elicit interaction scenes from the video data, and annotate whether the I-measure of a target person and a social atmosphere is high by observing the speed and distance of the face, head, hand, and shoulder movements. As the speed and distance of these movements differ from individual to individual, the coders needed to observe how fast and how long individuals move their bodies when they are considered to be heavily involved in the communication. Having watched the video three times, the coders annotated the I-measure states for each person. Although features of speech sound, such as pitch and power, may serve as useful clues for detecting I-measures, it is difficult to separate sound sources when several users are talking nearby.

We assumed that the social atmosphere related to the I-measure was high if more than two communicating participants had a high I-measure. The level of the atmosphere was calculated as the average of the I-measures of the communicating participants. For example, if three out of four participants in the communication had a high I-measure, the level of I-measure for the social atmosphere was calculated as the average of the three participants' I-measures. A person familiar with video annotation defined the social atmosphere associated with a particular I-measure.

We also used physiological indices to estimate the person's mental state, as it is known that biological reactions, such as brain waves, potential differences in cardiography, and variations in blood pressure, pulse waves, respiration, body temperature, muscle potential, and skin conductance, reflect the mental state of a person. We used SCR and respiration in this experiment. SCR measures the skin conductance caused by emotional sweating, which can occur as a result of excitement and mental stress and as concentration increases. At the same time, respiration becomes more rapid as a result of the mental state. We can detect laughing by rapid respiration. We did not use brain waves, since the actual measurement of brain waves often prevents participants from using natural communication as a head piece must be affixed to the

participant's head to measure the brain waves. Nor did we use pulse waves, as these often contain a large amount of noise.

Following Lin et al. (2005), we report an increase in SCR only if SCR increases more than 5 % per second and rapid respiration if respiration peaks are observed more than four times every 3 seconds. The latter criterion was introduced to detect changes in respiration other than those caused by laughing, and to avoid the detection of changes associated with normal communication.

8.3.2 Experiment to Record I-Measure Responses

We conducted an experiment to record the responses of participants in a high I-measure situation to videos and to measure their physiological indices.

8.3.2.1 Experimental Task

The participants were asked to answer quiz questions while being allowed to communicate freely. One quiz session consisted of ten questions. Each participant attended three sessions, i.e., s/he answered 30 questions. Of the ten questions, three questions required inspiration to solve, three questions involved logical structure, and four questions relied on general knowledge. The order of the questions was set so that the same kind of question was not presented sequentially. The participants were eligible for a prize depending on the number of correct answers.

8.3.2.2 Participants

Four participants composed a group, referred to as a "quartet" in what follows. Each quartet comprised undergraduate students, who were 21 or 22 years old and acquainted with one another. Three male quartets participated in the experiment.

Participants answered the quiz questions for about 45 min. The duration of the three sessions for Group A was 50 min, that for Group B was 40 min, and that for Group C was 50 min. The participants did not perform any other activities besides conversation when answering the questions and no participant answered the questions alone without interaction with others in the group. Therefore, the mental states as detected by the physiological indices, such as excitement, stress and cognitive load, during interaction in the task can be interpreted as the participants' I-measures during the interaction.

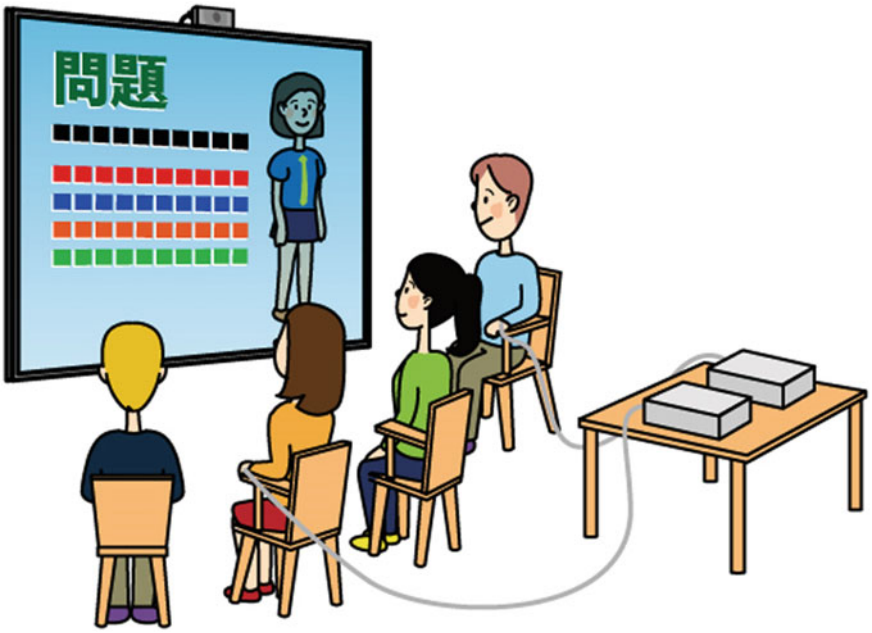


Fig. 8.5 Experimental settings for capturing human behavior and physiological indices. Adapted from Ohmoto et al. (2010). © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

8.3.2.3 Experimental Setting

An embodied conversational agent (ECA), referred to as the quiz agent, introduced the series of quiz questions, since we had identified a method to detect the I-measure in multi-user interaction which involved an ECA. A 100-inch screen displayed the quiz agent. Using a notebook computer, the investigator, who was located out of the participants' view, directed the quiz agent to start a quiz or accept an answer to a question. During the quiz, participants remained seated in a half circle in front of the screen to avoid introducing disturbances into the data. Participants were able to watch the screen and communicate with one another naturally (Fig. 8.5).

A video camera which was placed on top of the screen recorded the participants' behavior. A Polymate device, placed on a table out of sight behind the participants, was used to measure the physiological indices of the participants. We recorded two of the participants seated on the chairs. In this experimental setting, participants seated next to each other tended to talk to one another. The participants whose data were measured were separated to prevent them from communicating in only a single conversational field. Electrodes for measuring SCR were placed on the forefinger and ring finger of the left hand. The left arm was place on an armrest to support the left hand. Respiration was measured by a band placed around the chest. Participants did not experience any discomfort from this band.

8.3.3 Analyses of the Effects of an Atmosphere

The collected data consisted of visual information and physiological indices. First, we asked two coders to independently segment the video data and annotate whether a target person and the social atmosphere conformed to the state of the I-measure, based on the visual information. Then, the coders performed the same procedure on the physiological indices data. The annotation tool iCorpusFStudio (Sumi et al. 2010) was used for segmentation and annotation. Finally, we conducted a quantitative analysis to compare the annotations from the two coders and investigate the possibility of using only visual information to detect states of the I-measure and to identify the effects of the social atmosphere on the I-measure.

8.3.3.1 Adequacy of Annotations of the I-Measure Using Visual Information

We examined the adequacy of the annotations of I-measure by visual information. Coders annotated the level of the I-measure state on a scale from zero to one or one to five. They segmented the videos into I-measure units, in which the level of the I-measure state is the same. Consecutive I-measure units do not necessarily correspond with consecutive levels of the I-measure. For example, coders could annotate 1/4/1/3 in four consecutive I-measure units. Two coders, the researcher and a person with vast experience in video annotation (whom we refer to as the "reference coder"), segmented the videos and annotated the level of the I-measure state.

We compared the positions of annotations of the two coders to confirm their reliability. When coders annotates continuous data, the segmentations (the start and end positions) are often slightly different between the coders'. We thus regarded as the annotated positions are the same when more than half between the two coders' annotations were overlapped. In 82 % of the 214 annotations, the positions of the two coders' annotations overlapped, with more than half located between the two coders' annotations. This confirms the reliability of the positions of the annotations.

We compared the levels of the annotations on a scale of one to five to further validate the reliability, and found that the levels of 70 % of the two coders' annotations matched completely with the chance level set to 20 %. In addition, we found that a further 25 % of the remaining 30 % of annotations differed by one level between the two coders. This confirms the reliability of the levels of the annotations. Since the levels of the annotations are a ordinal scale, the baseline of the scale and the intervals between the scales are often different between the coders. To use the annotations for the analyses, the baseline and the intervals should be standardized between the coders to some extent. In this study, we carefully explained the annotation guideline to the coders for the standardization.

Table 8.1 Concordance rates between annotations using visual information and those using physiological indices

	Number of I-measure using physiological indices	Number of non-I-measure using physiological indices
(a) Group A		
Number of I-measure using visual information	96 (83 %)	20 (17 %)
Number of non-I-measure using visual information	42 (35 %)	78 (65 %)
(b) Group B		
Number of I-measure using visual information	38 (76 %)	12 (24 %)
Number of non-I-measure using visual information	30 (57 %)	23 (43 %)
(c) Group C		
Number of I-measure using visual information	44 (86 %)	7 (14 %)
Number of non-I-measure using visual information	29 (52 %)	27 (48 %)

8.3.3.2 Adequacy of the Annotations of I-Measure by Physiological Indices

We investigated the adequacy of the annotations of I-measure by physiological indices. We examined whether a participant was in a state of I-measure at each of the I-measure units that were segmented by coders. The "increase of SCR" and the "rapid respiration," were used to identify I-measure, while the levels of the I-measure state were not. We found that most of the annotations were matched between the coders, as the criteria were clear.

8.3.3.3 Analysis of Accuracy of I-Degree Detection by Using Visual Information

We investigated whether people could detect the I-measure state based on visual information. For this investigation, we compared the annotations with an I-measure level on a scale from zero to one based on visual observation, with those obtained using the physiological indices. Table 8.1 gives the results.

The average concordance rates between annotations by visual information and those by physiological indices were 74 % in Group A, 60 % in Group B, and 67 % in Group C. It allowed us to conclude that some people are better able to detect the I-measure states using visual information than others.

Conversely, the concordance rates were not that high, particularly for non I-measure annotations. In most cases, annotations using visual information indicated

non I-measure states, while those based on physiological indices indicated I-measure states. We conclude that low-level I-measure states are difficult to detect using visual information, but can be detected using physiological indices.

In addition, we analyzed videos of the experiment to investigate what caused low-level I-measures. Low-level I-measures were observed in several situations in which other participants had high I-measures. This means that a participant's I-measure can be used as the induced I-measure of the social atmosphere in those situations in which the I-measure of an individual can be detected using visual information. However, it is necessary to detect low-level I-measures using visual information with a low threshold.

As mentioned above, to interpret the target phenomenon by using different criteria and to compare them, we can reveal the advantages and disadvantages of each criterion and the features included in the target phenomenon. By using annotation data, we can apply the analysis to the ambiguous and intuitive criteria.

8.3.3.4 Analysis of the Effects of an I-Measure Atmosphere

When a multi-user interaction has the social atmosphere of the corresponding I-measure, it intuitively affects members who have not yet reached the respective I-measure state. We thus investigated whether an I-measure atmosphere could affect those members who are not directly involved in the I-measure state and whether participants could detect the I-measure state using visual information. For the investigation, using visual information, we selected scenes in which a participant was in the social I-measure atmosphere, but not in an I-measure state individually. We conclude that participants are affected by the social I-measure atmosphere, which is difficult to detect based on visual information, when annotations based on physiological indices indicate that the participant is in an I-measure state in the selected scenes. We conclude that the social I-measure atmosphere can affect participants who are not directly involved in the I-measure state. In addition, the affected I-measure state can be detected using visual information with a low threshold.

This analysis suggests that the visual information and the physiological indices captured the different features of an I-measure state from different perspectives. The visual information captured extrinsic I-measure and the physiological indices captured intrinsic I-measure. The situation in which the extrinsic I-measure affects the intrinsic I-measure of others is the I-measure atmosphere. We can show some findings, such as the features of I-measure, the methods to detect I-measure and what is the I-measure atmosphere through the analyses using annotation data based on ambiguous criteria.

8.3.4 Discussion

A method for detecting I-measures could be implemented as follows. First, the region of a user in the camera images is detected by image processing, such as background

subtraction. Second, moving distances and the speed of the user's body motions in the region are detected by an image processing method, such as optical flow. Third, the user is considered to be in an I-measure state if the moving distances and speed are above a certain threshold. Fourth, the level of the social atmosphere is determined when it is found that one or more members of the group are not in an I-measure state. Fifth, the non-I-measure members are judged again with a lower threshold to detect an affected I-measure state when the level of the social atmosphere is above a certain threshold. We expect that this method would be able to detect about 70 % of I-measure states.

For robust and accurate detection of I-measure states it is necessary to consider voice information and facial expressions as well. In order to use these, they must be measured in a multi-user interaction. However, currently there is no system that can measure these in a natural interaction where the interaction partners do not have to wear special devices. This remains as future work for researchers investigating a suitable measuring system.

8.4 Extracting Evaluation Criteria for Ballroom Dancing

Most forms of art including ballroom dancing are evaluated using subjective criteria defined by experts. When beginners learn an art, they have to understand what is important in the art; however, the criteria can be ambiguous and hard to understand, especially given the expression thereof in language only.

In ballroom dancing, for example, the instructor often points out mistakes and important features using verbal expressions and pointing gestures. (In this section, we refer to the learner's motion including the mistake as "practice motion: before.") In addition, the instructor may demonstrate better execution of the features that have been pointed out (hereafter referred to as "teaching motion"). After being corrected, the learner practices the dance motion considering the criteria. After much practicing, the instructor decides that the learner can perform the dance motion at an acceptable level (hereafter referred to as "practice motion: after"). Thus, we expect that dance motion criteria can be extracted from the data by comparing the "practice motion: before," "practice motion: after," and "teaching motion" data. Before the comparison, we can identify the important body parts related to the criteria based on verbal expressions and pointing gestures when teaching.

The methodology introduced in this chapter could possibly be applied to extract and express the evaluation criteria for ballroom dancing based on measured dance motions. We built a ballroom dance evaluation support (BDES) system to extract evaluation criteria from teaching interaction in ballroom dancing from the perspective of whether the learner rethinks important features of the dance. The BDES system extracts the important features of the dance from time series data of dance motions when teaching and learning ballroom dancing.

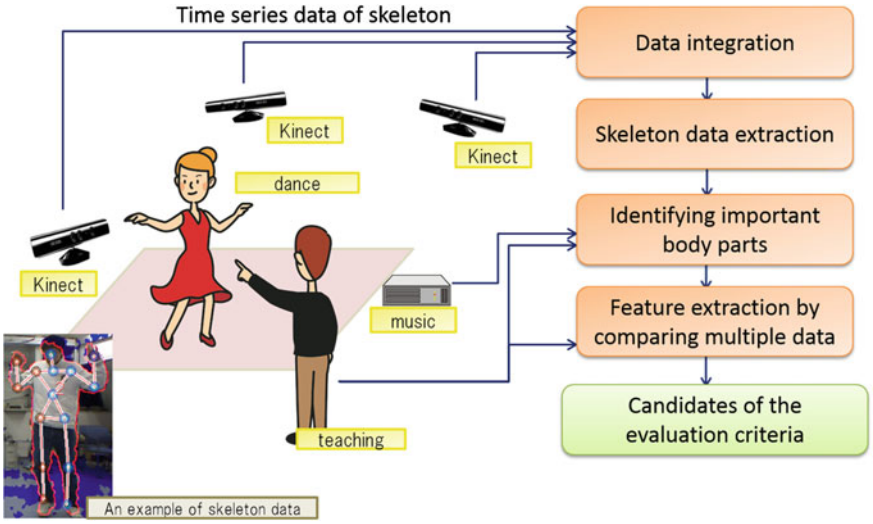


Fig. 8.6 Overview of the data flow in the proposed system. © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

8.4.1 Ballroom Dance Evaluation Support System

Execution of the system proceeds as follows: First, dance motion data while teaching are captured using the 3DCCbyMK system (introduced in Chap. 6). The captured data consist of 3D positions, speed vectors, and acceleration data for each human body joint (ten joints in total). Second, the system operator selects the teaching scenes and manually identifies important body parts in each scene based on verbal expressions and pointing gestures. Third, from the teaching scenes, the system extracts the dance motions categorized as "practice motion: before," "practice motion: after," and "teaching motion." Fourth, the system calculates body motion parameters, such as swing speed of arms, angles of arms and neck, acceleration of the center of the body, transitions of each joint, and so on, and the similarities among "practice motion: before," "practice motion: after," and "teaching motion." Finally, the system extracts several parameters as evaluation criteria in the teaching scene based on these similarities. Figure 8.6 depicts an outline of the data flow.

8.4.2 Evaluation Experiment

The purpose of this experiment was to examine whether the system can extract evaluation criteria when teaching ballroom dancing to a beginner. We captured actual data

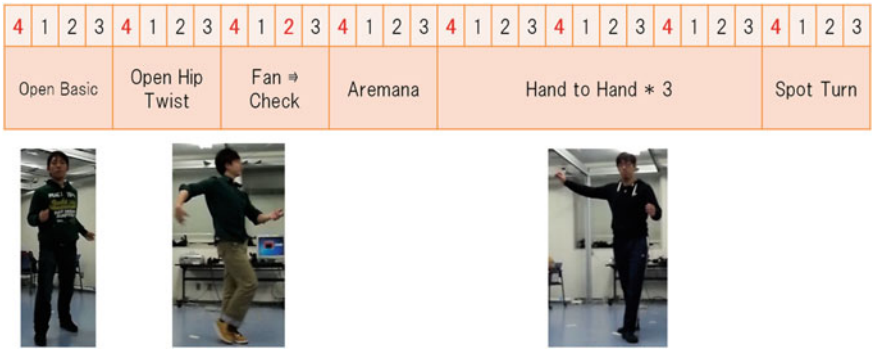


Fig. 8.7 The structure of a dance step

when beginners were being taught by a ballroom dancing instructor and evaluated three points in a ballroom dance step.

8.4.2.1 Task

In this experiment, participants learned one of the basic steps of the "rumba" dance. They practiced the step repeatedly until they could perform the dance motion at an acceptable level.

8.4.2.2 Evaluated Criteria

We extracted and evaluated criteria at three points in the dance step: point (1) posture in "HandToHand" count 2; point (2) swing speed of left arm and finish position in "HandToHand" count 2; and point (3) time series of postures in "OpenBasic." (See Fig. 8.7 for the structure of the dance step.) The three points identified above are important for performing the dance well.

8.4.2.3 Participants

The participants in this experiment comprised 12 Japanese college students (ten males and two females), who were beginner ballroom dancers. Since some of the dancers did not need to practice certain parts of the dance repeatedly, we obtained the following teaching data: data for eight participants at point 1, data for eight participants at point 2, and data for six participants at point 3.

8.4.3 Results and Discussions

In this analysis, we conducted leave-one-out cross-validation to evaluate whether the extracted criteria were reasonable by discriminant analysis.

8.4.3.1 Procedure

Three annotators annotated the beginner's ballroom dance, which was segmented into 15 parts. Each annotator annotated how good each of the body parts (right hand, left hand, right leg, left leg, and torso) and the complete motion were in each segment of the data by comparing the beginner's video with an expert video of the same dance.

To use the system, we measured time series data of the body motions using the 3DCCbyMK system (see Chap. 6) at the same time as recording the video. The trajectory, speed, acceleration, and angle of each body part were calculated based on the measured data. Thereafter, we calculated the similarities between the data for each body part of the beginners and that of the expert in each segmented section using the AMSS method. Those parts with high or low similarity were candidates for annotation criteria.

The procedure for extracting the annotation criteria is given below. First, the annotated scores of the level of proficiency of the ballroom dancer were normalized for each annotator. Second, the normalized scores were compared with those of the other annotators and the system extracted very different scores. We regarded the candidates for annotation criteria in those sections with different scores as clues for evaluating the level of proficiency of the ballroom dancer. Finally, the annotation criteria candidates in each section were shared among all annotators. The criteria were provided in the form of video and summary text documents, such as "the trajectory of the right hand is important but the speeds of both legs can be ignored" or "the angle between the left hand and torso is important but the speed of the torso can be ignored." The annotators determined whether the criteria were reasonable. The results of the judgment were also shared and used to construct consistent annotation criteria.

8.4.3.2 Results of the Leave-One-Out Cross-Validation

Table 8.2 gives the number of samples (good and bad) for each participant and the results of the leave-one-out cross-validation by discriminant analysis. The accuracy rates at point 1 and point 3 are high. Having compared the discriminant functions at each point, we found that the functions at point 1 and point 3 are similar for all participants at the respective point. From these results, we suggest that the system can extract reasonable criteria at point 1 and point 3. in particular, point 3 represents a typical situation in which the transitions of each joint are important in performing

Table 8.2 Results of the leave-one-out cross-validation

***a** Point1; head, neck, left foot*

Participant	A	B	C	D	E	F	G	H	Total
Bad sample	9	8	11	10	9	10	10	9	76
Good sample	8	12	9	15	13	10	10	8	85
Total									161
Cross validation									
Bad sample	6	8	10	10	9	4	10	4	61
Good sample	6	9	6	11	9	9	10	1	61
Total									122

***b** Point2; right shoulder, left shoulder, left elbow*

Participant	A	B	C	D	E	F	G	H	Total
Bad sample	7	8	11	23	14	15	12	9	99
Good sample	7	10	10	13	10	10	10	6	76
Total									175
Cross validation									
Bad sample	3	4	6	19	10	8	7	4	61
Good sample	4	8	6	9	4	8	2	4	45
Total									106

***c** Point3; left shoulder, right shoulder*

Participant	A	B	I	J	K	L	Total
Bad sample	13	12	12	12	12	12	73
Good sample	15	12	11	17	12	12	79
Total							152
Cross validation							
Bad sample	13	12	12	12	11	12	72
Good sample	14	12	11	15	12	12	76
Total							148

the dance motion well. We also confirmed that the system can extract time series features.

However, the accuracy rates at point 2 are relatively low and the discriminant functions are not similar for all participants at this point. To find the reason for this, we investigated the data for point 2 and found that the position of the left wrist was not captured correctly owing to an error in the measuring system. In addition, the system did not consider those criteria that changed depending on the body size and the body type, such as male or female, broad shoulders, slim, and so on. To incorporate adjustments due to these differences, we need to analyze and classify the criteria themselves.

Table 8.3 Accuracy of classification results using the discriminant function

Participant	A	B	C	D	E	F	G	H	Total
Bad sample	7	8	6	5	8	6	3	5	48
Good sample	6	9	6	5	6	5	5	5	47
Total									95
Cross validation									
Bad sample	7	8	4	2	8	1	0	2	32
Good sample	1	7	6	0	6	4	5	5	34
Total									66

8.4.3.3 Generality of the Extracted Criteria

To confirm the generality of the extracted criteria, the discriminant function was applied to data for another part of the dance step. Point 1 is conceptually an important point but we do not know whether the discriminant function supports the generality of the criterion. Thus, we used the function to confirm this. The discriminant function at point 1 was calculated using the data in "HandToHand" count 2. We then applied the function to the data in "OpenBasic," the results of which are given in Table 8.3. The function achieved about 70 % accuracy in classifying the data of another dance step. Therefore, the function has a certain degree of generality of the criterion.

8.5 Summary

In this chapter, we introduced a method of multi-modal interaction analysis towards better understanding of conversation and the case studies. First, we briefly described how to design experiments for analysis of multi-modal interaction. Especially, we explained the practical aspects of methodological issues involved in designing the experiment, use of annotation which was useful to analyze ambiguous phenomena and measurement of physiological indices which can be helpful in detecting changes of human mental processes. Next, we introduced three case studies of multi-modal interaction analysis. The first study concerned itself with the development of an "objective" physiological-signals based metric for measuring naturalness in conversational contexts. The second case study investigated measuring social atmosphere. In this study, we introduced analyses about ambiguous phenomenon mainly using annotation data. The final case study involved the analysis for extracting evaluation criteria for ballroom dance. In this section, we introduced an example of design for multi-modal interaction experiment and analysis using the obtained multi-modal data. There are diverse factors related to multi-modal interaction and methods to analyze multi-modal data. The experimental design based on clear hypothesis, however, is also important and foundational for the multi-modal interaction analysis.

Chapter 9
From Observation to Interaction

Abstract In this chapter, we will describe a framework of learning by mimicking for converting observation into proficient conversational behaviors. Individuals of some species can utilize the learning capacities of other individuals by mimicking their behavior. When this happens, biologists speak about culture. Humans are arguably the most sophisticated cultured species on earth and learning by imitation or mimicry lies at the root of their cultural abilities within which interaction and conversational behavior exists. This chapter starts by providing a general framework for learning by imitation covering its architecture and algorithmic details. It then moves on to describe applications of this framework in learning interactive behavior both as implicit and explicit interaction protocols. The insights learned from theoretical and experimental considerations are then utilized to provide a general framework for fluid agent-initiated imitation in the following chapter.

Keywords Imitation · Learning from demonstration · Theory of mind · Imitation in infants · Interaction babbling · Interaction structure learning · Embodied interactive control architecture

9.1 Imitation, Simulation and Conversation

As the reader will notice, this chapter is mostly about imitation in biological and artificial agents. Where does this *technology* fit within the conversational informatics world? This introductory section will focus on understanding this question and will try to provide an adequate answer to it based on three basic concepts: imitation, simulation and conversation.

9.1.1 What Is Imitation?

The term *imitation* is used by different research communities to mean different things but it always involves copying an otherwise improbable—or a novel—response (e.g.,

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_9

Zentall 2003). The mere behavior similarity between two agents does not necessarily imply imitation but may be a sign of one of the following related phenomena. According to Zentall (2003), these phenomena can be summarized as:

Contagion: Some behaviors are contagious but do not seem to involve an internal *representation* of the action. For example, yawning is contagious in humans. We do not consider contagious behaviors as full fledged imitation because they do not involve any form of *learning* and cannot—by themselves—provide a basis for cultural transfer. This does not preclude the possibility that contagious behaviors may have a social or psychological effect that is culture enhancing.

Social facilitation: The mere presence of the demonstrator may increase the probability that an agent will produce some behaviors. If this is working two ways for two agents, some behavior similarity may arise. This socially facilitated behavior similarity. For example the mere presence of another animal may increase the arousal levels of another animal leading to more bar pressing behavior. If the first animal was trained to press its bar, the two animals' behaviors may show higher than normal correlations that could be attributed to imitation. Nevertheless, in this case—similar to the contagion case—there is no *learning* involved and the behavioral similarity cannot be considered as full imitation.

Stimulus facilitation: If the demonstration involves object manipulation, these manipulations may move the object resulting in higher saliency. This higher perceptual saliency of the object may in turn induce higher probability of interaction with the object in a watching agent which may appear as an imitative act. Stimulus facilitation is very hard to control for in studies looking for imitative behavior in animals and infants and is specially problematic if the objects involved can be manipulated only in few ways (e.g., a bar or button that can only be pressed).

Learned Affordances: Watching a sequence of actions involving an object allows the watcher to learn the affordances of this object. If these affordances are limited, the watcher may later interact with the object in ways similar to the demonstrator without the *intention* to imitate. This phenomenon is related to emulation in which what is learned is the change in the object's state rather than the specific action that caused this change.

Depending of the goal of the study, we may focus on the *copying* aspect emphasizing that the learner must learn the specific response topography (i.e., the specific actions by which the response is made) which is normally the case in studies of imitation in animals (Zentall 2003). The emphasize may also be put on the *novel response* part which is a more pragmatic stance that can be found in robotics research (for a survey, please refer to Argall et al. 2009) that has some difficulty distinguishing imitation from learned affordances. In this chapter we employ the more pragmatic stance and accept as imitation behavior similarity that requires the existence of the imitated demonstration and that involves a novel behavior or behavioral sequence.

This can be justified by the fact that we are considering imitation as a *tool* for cultural learning and as such learned affordances or emulation can qualify as some form of imitation for our purposes if they can cause some skill transfer.

The lines between these four types of behavior similarity causes is not as clear cut as may be implied by the aforementioned definition is the sense that an agent may actually employ several of them in the same time. For example it was shown that children of ages 3 and 4 employ in some cases proper imitation in the immediate re-generation of behavior but use emulation for delayed reproduction (Simpson and Riggs 2011). This may indicate that children use at least two representations of the learned behavior and that memory affects which of the two representations is utilized for behavior generation.

Imitation is not widespread in the animal kingdom. For some time it was believed that it may exist only in humans and great apes. Nevertheless, some studies have demonstrated some forms of imitation in birds (e.g., pigeons and Japanese quail) (Dorrance and Zentall 2002).

9.1.2 Imitation in Infants, Children and Adults

Infants are known to participate in behavior copying early in life (Meltzoff and Moore 1997). Even though there is some debate about when exactly does infants have some intentional imitation capability, it is clear that at least adult humans use imitation frequently as a major learning methodology in new situations.

One early model of imitation in humans was Meltzoff's *active intermodal mapping* (AIM) model (Meltzoff and Moore 1997). A main assumption behind AIM is a nativist approach to imitation assuming that infants are born with a unitary capacity for imitation. This assumption is based on early studies in infant imitation that established behavior matching at early ages (few days) (Meltzoff 2005).

The nativist approach of Meltzoff and and Moore (1997) can be characterized as a *nativist-starting-point* position in which the infant comes to the world equipped with some innate (or womb-acquired) mental constructs that allows it to imitate some bodily actions as early as 42 minutes after birth. In that these accounts are different from the Skinnerian and Piagettian accounts that treat the infant mind at birth as if it was mostly empty of any structure. Nevertheless, these accounts are not nativist in the sense of Fodor according to which most of the psychological structure of the mind is already inborn. This means that these accounts accept a role of development in shaping the innate abilities of infants and modifying them during the course of life and interaction with the world and others. This is clear from the role of motion babbling and organ identification as developmental steps in Meltzoff's theory.

Recently, several studies have challenged the assumption that behavior matching in very early infancy are true imitation leading to a new position—that is not yet fully articulated—that employs a dynamical systems account of imitation taking the position that imitation develops during the first 2 years of the infant's life emerging out of infant's acquisition of different kinds of knowledge and motor, cognitive and social skills (e.g., Jones 2009; Ray and Heyes 2011). Nevertheless, AIM is still a viable model that is supported from some recent studies. For example, Soussignan et al. (2011) found that infants increased significantly their tongue protrusion when seeing disembodied human and artificial tongue movements, but not when seeing a 2D full-face protruding tongue. This result was interpreted as revealing the exploration of top-heavy patterns of the 2D face that distracted infants' attention from the tongue. Results also showed progressively more accurate matching (full tongue protrusion) throughout repeated exposure to each kind of stimulus.

It seems that the debate about neonatal imitation is still inconclusive and no consensus have been reached even after three decades of research. A recent review by Oostenbroek et al. (2013) reported on three main alternatives for understanding neonatal imitation: (1) neonatal imitation is a genuine act of social communication mediated through an abstract representational system; (2) the phenomenon is actually an involuntary, inborn reflex limited to tongue protrusion; and (3) imitation in newborns is a product of arousal stating that:

> These views continue to be maintained without much promise of resolution.

The most important point in this debate for our purposes is that imitation has a developmental aspect (in both accounts) which means that the social interaction shapes imitative ability as much as imitation shapes the social interaction and that—in turn—reflects on conversational intelligence.

Moreover, imitation may be at the heart of the earliest form of *dialog* that human beings are able to engage in. For example Nagy and Molnar (2004) have shown that infants are not passive imitators even in controlled imitation studies but they provoke the experimenter by spontaneous deferred reproduction (imitation) of previous acts of the demonstrator eliciting an imitation response in a form of a turn-taking that may be a rudimentary form and a precursor to verbal dialog in later stages of life.

One of the earliest and most widely acknowledged theoretical models of imitation in infants is the Active Intermodal Mapping (AIM) proposed by Meltzoff and Moore (1997) as an effort to understand infant facial expression imitation which was documented in several experiences at ages as small as 42 minutes. Figure 9.1 shows a schematic representation of the main hypothesis underlying this model. The infant uses exteroceptive sensors (e.g., vision) to perceive the facial expression of the other and represents it in what they called a supramodal representation space which is a form of a generalized representation space that matches both sensory and action spaces. The infant then uses proprioceptive sensors to measure its own organ pose (e.g., facial expression) and represents it on the same supramodal representation space. This step is crucial for AIM as this representation in a common space is an important step to allow the infant to compare its own behavior and the perceived behavior of the other. An equivalence detector is then used (in this common

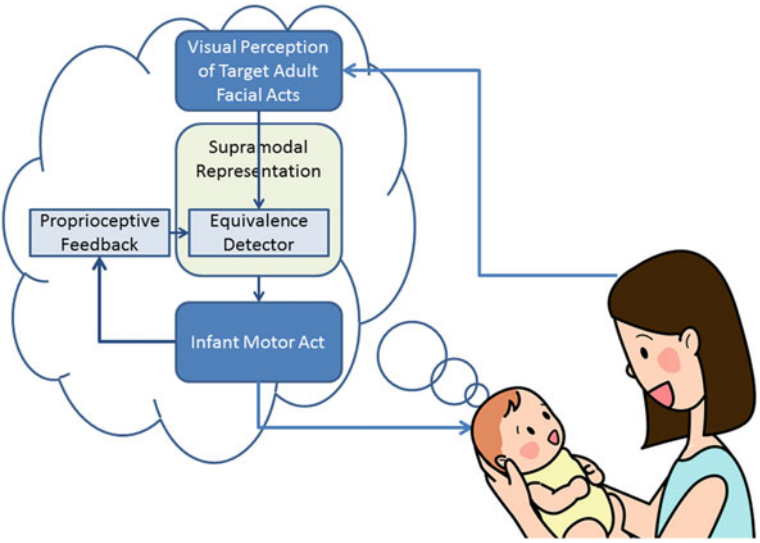


Fig. 9.1 Meltzoff's AIM model for infant facial expression imitation. © 2014, Yasser Mohammad and At, Inc. 2014. Reproduced with permission

supramodal space) to measure the difference between the two representations. This comparison allows the infant to infer that the perceived other is *like-me* which is the core of Meltzoff's theory of intersubjectivity.

Meltzoff and Moore (1997) propose (based on a wide range of experiments) ten main features of early imitation. Infants imitate a range of acts at this stage (this feature is challenged recently as we pointed out earlier). Imitation is specific in the sense that the infant does not confuse the organ doing the act or different acts of the same organ. Imitation is literal in the sense that it reproduces the same act perceived as faithfully as possible and—more importantly—infants correct their actions to produce a more and more faithful imitation but they quickly activate the appropriate body part. Even novel acts, absent actors, and static gestures can be imitated. It is also important that infants recognize that they are imitated and that imitation has a developmental aspect. These features were the basis for the assumption that imitation is representation mediated which led to the supramodal representation space idea.

Rao et al. (2004) propose four stages of imitation development: (1) body babbling, (2) imitation of body movements, (3) imitation of actions on objects and (4) imitation based on inferring intentions of others. These stages can be utilized not only for learning object manipulations as been suggested by Rao et al. (2004) but also can be the basis for learning interaction protocols in order for an agent to interact naturally in the social space created by human interactions.

The previous discussion may have led the reader to believe that imitation is an indispensable learning mechanism for humans and that it is always a useful aspect of human behavior. Although we agree with the first conclusion (imitation is an indis-

pensable learning mechanism), it is not the case that imitation is always beneficial due to a phenomenon called usually over-imitation (Kenward 2012). Over-imitation is defined as *the imitation of elements of an action sequence that are clearly unnecessary for reaching the final goal*. This peculiar human tendency has been argued to enable the development of unique aspects of human culture such as cumulative cultural evolution. In general, over-imitation is absent in the behavior of infants and it appears in early childhood increasing with increased age of the child.

Several hypotheses have been suggested to explain over-imitation ranging from social aspects of the interaction (e.g., the child's desire to look like the adult or to communicate an affiliation with her through faithful copying of actions), to casual hypotheses (e.g., the child's understanding that the action is causally related to the goal in some mysterious way) and senorimotor hypotheses (e.g., the child unconsciously imitate the irrelevant action(s) because of the underlying tendency to imitate demonstrations). More recently some researchers suggested that the child may be encoding the intentional irrelevant action as a normative action that *should* be done even though it does not have a direct causal connection with the stated goal of the behavior (Kenward 2012).

Imitation and over-imitation in humans are by no means phenomena of infants and children only. Adult humans also use imitation for learning new tasks (sometimes even leading to over-imitation) consciously and sometimes unconsciously. For example in a recent study McGuigan et al. (2011) allowed children and adults to watch a model doing some actions some of which are relevant (goal-directed) and others are irrelevant for a task of getting reward from a transparent puzzle box. The main finding was that both adults and children did over-imitate (imitated irrelevant actions as well as relevant ones) and over-imitation was even worse with older adults when the model was adult. This suggests that humans do not develop beyond imitation in some sense but they develop their ability to imitate more faithfully once they decide that the model is to be imitated. This may be one of the factors that generate the strong adhesive forces keeping together human cultures in different societies.

The role of imitation in understanding the development of human interactive and conversational skills got a new boost in the 1990s due to the discovery of mirror neurons which did not only shed light on a possible mechanism for *how* imitation works but also made it possible to understand *why* it is important. The answer seems to be that imitation (may be through the mirror neuron system) can be the basis of our ability to form and maintain a theory of mind which is an essential step in understanding others. The following section briefly introduces the concept of ToM (theory of mind) and relates it to our current discussion of imitation.

9.1.3 Understanding Others: Simulation

Theory of Mind (ToM) refers to the cognitive capacity to attribute mental states to self and others. Other names for the same capacity include *commonsense psychology*,

nave psychology, *folk psychology*, *mindreading and mentalizing* (Margolis et al. 2012).

To understand the intentions of other people, humans develop a theory of mind that tries to understand the actions of interacting partners in a goal directed manner that does not only utilize observable behaviors but also internal mental states like beliefs, desires, and emotions. Failure to develop this theory of mind is hypothesized to be a major factor in developing autism (ASD) (see for example the broken mirrors theory of ASD (e.g., Ramachandran and Oberman 2006; Biscaldi et al. 2013) and other interaction disorders).

The term *theory of mind* implicitly indicates that there is an object/representation that is used for prediction (i.e., a *theory*). This term is biased toward one of the two major attempts to explain how the mind can do this attribution of mental states as will be clear shortly. The terms mentalizing and mindreading are less biased because they reflect an active process. Even though the process understanding seems more appropriate for us, we will still use the term ToM as interchangeable with mentalizing and mindreading for the rest of this chapter due to its widespread use specially when related to artificial agents and robots.

Two major theories are competing to explain how humans do this sophisticated mentalizing process namely the theory of theory and the theory of simulations (Breazeal et al. 2005). The theory of theory hypothesizes that a separate recognition mechanism is available that can decode the partner's behavior while the theory of simulation suggests that the same neuronal circuitry is used for both generation of actions and recognition of those actions when performed by others (Davies and Stone 1995; Sabbagh 2004).

The theory of theory was the earlier to develop and can be traced back (in modern ages) to the famous essay "Empiricism and the Philosophy of Mind" by Sellars et al. (1956). The main idea here is that we attribute mental states to other people utilizing a theory (sometimes called folk psychology) that has three types of laws. The first type connects environmental states and history to mental states (e.g., people who did not eat for a long time are usually feeling hungry), the second connects internal states with each other (e.g., people who feel hungry will want to eat) and the final type connects internal states with external behavior (e.g., people who want to eat and know that there is food in the fridge will open the fridge). Using these three types of laws, it is possible to understand/predict future behavior as well as past behavior.

The nature of the folk psychology theory is in many cases considered similar to scientific theories in as much as it involves generating hypotheses and testing them empirically. It is entirely conceivable that humans have some internal representation of this theory and that it is the basis of the ToM (or it is the ToM). Nevertheless, as we will see later in this chapter, we believe that some form of simulation theory can better explain the wealth of evidence now available about mentalizing in humans.

One early finding that was very important for the development of studies of ToM both methodologically and theoretically was by Wimmer and Perner (1983). The reported a systematic cognitive difference between 3 and 4 years old children in a task called the false belief task. In one specific example of this task children are shown a chocolate bar which a puppet called *Alice* leaves in a counter and leave the

scene, then they see another puppet (e.g., *Bob*) coming into the scene and moving the puppet from the counter to a box. The children are then asked: *where will Alice look for the chocolate bar when she comes back?*

Knowing the correct answer (that Alice will look into the counter not the box) requires the child to ignore her own knowledge and ascribe to Alice a false belief in the location of the chocolate bar. Prior to age four, children typically fail this test (answering that Alice will look in the box) but starting at age four or five children easily pass the task.

Theory theorists explain this change by an internal change in the *theory* that the children use to understand the behavior of other agents (Alice). Before the age four, children are assumed to have a theory that understand desires and beliefs as relations between an agent and the world that cannot be *wrong*. At age four, the argument suggests, children change their conception of desires and beliefs and relate them to internal propositional representations that can be true or false allowing them to pass the false belief task. Even though this proposal is viable, it leaves several questions unanswered. What is the event that triggers this change in the theory? why is it that most children undergo this change at roughly the same age? How is this theory represented in the neural substance of the brain?

Several other questions can be raised now concerning this internal theory in general (not only in relation to the false belief task). The answer to these questions can be used to classify the theory-of-theory advocates into more groups.

The first of these questions is whether or not there is an underlying principle that limits the possible kinds of folk psychologies? Consider our third law from the previous examples: *people who want to eat and know that there is food in the fridge will open the fridge*. Why is that so? we can easily conceive of some kind of aliens that can have all of these believes and desires but will not be *rational* enough to open the fridge. Nevertheless, it is not likely that we will be able to meet any members of this alien species because there *irrationality* would have caused them to starve to death long time ago. This simple example suggests that *rationality* is expected to be present in any successful theory of mind because individuals who develop irrational ToMs will not be able to survive long enough in the complex physical and social world that the homo-sapiens species inhabit.

As an example of this rationality theory consider Dennett's *intentional stance*:

> [I]t is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations. Folk psychology, then, is idealized in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do (Dennett 1989).

One problem with these *rationality* theories in general is that it is not clear how much *rationality* should we ascribe to other agents. For example, it seems consistent with the ToM to ascribe to some one the belief that he may be wrong in some of his own believes and in the same time ascribe to him the belief of everyone of these believes but this is clearly a contradiction and cannot be *rational*. Another problem is that these theories were designed only with belief in mind (they may cover desire and

intention) but it is not clear how can these theories cover sensations and emotions. We can easily attribute feelings of pain to people but when is feeling pain a *rational* response?

A second question concerning the nature of the internal theory is how it relates with other abilities of the mind. One possible answer is that it depends on a general cognitive ability which seems to be the position taken by rationality theories. The modular-nativists take the opposite position that the ToM is a specific module (usually in the Fodorian sense) that uses its own internal representation and computations. It is conceivable to assume that ToM is modular but shaped by experience (which makes it a non-Fodorian module) yet most of the existing literature supporting modularity also ascribes to some form of nativism according to which the ToM is there from birth or (more frequently) develops as a form of maturation mostly independent of the social and environmental context.

One supportive evidence that is usually utilized in support of these modular-nativist theories is the specific impairment of the mindreading ability (ToM) in autistic children as compared with normal children or even children with Down syndrome. An early study for the support of this hypothesis was conducted by Baron-Cohen et al. (1985) in which they compared the performance of normal pre-school children, Down syndrome children, and autistic children on a false-belief task. All children had a mental age of above 4 years, although the chronological age of the second two groups was higher. Eighty-five percent of the normal children, 86 % of the Down syndrome children, but only 20 % of the autistic children passed the test. This specific impairment (the argument goes) suggests that ToM is a specific module that is affected in autistic children.

In its Fodorian form, the modular-nativist theory will have problems concerning the nature of this module. Firstly, modules—at least in the Fodorian sense—encapsulate their information which is—arguably—the most important feature of modules. The ToM cannot satisfy this condition except if it has limited access to information contained in other modules but this does not seem to be possible. For example, for the child to pass the false belief task, she must understand that the chocolate bar put in the counter will stay there which is hardly a mindreading task. Another problem is that modules should be domain specific: *only a restricted class of stimulations can throw the switch that turns [the system] on* (Fodor 1981). It is not clear what kind of *stimulus* satisfy this condition for ToM. Moreover, the specificity of mindreading failure in autistic children is not enough to support a purely modular theory because autism involves only motor problems (Biscaldi et al. 2013). Of course it can be possible that there is multiple module impairments in autistic children but a simpler approach that links these two failures has been suggested after the discovery of mirror neurons which supports a simulation theoretic understanding of mindreading. This shows that it is possible to understand this specific impairment without the requirement of having a strictly Fodorian module for ToM.

Based on the previous discussion, we argue with Goldman (2006) that a purely theory based explanation of mindreading in any of its incarnations is not completely satisfactory to understand this phenomenon. Nevertheless, there are several aspects of these *theory* theories that can be of value both in understanding mindreading and

in developing artificial agents that can develop a human-like ToM allowing them to achieve our final goal of artificial empathy (see Chap. 1). From the rationality theories, we salvage the importance of choice in attributing the *intentional stance* to agents and the requirement that agents must behave in a somewhat consistent manner to nudge their human partners to attribute some form of *agency* or rationality to them. This can inform artificial conversational agents designers in selecting the behavioral repertoire and algorithms guiding these agents. From the modular-nativist approach (as well as studies in early neonatal imitation) we salvage the idea of the existence of a-priori structures that allow agents to develop their ToMs during natural interaction with people. These structures need not be encapsulated in Fodorian modules (as we argued before) because of the need to interact heavily with other components of the cognitive structure of the agent, yet it is still possible to achieve some limited form of separation—at least in the computational resources—that allows us to focus on the development of ToM components without having to implement the complete cognitive structure of the agent. This is mostly an Engineering decision to make it possible to study artificial agents with limited forms of ToM without having to implement general intelligence in them.

The major other candidate for explaining mindreading (other than the *theory/folk psychology* theory with all its incarnations we discussed) is the simulation theory upon which we base a large portion of our work. We now turn to this theory and try to introduce it along with—what we think—is some of its supporting evidence. Figure 9.2 shows a simplified basic structure of simulation and theory theories for understanding how can 4-years old children pass the false-belief task. The most important difference is that while the simulation baby *puts herself in the shoes of Alice*, the theory baby uses *generalized laws of mental states* to deduce what Alice *must* be believing. Both agree on the final assessment but through very different routes. The main idea of the simulation theory is that we understand others by putting ourselves in their position and trying to see what will happen inside our heads if we were them (i.e., by simulating them). A major difference between the simulation theory and the theory of theory is that there is no need for generalized science-like laws that we generate by observing other's behavior (or that we have innately). The understanding is then driven by this simulation rather than propositional inference or any kind of inference.

Goldman (2006) tracked the history of simulation theory to Hume, Adam Smith and Kant. In modern times, the theory in its current form and the use of the term *simulation* can be traced back to Robert Gordon when he suggested that we predict others' behavior by answering the question: *What would I do in that person's situation?* (Gordon 1986).

If the simulation theory is true, then we run internal simulations of other agents using the *same* brain that is used for generating our own beliefs, emotions and motions. This immediately suggests two modes of failure. The simulation may leak into our own beliefs and emotions (contagion) and our own beliefs and emotions may leak into the simulation (attributing to the other some of our own beliefs, desires and emotions). There is no clear reason why these two kinds of failure should exist in the theory theories. As the reader may have already noticed, these two types of failures

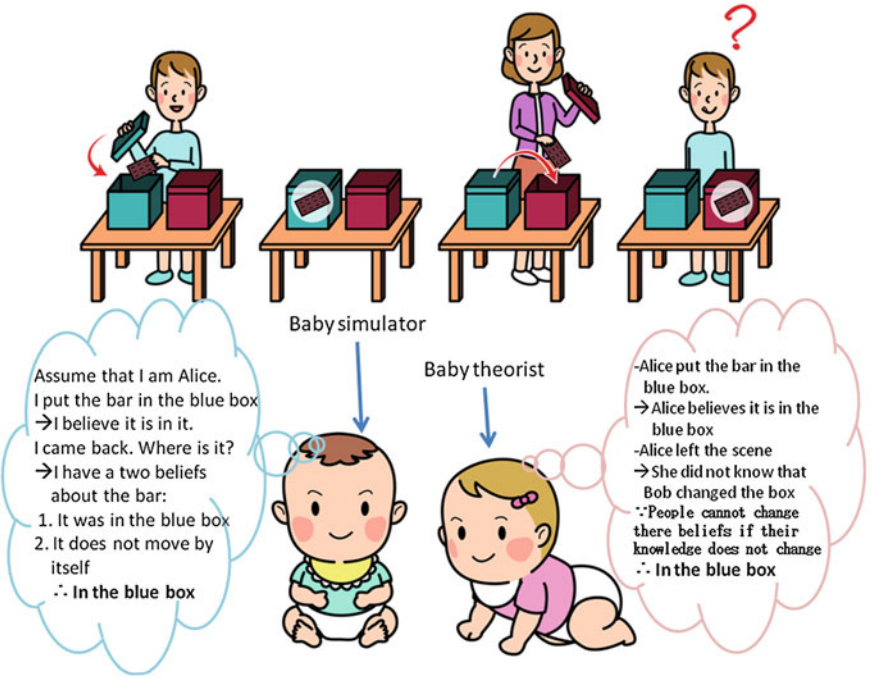


Fig. 9.2 A simplified basic structure of simulation and theory theories for understanding how can 4-years old children pass the false-belief task. The most important difference is that while the simulation baby *puts herself in the shoes of Alice*, the theory baby uses *generalized laws of mental states* to deduce what Alice *must* be believing. Both agree on the final assessment but through very different routes. © 2014, Yasser Mohammad and At, Inc. 2014. Reproduced with permission

are common enough to give—at least some weak—support to the simulation theory. The first type of failure mentioned may be—at least partially—behind the very known forms of contagion in human behavior ranging from yawning to emotional contagion and adult unconscious imitation that we mentioned in the previous section. The second type of failure in which we project our own desires, emotions and beliefs upon others is very common in everyday life and may be one of the contributing factors behind the failure of 3-years old children to pass the false belief task.

This is of course one a circumstantial evidence and a stronger empirical support is needed. Several strands of evidence are now available. For example, Kanakogi and Itakura (2011) found that the onset of infants' ability to predict the goal of others' action was synchronized with the onset of their own ability to perform that action. Moreover, action prediction ability and motor ability for some action were found to be correlated. This renders some support to a relation between action prediction (*understanding actions in goal directed manner*) and action performance; but that is what we would expect from a simulation-theoretic mind. The evidence of a correspondence in the impairment of mindreading (ToM) and motor abilities in autistic

children can also be marshaled as a supporting evidence for this point (Biscaldi et al. 2013).

A more direct supporting clue was found in the 1990s with the discovery in Parma of mirror neurons in the F5 premotor cortex area of macaque monkeys (Gallese et al. 1996). The myth goes that while one of the researchers was licking an ice-cream, the electrodes inserted in this area of a monkey's brain watching the act starting to buzz (Iacoboni 2008). Whether or not this ice-cream event was the onset of it, the discovery of mirror neurons was a landmark event in several areas of neuroscience and cognitive science. It mended together perception and action in a way compatible with the simulation theory (Iacoboni 2008).

It should be emphasized that the exact role of mirror neurons in learning and whether or not they imply a simulation theoretic ToM are not ubiquitously accepted by all researchers. In a short review, Molenberghs et al. (2009) suggested that the regions related to imitation in humans extend beyond the traditional mirror neurons areas.

Simulation theory in its first incarnations (e.g., Gordon 1986) dealt mainly with propositional attitudes like beliefs, desires and intentions (similar to rationality theories) but after the discovery of mirror neurons and the important of mirroring that we discussed earlier, it started to be generalized to cover *all* forms of mental state attribution including sensations and emotions (Gallese and Goldman 1998). This generality of the theory along with the supporting clues provided in previous paragraphs imply that it can be a basis for a natural conversational agent that can—at least in principle—evoke an empathetic agency response from its partners and hope for one day attaining a form of artificial empathy that we target as discussed in Chap. 1.

One major criticism of the simulation theory is that it is not adequate for explaining self-attribution of mental states (at least in its pure form) as discussed in details by Goodlman (2006) who proposes a hybrid simulation-theory approach that emphasizes the simulation aspect in much of the same spirit as the proposal we put forward computationally in this chapter. Nevertheless, we can—at least for now—ignore this criticism in the case of conversational agents because we are targeting the implementation of a ToM in the agent that corresponds to the human ToM in order to make it transparent to the partner allowing her to understand the artificial agent's behavior as naturally as possible. We are not interested in the self-attribution of the artificial agent (or even on the possibility of such self-attribution of mental states for an artificial algorithmic agent).

The following section summarizes our discussion of imitation and simulation theory and connects them to the theme of this book (conversational informatics).

9.1.4 The Road to Conversation

Conversation is a unique cultural construct of homosapiens. Imitation also seems to be a homosapiens speciality. Even though other animals communicate, their communication cannot be thought of as full fledged conversation and even though some

forms of imitation exist in the animal kingdom, it is also not on the same bar as human imitation and certainly does not have the same cultural implications for the species as a whole.

The relation between conversation and imitation is not just that they both seem to be uniquely human or even that both are essential for generating and maintaining human culture. More interestingly, imitation plays a crucial rule in learning how to conduct conversations both on the verbal and nonverbal dimensions and imitation can be improved through conversation.

Several studies looked at the effect of imitation on later language development. For example, Siegel (1981) showed in an early study of 148 infants that Uzgiris-Hunt scales—administered at 4, 8, 12, and 18 months—were significantly correlated with cognitive and language development at 2 years. Among the dimensions measured by these scales are gestural and vocal imitation. Charman et al. (2000) followed a sample of 13 infants longitudinally to 44 months after collecting measures of play, joint attention and imitation at 20 months of age. They found that Imitation ability at 20 months was longitudinally associated with expressive, but not receptive, language ability at 44 months which led them to speculate that joint attention, play, and imitation, and language and theory of mind, might form part of a shared social-communicative representational system in infancy that becomes increasingly specialized and differentiated as development progresses.

These results and many others show that imitation in infants is related to language development which is a necessary precursor for conversation. It can be argued that there is no enough evidence of a causal link between imitation and language development. Nevertheless, the mere correlation between the two phenomena is enough for our purposes here as it justifies looking at imitation as a factor that is at least related—if not causally connected—to verbal communication.

Conversation is not only about language. Nonverbal behavior plays an important role in conversations through what we call *interaction protocols* that enable partners in a conversation to get feedback about their utterances through the nonverbal behavior of other partners. Imitation is also an important factor in learning nonverbal interaction protocols. Later in this chapter, we will propose a computational framework for implementing an agent that can learn natural interaction protocols through imitation.

The relation between imitation and conversation is not one-way as may appear in the last few paragraphs. Conversation also can play a role in improving imitative skill. Consider two children trying to learn how to make a good tennis serve. The child with better conversational skills is expected to have a better chance at improving her skill by just conversing with the coach.

In this chapter we develop a computation model of imitation and an architecture that allows a robot or an agent to learn through imitation how to interact with humans after watching human-human interactions. Even though we will focus on nonverbal aspects of the interaction that we define as interaction protocols, the proposed approach is extendible in principle to verbal content of the conversation. For this chapter we define an interaction protocol as *a set of rules that govern the verbal/nonverbal behavior of the participants during interaction or conversation*.

Conversation cannot be authentic without a form of mental attribution which is provided by a theory of mind as discussed in the previous section. The simulation theory of ToM provides several advantages including being more supported by recent evidence of neuroscience (mirror neurons) and developmental studies (correspondence between action proficiency and understanding), fitting better with imitation as a driving force behind social cognition as well as a result of it (Heyes 2010) and being general enough to cover sensations, emotions and propositional attitudes. This suggests that a simulation theoretic agent is our best chance toward implementing a natural conversational agent that can aspire to achieve some form of artificial empathy naturalizing its interaction with human partners.

The connection between imitation, simulation and conversation that we tried to develop in this section is the heart of our proposed approach to the implementation of autonomous conversational agents. Simulation is used as the basis for agent's behavior generation while imitation is utilized to develop the internal processes of the agent that are used by the simulation engine. We call this approach SILI standing for Simulation based Interaction Learning through Imitation. The rest of this chapter discusses this model in more details and reports our initial attempts at implementing and evaluating it in real-world situations. Before that we will take a short detour to discuss some of the uses of imitation and simulation theory in artificial agents (specially robotics) research in order to better situate our proposed system.

9.2 Imitation in Artificial Agents

Imitation (also called learning from demonstrations or programming by demonstration) is becoming an important research area in robotics (Aleotti and Caselli 2008; Argall et al. 2009; Abbeel et al. 2010) because it allows the robot to acquire new skills without explicit programming.

There are two main directions in robotic imitation research. The first direction tries to utilize imitation as an easy way to *program* robots without explicit programming (Nagai 2005). This use usually goes by other names like *learning from demonstration* (Billing 2010), *programming by demonstration* (Aleotti and Caselli 2008) and *apprenticeship learning* (Abbeel et al. 2010). Researchers here focus on task learning. The second direction tries to use imitation to bootstrap social learning by providing a basis for mutual attention and social feedback (e.g., Nagai 2005; Iacoboni 2009). Researchers here focus on interaction learning. We can say that, roughly, in the first case, imitation is treated as a *programming mode* while in the second, it is treated as a *social phenomenon*. In this chapter, we focus on the task-learning aspect but we try to extend it to allow robots to learn from unaware teachers and from continuous streams of data without predefined action boundaries imitation in infants and animals.

As discussed before in this chapter, in some animals including humans, imitation is a social phenomenon (Nagai 2005) that was studied intensively by ethologists and developmental psychologists. Social psychology studies have demonstrated

that imitation and mimicry are pervasive, automatic, and facilitate empathy. Neuroscience investigations have demonstrated physiological mechanisms of mirroring at single-cell and neural-system levels that support the cognitive and social psychology constructs (Iacoboni 2009). Neural mirroring and imitation solves the *problem of other minds* and makes inter-subjectivity possible, thus facilitating social behavior. The ideomotor framework of human actions assumes a common representational format for action and perception that facilitates imitation. Furthermore, the associative sequence learning model of imitation proposes that experience-based Hebbian learning forms links between sensory processing of the actions of others and motor plans.

One of the major differences between learning through imitation and traditional supervised learning, is the availability of a limited number of training examples for the learner. This limits the applicability of traditional machine learning approaches like SVMs and BNs. Another major difference—that is usually ignored in LfD research—is that in real world LfD situations, the learner may have to detect for itself what behaviors it needs to learn as the demonstrator may not be always explicit in marking the boundaries of these behaviors or the dimensions of the input space that are of interest for learning.

For a robot to be able to learn from a demonstration, it must solve many problems. Most important of these problems are the following six challenges:

- *Action Segmentation*: How can the learner segment the continuous stream of actions (e.g., motion in the trajectory space) perceived from the demonstrator into discrete *behaviors* (e.g., a tennis serve, opening a door, etc.)?
- *Behavior Significance for Imitation*: How to know the interesting behaviors that it should imitate? What of the actions and state components (e.g., pose) of the demonstrator is related to the behavior to be imitated? This encapsulates the *what* and *who* problems (Nehaniv and Dautenhahn 1998).
- *Perspective Taking*: How is the demonstrator perceiving the situation?
- *Demonstrator modeling*: What are the primitive actions (or actuation commands) that the demonstrator is executing to achieve this behavior? What is the relation between these actions and the sensory input of the demonstrator?
- *Correspondence Problem*: How can the sensory input and actions of the demonstrator be mapped to the corresponding spaces of the learner?
- *Evaluation Problem*: How can the learner evaluate its performance after the imitation and use this evaluation to improve its performance? This evaluation would usually require feedback from the demonstrator or other agents and can utilize social cues (Breazeal and Scassellati 2002).

9.3 Implications of Simulation Theory for Interaction Modeling

Learning interaction protocols using imitation can be considered a process of identifying the relation between intended partner behavior and agent's behavior. This entails a theory of mind of the partner and the development of such theory can then be done using a computational implementation of either of the discussed two theories. As we discussed earlier, we subscribe to a modified form of the simulation theory in which we have reflective element that may be considered a precursor of a generalized *theory* but without the deliberative and propositional connotations usually attached to this term.

The main insights behind the design of this architecture are the following:

1. Nonverbal interaction protocols (specially spontaneous ones) are not specified at a single time resolution or abstraction level but should be specified at multiple layers of abstractions corresponding to multiple time scales. The proposed architecture can achieve that using the idea of a multiple layers called interaction control layers.
2. From the point of view of the cognitive processes, behaving in whatever role of the interaction should have similar if not the same computations inside the agent. This is a more involved point that is based on both the theory of simulation in developmental psychology and mirror neurons in neuroscience. The proposed architecture achieves this indistinguishability.
3. Behavior generation in humans tends to employ both bottom–up and top–down activation directions as well as both reactive and deliberative processes. A robotic architecture capable of human-like natural interaction would allow these combinations.

The following subsections will explain each of these points in more details.

9.3.1 Simultaneous Role Learning

Let's consider a simple interaction context in which an instructor is explaining to a listener how to operate some device. During this kind of interaction, many types of nonverbal behaviors can be found from both partners (roles). For example when the instructor focuses on some object for a while, the listener tends to look at this object causing a mutual attention event. Some times the instructor focuses on the face of the listener (e.g., to measure how much the listener follows the interaction), and the listener may respond by sharing gaze (a mutual gaze event) and may be nods to indicates that she is following the explanation. The instructor then may respond by moving her attention to some other object. These kinds of nonverbal behaviors can be considered as a *protocol* and the goal of SILI is to learn this (and similar) protocols.

An essential feature of these protocols is that they represent dynamical couplings of the intention functions of the two partners. As such these protocols cannot be just represented fully with sets of *if-then* rules because these sets of rules ignore the temporal synchronization aspect inherent in interaction protocols. This means that we can think of these protocols as forms of synchrony and learning them requires learning a representation of this synchrony.

An important feature of synchrony is that the behavior of one partner is related to the behavior of other partners and so it is impossible to know how to behave as one of them without knowing how to behave as the other at least partially. For example in the previous example there is no way to say when should the listener activate the *look-at-object* behavior without saying something about the behavior of the instructor (e.g., when the instructor focuses on the same object for more than 3 s and in this case it is expected that the instructor will break the mutual attention by a call for mutual gaze etc.). This implies that if we accept the reactive theory of intention and its subsequent that nonverbal interaction protocols are forms of synchrony then we have to accept that there is no way to model the behavior of one partner without modelling the behavior of other partners. This is what we mean by simultaneous role learning.

There are many available algorithms that can be considered for learning nonverbal interaction protocols but as we will discuss now they usually tend to model the behavior of each role separately from the other missing the critical dynamical aspect of synchrony that we discussed in the previous paragraph and this means they cannot support synchronous role learning.

One possibility is to represent the protocol as a set of rules like: $I_i \rightarrow L_j$ where I_i represents behavior i from the instructor, and L_j represents behavior j of the listener. A logical induction algorithm or a tree growing one (like CART or C4.5) can then be used to learn these rules. Even though this may work for explicit protocols (e.g., using iconic gestures) for spontaneous protocols, it will be hard to achieve the required speed needed for reactive behavior as shown convincingly by Brooks (1986).

Another possibility is to use a Bayesian network to encode the protocol which can take care of the probabilistic nature of nonverbal protocols and the differences between individuals. This approach can work as long as the interaction protocol works in a single time scale but real world human-human interaction needs multiple levels of synchrony that cannot directly be encoded in this framework. This will be discussed in details in the following section.

Even in the previous cases, the listener needs some way to detect instructor actions from its sensor data (I_i in the previous example). This means it must know something about instructor's behavior. If there is a way to automatically convert this detection algorithm to a synthesis one (i.e., generating I_i behavior using knowledge about how to detect it) then it is possible to learn both protocols for the instructor and listener simultaneously. This is simultaneous learning can provide more timely protocol implementation and also provides faster learning (especially if the number of roles is even more than two). One goal of SILI is to provide this simultaneous learning.

9.3.2 Hierarchical Interaction Layers

Natural human-human interaction requires synchrony in multiple time scales ranging from spontaneous fast body alignment (Nagaoka et al. 2006) to high level deliberative turn taking (Wiemann and Knapp 2007). If interactive robots are to achieve human-like natural behavior they are required to be able to achieve this synchrony with their partners at all time scales. This means that the interaction protocol cannot be encoded at a single time scale or abstraction level but needs to be specified at multiple scales and there is a causal relation in both direction in this hierarchy.

One of the goals of SILI then is to encode these levels of interaction in what we call interaction control layers. The details of these layers are given in Sect. 9.4.

9.3.3 Cognitive Indistinguishability Between Roles

This point is more subtle that the previously discussed ones. The theory of simulation suggests that the same mechanism is responsible for both detection and generation of behavior as we discussed in details previously (Davies and Stone 1995).

As learning the interaction protocol can be considered a process of identifying the relation between intended partner behavior and agent's behavior it entails a theory of mind of the partner and the development of such theory can then be done using a computational implementation of either of the discussed two theories. If the assumption of the theory of simulation are right then taking into its limits there should be no difference in the subconscious level between behaving in one role while understanding the behavior of other partners (in all other roles) and behaving in any other role of the interaction. For example, in the explanation scenario in which an instructor is teaching a listener how to achieve some task, there should be no internal difference in the lower levels of the architecture between acting as a listener and as an instructor. The only difference should be in how the final behavior commands are transferred to the actuators and how the sensor data are interpreted.

In SILI we decided to follow this assumption of the theory of simulation and design the system so that internally there is no difference between acting in any role of the interaction (e.g., listener, instructor, bypasser, etc.). This property combined with simultaneous learning of all interaction roles are unique features of SILI that cannot be found in available HRI architectures. As will be shown later, these features enables the DUD mechanism to naturally combine both top–down and bottom–up behavior generation approaches.

9.4 Interaction as Simulation: System Architecture

In this section, we provide details of the system architecture of SILI. The proposed system provides a simulation theoretic mechanism for recognizing interaction acts of the partner in a goal directed manner. This mechanism is augmented by a separate recognition mechanism to enable learning the interaction protocol using the interaction itself as will be explained later in this section.

Figure 9.3 shows the main building blocks of the system. Every interaction has a set of types of participation called *roles*. The cognitive structure of our system consists of one role for each type of participation with one of them being special in that it represents the self. The forward processes are processes that represent different behavior generating functions in the agent (e.g., there can a forward process representing looking at the face of a partner). Depending on when these processes happen to be during the interaction (i.e., as parts of which role) they constitute either the *control* component (if it happen to be in the self role) or the simulation component (if it was in any other role). This is the basis of cognitive indistinguishability between roles as will be clarified soon. The other types of processes running in the system are reverse processes that represent discovery of the activation of forward processes (e.g., there is a reverse process associated with the looking-at-the-face-of-the-partner forward process that predicts its activation). These processes represent the reflective part of the agent. They provide a proto-theory yet without deliberative or propositional connotations. Again reflective/reverse processes can occur in both the role of self and the role(s) of other(s). The other-reflection component in Fig. 9.3 represents reverse processes that appear in roles occupied by others in the current

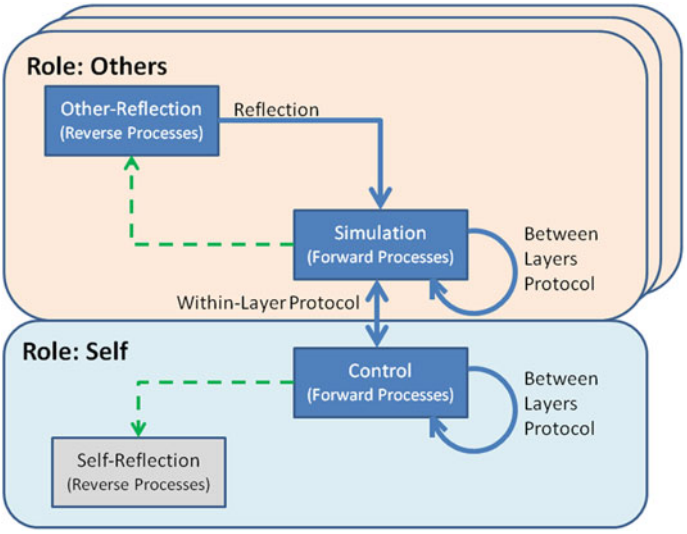


Fig. 9.3 A simplified version of the proposed architecture (SILI)

interaction and provide a complementary view of their behavior to support the simulation forward processes. The self-reflection component represents reverse processes that appear in the role occupied by the self in the current interaction and can be the basis of a form of self-identification and attribution of mental states to the self (in lines with Goldman's proposal for a hybrid simulation-theory approach discussed earlier). Nevertheless, in this book we are only concerned with the conversational aspect of intelligence and will not dwell any further on the possible utilization of this self-reflection component and will assume that it is disabled for the rest of the chapter. Nevertheless, it is important to notice that this component naturally arise in our architecture as dictated by the symmetry between the role occupied by the self and other roles in the interaction without any need to add it by hand to the architecture in the future to account for self-reflection in a life-like agent.

There are three main information flow channels in the proposed architecture. The first is what we call the within-layer protocol which represent the interaction at a specific abstraction level (and time resolution) between different roles (e.g., the joint attention behavior should emerge from the connection between processes in the self and the other roles activated in the same time and pointing the gaze toward the same object). Notice that this protocol is bidirectional. The Between-Layers protocol is the second information flow channel and it is restricted to within a single role in the interaction. This protocol represents the active simulation of the partner (if in one of the others roles) or the active control of the self at increasingly higher levels of abstraction (e.g., a follow object process in one layer will activate and deactivate head orientation control processes at lower layers in order to follow the face of a partner using this kind of protocol). Finally, the third flow is the reflection component which represent the ongoing abstraction of the interaction into behaviors of increasing timescale and is based on the reverse processes that represent the mirror of forward processes in the proposed architecture.

This bird's eye overview may not be all that clear at this point (which is why this section is just starting). Figure 9.4 gives a more detailed version of the current implementation of the SILI architecture. The most important parts of the architecture for the purposes of this book are as follows:

Perspective Taking Processes (PTPs): For every interacting partner a set of Perspective Taking Processes are spawned to provide a view of the interaction from the partner's point of view. Those processes generate the same kinds of signals generated by the agent's interaction perception processes but assuming that the agent is in the position of the partner.

Forward Basic Interaction Acts (FBIAs): The basic interactive acts that the agent is capable of. In this work we use Recurrent Neural Networks for these processes.

Reverse Basic Interaction Acts (RBIAs): Every FBIA has a reverse version that detects the probability of its execution in the signals perceived by the robot.

Interactive Control Processes (ICPs): These processes constitute the higher interactive control layers. Every interactive control process consists of two twin processes in the same way as FBIAs and RBIAs.

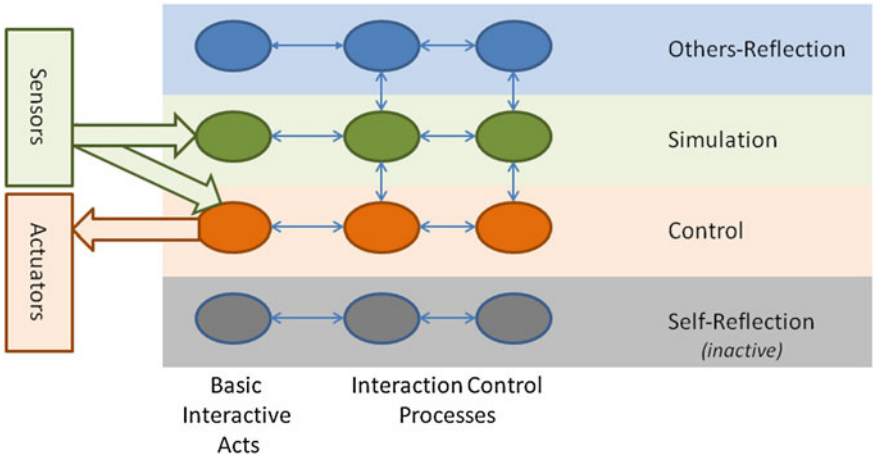


Fig. 9.4 A more detailed view of SILI architecture showing the main computational components and their relations. Also shown are the three main information flow pathes previously shown in Fig. 9.3

During interactions the processes of every layer are divided into two sets based on the role of the agent in the current interaction. The first set is the running interactive processes that represent the self. This set or processes runs only in the forward direction. The reverse direction can be used for self-reflection but this is beyond the scope of this book. The other set is running interactive processes representing other roles in the interaction. This set is further divided into processes running in the forward direction (down the hierarchy) representing the *theory* part of the system and processes running in the reverse direction (up the hierarchy) representing the simulation of other roles involved in the interaction. This gives rise to the three information processing paths distinctive of our approach and is shown schematically in Fig. 9.3.

Three factors (coming from the three information paths) can affect the activation level of interaction processes in the system: The first factor is the activation levels of various FICPs of the partner(s) (e.g., the frequency of looking at the partner is partially determined by whether or not this partner is speaking). The second factor is the activations levels of the FICPs representing the agent in the higher layer of the architecture (e.g., the frequency of looking at the partner is partially determined by whether the robot/agent is currently listening to this partner or is busy doing some other task). The third and final factor is the output of the corresponding RICPs as estimated from the behavior of the partner. This factor affects only the FICPs representing the other partners not the agent/robot itself. Formally,

$$act\left({}_iFICP_j^l, n\right) = w_w \times_i WP_j^l(n) + w_h \times_i HP_j^l(n) + w_t \times_i T_j^l(n), \qquad (9.1)$$

$$ {}_iWP_j^{\,l}(n) = \sum_{r \in \{roles\} - i} \left(\sum_{m=0}^{{}_rn_l} w_p \times act\left({}_rFICP_m^l, n-1 \right) \right), \tag{9.2} $$

$$ {}_iHP_j^{\,l}(n) = \sum_{m=0}^{{}_in_l} act\left({}_rFICP_m^{l+1}, n \right) \times F\left({}_rFICP_m^{l+1}, n \right), \tag{9.3} $$

$$ {}_iT_j^{\,l}(n) = R\left({}_iRICP_j^l \right), \tag{9.4} $$

$$ F\left(iFICP_m^l, n \right) =_i f_m^l \left(\left\{ act\left(\begin{matrix} {}_iFICP_{\{roles\}}^{l-1}, \\ n-1 : n-K \end{matrix} \right) \right\} \right), \tag{9.5} $$

$$ R\left({}_iRICP_m^l, n \right) =_i g_m^l \left(\left\{ act\left(\begin{matrix} {}_iRICP_{\{roles\}}^{l-1}, \\ n-1 : n-K \end{matrix} \right) \right\} \right), \tag{9.6} $$

where ${}_iRICP_k^l$ is the RICP number k of layer l in the hierarchy representing role i, ${}_iFICP_k^l$ is the FICP number k of layer l in the hierarchy representing role i, $act(P,t)$ is the activation level of process P at time t, w_* are constants, ${}_rn_l$ is the number of processes in layer l of the hierarchy representing role r, and $\{roles\}$ is the set of all defined roles in the interaction.

This feedback generates a dynamical system that evolves over time joining the behaviors of the partners at a level of abstraction that increases with the layer number.

For simplification a two-agent interaction scenario (e.g., a listener-speaker scenario) will be considered in this section. Generalization to interactions that involve more than two agents is straightforward. A set of designer supplied sensor processes continuously translate the input stream into the partner's frame of reference, while the reverse basic interaction acts are measuring the most probable value of the actionability of various basic interaction acts of him/her/it. This is then fed to the reverse processes in the higher layers to generate the expected actionability of all the ICPs. This constitutes the theory about the intention of the other agent at different levels of detail based on the learned interaction structure. This is moving from bottom up in the interactive control layer hierarchy.

The forward direction of processes representing the partner is also executed at the whole hierarchy to generate the expected actionability of each of them according to the simulation of the partner. This is moving from the top down in the hierarchy. The difference between the theory and the simulation is used at every layer to drive the adaptation system only if the difference is higher than a threshold that depends on the age of the agent. Currently, we use a threshold that increases linearly with the age (Mohammad and Nishida 2008b). After adaptation mirror training is used to bring the reverse and forward processes of the simulated partner together. The DUD behavior generation mechanism is then used to generate final behavior as described in Sect. 9.6.

9.5 Simulation Based Interaction Learned Through Imitation

One of the main advantages of the proposed architecture is that all the BIAs and ICPs are learnable through imitation. The learning process progresses in two distinct phases. During the first phase, the system learns the basic competencies needed to achieve the required behavior by watching human-human or other human-robot interactions and mining these interaction records for recurrent patterns called motifs that can be used later to develop specialized controllers allowing the robot to achieve these basic behaviors. These behaviors are encoded as FBIAs in the proposed system. The second stage allows the robot to build increasingly more complex behaviors by controlling the activation levels of FBIAs. This stage ends by learning the complete ICP hierarchy and by its end, the robot becomes ready for actual human-robot interactions. The following subsections briefly describes these two stages. For more details please refer to Mohammad and Nishida (2009a).

9.5.1 Interaction Babbling: Learning BIAs

The first stage of development of the robot/agent aims at learning the forward basic interactive acts (FBIAs). This stage is called interaction bubbling to emphasize its relation with motor bubbling that allows new born babies to explore their motor abilities and learn the basic motor functions they can do. Similarly during interaction bubbling the robot (agent) learns how to use its sensors and actuators to achieve basic behaviors related to interacting with humans. The details of the algorithms used during this stage were given by Mohammad and Nishida (2008a). Here we provide an overview of the proposed technique.

The input to the learning mechanism are records of natural human-human or human-robot interactions. The robot first tries to discover recurrent patterns in the behavior of different actors (roles) in these interactions. The robot then associates a controller (dynamical system) with each of the discovered patterns capable of generating the required behavior (each such controller is a forward basic interactive act FBIA). Finally the mirror trainer is invoked to learn the reverse basic interactive act corresponding to each of the learned FBIAs.

The most critical step in this algorithm is the discovery of recurrent behavioral patterns (motif discovery). Given that the input to the robot is a multidimensional time series representing the behaviors of interacting agents, the problem can be coined as motif discovery from time series. There are many available techniques for solving this motif discovery problem as we discussed in Sect. 4.5. There is a common problem to all of these available algorithms which is their inability to utilize constraints or domain knowledge to speed up the discovery process which results in superlinear operation in all cases. Given that the length of the time series involved is usually high (e.g., hundred's of thousands or millions of time steps) in order to represent fast nonverbal behaviors like gaze shifts etc., a superlinear solution is too slow for our

application. Also in this application the relations between the behaviors of interacting partners can be a useful clue for the probable locations of recurrent patterns (motifs) that are related to the interaction and can be useful for rejecting motifs that are not important for the interaction. Again, most available algorithms cannot utilize such relations to increase the accuracy and relevance of the discovered motifs. Constrained motif discovery (CMD) algorithms do not have these two limitations because they can utilize constraints evaluated from the signal itself or provided using domain knowledge to speedup the search for recurrent patterns. We discussed CMD earlier in this book in Sect. 4.5. For more detailed explanation of two such algorithms, the reader is referred to Mohammad and Nishida (2009a).

The constraints used with the CMD algorithm are usually derived from a change point discovery algorithm. The following subsection introduces change point discovery and describes some algorithms for solving it.

9.5.1.1 Change Point Discovery

The research in change point (CP) discovery problem have resulted in many techniques including CUMSUM (Page 1954), wavelet analysis (Kadambe and Boudreaux-Bartels 1992), inflection point search (Hirano and Tsumoto 2002), autoregressive modeling (Gombay 2008), discrete cosine transform, and singular spectrum transform (SST) (Ide and Inoue 2005). Most of these methods with the exception of SST either require ad-hoc tuning for every time series (e.g., wavelet analysis), discover a single kind of change (e.g., CUMSUM discovers only mean shifts), or assumes a restricted generation process (e.g., Gaussian mixtures). The main disadvantages of SST though are the sensitivity to noise and the need to specify five different parameters.

Change point discovery (CPD) has a very long history. Survey papers for this problem can be found as early as 1976 (Willsky 1976). One of the earliest methods for CPD was based on modeling the signal by two models: A long term model that is based on a growing window (M_0) and a short term model that is based on a sliding window of fixed length. Both models used were autoregressive models with additive Gaussian noise. The difference between the predictions of the two models was measured using Kullback's divergence and this difference gave an estimate of the change point score at every point. Andre-Obrecht (1988) applied this method to speech segmentation as early as 1988.

Early on (e.g., Willsky 1976; Basseville and Kikiforov 1993), it was recognized that CPD involves two different subproblems. The first one is generating a change score at every point (sometimes called *residual*). This score varies smoothly in most cases and may start to rise before the actual change and fall after it. Ideally this change score is zero everywhere but at change points. The second problem is to use these change scores to localizes specific change locations in the input signal (this is called the *localization* problem). In some applications, it may be enough to have the scores (see for example the case study presented in Sect. 8.2).

A general class of solutions for CPD involves converting the signal into a stochastic process with parameterized probability distribution ($P_\theta\ (x_i|x_{i-1}, x_{i-2}, \ldots, x_0)$) then deciding whether at some point in the input signal (or a sliding window over it for online approaches) there is a change in θ from θ_0 to θ_1 against the alternative that a single parameter vector θ explains the whole signal. This is a likelihood ratio test and there exists statistical method to decide it (e.g., Basseville and Kikiforov 1993).

One of the earliest approaches that follow this scheme is the CUMSUM algorithm (Page 1954) which is sometimes called *Page-Hinkley* stopping rule. The goal here is to detect a sudden change of the mean of some process. This can be used in conversation analysis for detecting the times at which a sudden change in gaze direction occurred. Because of noise in gaze direction detectors, the gaze direction can be modeled as a random walk around a mean value that represent the target. Sudden change in gaze can then be modeled as a change of this mean.

For simplicity let's start by assuming that both the mean before the looked-for change μ_0 and after it μ_1 are known. This means that we can model the signal using the following very simply model:

$$y_i = \mu_i + \mathcal{N}\left(0, \sigma^2\right), \tag{9.7}$$

where σ is the (possibly unknown) standard deviation of the Gaussian noise added to the means and μ_i is equal to μ_0 before the point of change (r). If a change happens at the point r, then μ_i will equal μ_1 for $i \leq r$ otherwise it will equal μ_0.

In this case our hypotheses can be stated rigourously as:

$$\begin{aligned} H_0 &: r > n, \\ H_1 &: r \leqslant n, \end{aligned} \tag{9.8}$$

where r is the change point we are looking to locate and n is the length of our signal (or a sliding window on it).

Given the assumptions of this problem, we can define the likelihood ratio between the two hypotheses by:

$$\prod_{k=r}^{n} \frac{p_1(y_k)}{p_0(y_k)}, \tag{9.9}$$

where $p_j = \mathcal{N}\left(\mu_j, \sigma\right)$.

Now the log of this likelihood can be written as:

$$\mathcal{L}_n(r) = \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{k=r}^{n} \left(y_k - \frac{\mu_1 + \mu_0}{2} \right) = \frac{1}{\sigma^2} S_r^n(\mu_0), \tag{9.10}$$

where

$$S_r^n(\mu) = (\mu_1 - \mu_2) \sum_{k=r}^{n} (x_k - \mu - \zeta/2). \tag{9.11}$$

This assumes that we actually know r but in reality that is what we are after (the change point). For this reason, we can replace r by its maximum likelihood estimate $\hat{r}$ which can be calculated as the value that maximizes S_r^n (notice that σ is independent of this value). This gives us the following estimate for $\hat{r}$:

$$\hat{r} = \underset{1 \leqslant r \leqslant n}{\operatorname{argmax}} \left(\prod_{i=0}^{r-1} p_0(y_i) \prod_{j=r}^{n} p_1(y_j) \right) = \underset{1 \leqslant r \leqslant n}{\operatorname{argmax}} \left(S_r^n(\mu_0) \right). \tag{9.12}$$

After calculating $\hat{r}$, we can then decide that a change happens within the n-point signal (or window) if $\max\limits_{1 \leqslant r \leqslant n} \left(S_r^n(\mu_0) \right)$ was higher than some predefined threshold τ. This leads to the CUMSUM statistic which can be computed recursively using:

$$g_n = (g_{n-1} + y_n - \mu_0 - 0.5(\mu_1 + \mu_0)). \tag{9.13}$$

Whenever g_n exceeds the predefined threshold τ we announce a change at the point $\hat{r}$ estimated as described above.

Of course in real life we do not know μ_0 and μ_1 but we can use the same strategy used for r and use their maximum likelihood estimates from the data.

9.5.1.2 Robust Singular Spectrum Transform

As the discussion of CUMSUM showed, most CPD algorithms require the specification of some threshold (e.g., τ in CUMSUM) that is used for localization. They also assume some predefined model of the signal (e.g., a constant value corrupted by a Gaussian noise in CUMSUM). In our application for motor-babbling, the signals we deal with are very different ranging from speech to motor commands to motors. It is very hard to have a model for each kind of these signals and it is also very difficult to decide apriori appropriate threshold values. A promising approach that can overcome all of these problems is based on SST. There are several SSA based CPD algorithms in literature (for a survey, the reader is referred to Mohammad and Nishida 2011b). Here we discuss one of the simplest versions that was used in early implementations of SILI.

To overcome the aforementioned problems, we proposed the RSST transform (Mohammad and Nishida 2009e). The essence of the RSST transform is to find for every point $x(i)$ the difference between a representation of the dynamics of the few points before it (i.e., $x(i-p) : x(i)$) and the few points after it (i.e., $x(i+g) : x(i+f)$). This difference is normalized to have a value between zero and one and named $x_s(i)$.

The dynamics of the points before and after the current point are represented using the Hankel matrix which is calculated in two steps:

1. A set of subsequences $seq(t)$ are calculated as:

$$seq(t-1) = \{x(t-w), \ldots, x(t-1)\}^T . \tag{9.14}$$

2. The Hankel matrix is calculated as the concatenation of n overlapping subsequences:

$$H(t) = [seq(t-n), \ldots, seq(t-1)] . \tag{9.15}$$

Singular Value Decomposition (SVD) is then used to find the singular values and vectors of the Hankel Matrix by solving:

$$H(t) = U(t)S(t)V(t)^T, \tag{9.16}$$

where $S(i-1, i-1) \leq S(i, i) \leq (i+1, i+1)$.

Only the first $l(t)$ left singular vectors ($U_l(t)$) are kept to represent the past change pattern as the hyperplane defined by them. Ide and Inoue (2005) showed that this hyperplane encodes the major directions of change in the signal. In RSST the value of $l(t)$ is allowed to change from point to point in the time series depending on the complexity of the signal before it. To calculate a sensible value for $l(t)$ we first sort the singular values of $H(t)$ and find the corner of the accumulated sum of them ($l_{inf}(t)$) (the point at which the tangent to the curve has an angle of $\pi/4$).

To find a first guess of the change score around every point, RSST tries to utilize as much information as possible from the future Henkel Matrix ($G(t)$) by using the $l_f(t)$ Eigen vectors of $G(t)G(t)^T$ with highest corresponding Eigen values ($\lambda_{1:l_f}$). The value of $l_f(t)$ is selected using the same algorithm for selecting $l(t)$.

$$G(t)G\,(t)^T\, u^g = \mu u^g. \tag{9.17}$$

$$\beta_i(t) = u_i^g,\ i \leq l_f \quad \text{and} \quad \lambda_{j-1} \leq \lambda_j \leq \lambda_{j+1} \quad \text{for} \quad 1 \leq j \leq w. \tag{9.18}$$

Each one of these l_f directions are then projected onto the hyperplane defined by $U_l(t)$.

The projection of $\beta_i(t)$s and the hyperplane defined by $U_l(t)$ is then found using:

$$\alpha_i(t) = \frac{U_l^T \beta_i(t)}{\left\| U_l^T \beta_i(t) \right\|}, \quad i \leq l_f. \tag{9.19}$$

The change scores defined by $\beta_i(t)$s and $alpha_i(t)$s are then calculated as:

$$cs_i(t) = 1 - \alpha_i(t)^T \beta_i(t). \tag{9.20}$$

The first guess of the change score at the point t is then calculated as the weighted sum of these change point scores where the Eigen values of the matrix $G(t)$ are used as weights.

$$\hat{x}(t) = \frac{\sum_{i=1}^{l_f} \lambda_i \times cs_i}{\sum_{i=1}^{l_f} \lambda_i}. \tag{9.21}$$

After applying the aforementioned steps we get a first estimate $\hat{x}(t)$ of the change score at every point t of the time series. RSST then applies a filtering step to attenuate the effect of noise on the final scores. The filter first calculates the average and variance of the signal before and after each point using a subwindow of size w.

$$\mu_b(t) = \frac{\sum_{i=0}^{w-1} \hat{x}(t-i)}{w}. \tag{9.22}$$

$$\sigma_b(t) = \frac{\sum_{i=1}^{w} \left(\hat{x}(t-i) - \mu_b(t)\right)^2}{w-1}. \tag{9.23}$$

$$\mu_a(t) = \frac{\sum_{i=1}^{w} \hat{x}(t+i)}{w}. \tag{9.24}$$

$$\sigma_a(t) = \frac{\sum_{i=1}^{w} \left(\hat{x}(t+i) - \mu_a(t)\right)^2}{w-1}. \tag{9.25}$$

The guess of the change score at every point is then updated by:

$$\tilde{x}(t) = \hat{x}(t) \times |\mu_a(t) - \mu_b(t)| \times \left|\sqrt{\sigma_a(t)} - \sqrt{\sigma_b(t)}\right|, \tag{9.26}$$

where $\tilde{x}(t)$ is then normalized to get $x(t)$ which represents the final change score of RSST.

9.5.2 Imitation's Road to Interaction: Learning ICPs

ICP learning can be achieved using the Interaction Structure Learner (ISL). The Interaction Structure Learner is invoked by a set of training examples and a set of verification examples. Every example (c) is a record of the sensory information (${}^{c}s$) and body motion information (${}^{c}_{i}m$) of every agent i during a specific interaction c of the type to be learned.

FBIAs are already learned using constrained motif discovery, while RBIAs can be learned using the mirror trainer (Mohammad and Nishida 2009f).

The outputs of the interaction structure learner are: the number of layers needed to represent the interaction ($l_{\max}$), the number of ICPs for every role in each layer ($_r n_l$), and parameter initialization of every FICP ($_r FICP_j^l$) in every layer l for every role in the interaction r and the corresponding RICPs ($_r RICP_j^l$). These initial values of the parameters can then be adapted while the agent is interacting in various roles using Interactive Adaptation Manager (IAM) as described in detail by Mohammad and Nishida (2008b).

One of the most important parameters of the Interaction Structure Learner is the expected set of synchronization frequencies in the interaction. This can be found from known human-human interaction data like the H^3R Explanation Corpus (Mohammad et al. 2008). If this set is not available the system must try many frequencies until a good synchronization frequency is found. Currently the system simply uses multiple line searches starting from random frequencies.

The ISL algorithm involves the following steps:

1. Project sensor readings to the perspective of every partner in the interaction using the PTPs.
2. Convert the outputs of the PTPs of all interactions into activation levels of the FBIAs using the RBIAs.
3. Learn the interaction control processes starting from layer l incrementally one layer at a time until one layer contains at most one process for every agent as follows:
   (a) Apply constrained motif discovery for the activation levels of layer $l-1$ processes to get $_i n^l$ motifs characterizing the building blocks of the behavior at this level of abstraction.
   (b) For layer l use the occurrences of these motifs to train $_i n^l$ RBFNNs representing the FICPs of every role i in this layer.
   (c) Apply this learned FICPs to the validation set and find the difference between its outputs and the action activation levels at layer $l-1$ and accept every FICP if this error level is acceptable.
   (d) Learn the within-layer protocol connections of every forward process in layer $l-1$ as a linear function of the one time step delayed activation levels of the partner representing processes in layer $l-1$.
   (e) Invoke the mirror trainer to learn the RICPs corresponded to the accepted FICPs

9.6 Simulation Based Behavior Generation

This section presents the behavior generation mechanism used to generate the final behavior after the system is already learned through the process described in Sect. 9.5. Behavior generation in SILI does not follow a serial top–down or bottom–up

approach. It is rather a distributed mechanism in which the activation level of each process is calculated based on the state of other processes connected to it.

Figure 9.3 shows the three information paths active during behavior generation called the *theory*, *simulation* and *control* paths.

The *theory* path is composed from reverse interactive control processes that represent the agent's understanding of the partners behavior. This is a bottom–up information path with information passing from lower to higher layers of the interaction protocol. Conceptually, this path represents the answer to the question: *what are my partners doing giving what I perceive from his behavior?*. Notice that not only data but also activation goes bottom–up in this information path. This is the theory aspect of SILI.

The *simulation* path is constituted of the corresponding forward interactive control processes associated to all roles in the interaction other than the one currently occupied by the agent. Information and activation are going from the top down in this path. Conceptually, this path represents the answer to the question: *what would I be doing if I was in the position of my partner given my current high level goals?*. This is the simulation aspect of SILI.

The *control* path is the current controller that directly generates the behavior of the agent. Information and activation are going from the top down in this path as well. Conceptually, the question answered by this path is: *what should I do?*. Notice that even though this is the only path of control that directly affects the actuators of the robot, its behavior is in part controlled by the other two active paths which means that the final behavior is indirectly affected by the simulation running in the simulation path and theory represented by the theory path. This integration of these three aspects of situation understanding and control are what gives SILI its power.

Figure 9.5 shows the factors affecting the activation levels of FICPs and FBIAs. The only difference between FICPs and FBIAs in this regard is that for FBIAs the activation level is translated to intentionality value in EICA (see Sect. 10.1.1) terminology while for FICPs it is translated to actionability. The actionability of FBIAs is connected always to 1 as long as this interaction protocol is activated by higher level deliberation layers which are not a part of SILI. For simplicity we assume that the robot has a single interaction protocol implemented. For all other purposes FBIA and FICP activation is the same so we will speak about FICPs only in the rest of this section but it should be clear that concerning activation level, the same rules apply for FBIAs.

As Fig. 9.5 shows, the activation level of each FICP is controlled through three factors:

1. The activation level of the corresponding RICP from the theory or inactive paths. This is called the theory factor. In case of the ICPs representing the current role, the theory factor is zero because the activation level of all inactive RICPs is set to zero.
2. The activation level of other FICPs corresponding to other roles in the interaction (from the control or simulation paths). This is called the within-layer protocol factor. For the ICPs corresponding to the current role this within-layer protocol

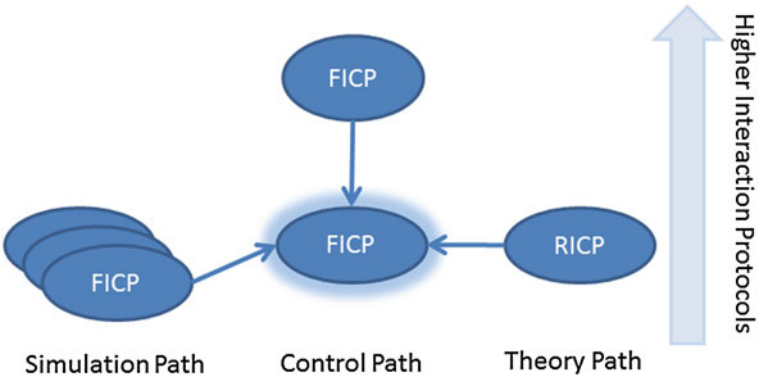


Fig. 9.5 The factors affecting the activation level of basic interactive acts and interaction control processes

factor is coming from ICPs in the simulation path while for ICPs representing other agents in the interaction, this within-layer protocol factor comes from ICPs in the control path. This within-layer protocol factor represents the protocol at one time scale.

3. The output of higher level ICPs corresponding to the self for the control path and to the same other agent in the simulation path. This is called the inter-layer protocol factor. This factor represents the top–down activation representing the relation between the interaction protocols at different time scales.

The architecture itself does not specify how to combine these factors. In most of our implementations though we simply average the three contributions using the activation levels of their sources as weights (Mohammad et al. 2010).

After understanding how each ICP and BIA is activated we turn our attention to the meaning of the inputs and outputs of them. The input–output mapping implemented by BIAs and ICPs is quite different and for this reason we will treat each of them separately. Moreover reverse and forward versions are also treated differently.

The input for the FBIAs representing the current role (in the control path) are taken directly from the sensors while its output go to the actuators. The inputs for the FBIAs representing all other roles in the interaction (the simulation path) are taken from perspective taking processes and the output is simply discarded. In actual implementation the FBIAs corresponding to other agents are disabled by having their activation set to zero to save their computation time. Even though in this discussion it seems that there is a difference in the wiring between FBIAs corresponding to current role and other FBIAs which does not agree with the cognitive indistinguishability rule described in Sect. 9.1.3, this difference is just an effect of the simplification in drawing the figure. Using the set of MUXs and the DEMUX, as was done by Mohammad et al. (2010), the wiring of all FBIAs is identical and the role variable controls at run time which FBIAs are given the sensor data directly and which are

given the output of perspective taking processes. Also the same variable controls which FBIAs are connected to the actuators of the robot and which are not.

The FICPs in control and simulation paths have exactly the same inputs and outputs. This feature is in agreement with the cognitive indistinguishability suggested by the theory of simulation. The inputs of the FICP are the last n_z activation levels of the FICPs in the same path but the lower interaction control layer (the output of their effect channels connected to their activation level). n_z controls the relative speed of the FICPs running in this layer and the lower layer. In general each FICP can use a different n_z value. The output of each FICP is the inter-layer protocol factor affecting the FICPs of the lower layer in the same path. This means that FICPs of layer i implement a higher level slower protocol on top of the protocol implemented with the FICPs in layer $i-1$. Mohammad and Nishida (2009c) described how the number of layers, the number of FICPs in each layer and the strengths of the inter-layer protocol factors are learned.

Keeping in mind that all RBIAs and RICPs in the inactive path are disabled, we will describe only the operation of the RBIAs and RICPs in the theory path. The reverse processes in the inactive path have exactly the same wirings but with the control layer and they are not considered here because they are always disabled (when they are enabled the role is reversed and the control and simulation paths are also reversed and in this case the reverse processes currently in the inactive path become the theory path and vice versa). The inputs to RBIAs in the theory path are the last n_z values of all the sensors sensing the behavior of the partner. The output is the probability that the corresponding FBIAs are active in the partner control architecture assuming that the partner has the same control architecture (the central hypothesis of the theory of simulation). This output is the theory factor affecting the activation level of the corresponding FBIA. It is also used for adapting the parameters of this FBIA. RICPs are doing exactly the same job with relation to FICPs. The only difference is that because of the within-layer and theory factors the RICPs have a harder job detecting the activation of their corresponding FICPs compared with RBIAs. To reduce this effect, the weight assigned to the RICPs is in general less than that assigned to RBIAs in the effect channels connected to their respective forward processes.

9.7 Summary

Our discussions in this chapter wandered between cognitive science, neuroscience, psychology and developmental studies looking for threads that unify the underlying competencies required for interactive agents. Based on this analysis of existing research in these fields we argue for an intimate relation between imitation as a learning strategy, simulation as a behavior generation mechanism and interaction or conversation as basic constructs of human sociality. Based on this relation, we proposed an architecture for conversational agents that utilizes simulation as the main building block of its behavior generation allowing it to combine seamlessly information from lower and higher layers in the cognitive hierarchy. The proposed

architecture also utilizes imitation to learn the computational processes (to be used for control and simulation during behavior generation). The following chapter will show examples of how can this approach be used to endue robots with human-like interactive abilities that are rooted in the same cognitive mechanisms of humans' sociality.

Chapter 10
Applications of Simulation and Imitation for Interaction Learning

Abstract In this chapter, we will describe two case studies that utilized the architecture presented in Chap. 9 which utilizes ideas from the simulation theory of mind for behavior generation during interaction and imitation learning as a technique to develop the required computational processes needed. The first case study will be concerned with gaze behavior during face-to-face interactions while the second will be concerned with a newly proposed paradigm for learning from demonstration that we call fluid imitation. Fluid imitation allows the agent to learn interactive behavior (or any kind of behavior) not only through intended demonstrations but from unintended ones during day-to-day operation. The chapter concludes with some ideas for other possible applications of the architecture to other aspects of conversational informatics.

Keywords Gaze control · Fluid imitation · Self-initiated imitation · Behavior significance · Reactive gaze control · Dynamic structure gaze controller · Interaction dimensions

10.1 Case Study: Learning Gaze Behavior

Gaze is an important nonverbal interaction modality in human–human conversational situations (e.g., Kendon 1967; Argyle 2001). Studies in HRI have shown that gaze behavior is of similar importance in the human–robot interaction case (e.g., Imai et al. 2002; Sidner et al. 2004; Yamazaki et al. 2008). People were shown to accurately perceive and interpret that robot's focus of attention by utilizing cues from the robot's gaze behavior (Imai et al. 2002). The fact that people assign a *focus of attention* at all to robots is interesting because of its implications for anthropomorphic attribution of a cognitive faculty (attention) to what is basically a machine. In our experience, this kind of attribution can be elicited with minimal design effort as long as the behavior of the robot is interpretable in using some form of the intentional stance.

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_10

It is not only that people can interpret gaze behavior of robots. Sidner et al. (2004) have shown that a well-designed gaze controller for a robot can significantly increase the engagement level of its human partners in face-to-face situations. The mere belief that an observed agent is an intentional system (human or robot) was shown by Wiese et al. (2012) to significantly increase gaze cuing effects and it did not matter whether the agent had a face of a human or that of a robot.

These studies and a wealth of other studies support the need for a natural gaze controller for robots specially humanoids. There are several studies that were reported in literature for designing and implementing gaze controllers for robots that targeted achieving human-like behavior (e.g., Atienza and Zelinsky 2005; Yonezawa et al. 2007). In this book though—and being true to the spirit of focusing on data-intensive approaches to conversational informatics—we report a case study the focus mainly on developing such controllers unsupervisedly from human–human interaction data based on the SILI architecture introduced in Chap. 9.

As a base line system for testing the performance of SILI based gaze control, we use a reactive system developed earlier by Mohammad and Nishida (2010a) which used a floating point genetic algorithm and a human-inspired model-based reactive architecture for building the gaze controller. Learning in this case was used to adjust the parameters of a general architecture that was designed to be suitable for gaze behavior and the system relied heavily on a specific algorithm for gaze-map generation and maintenance. Even though this is a step in the direction of truly autonomous learning of gaze behavior, it still requires the design of a set of computational processes based on a model of human–human interactions in similar situations. In this case, the model was based on a known approach-avoid architecture in human spatial behaviors during close encounters (Mohammad and Nishida 2010a).

In this chapter we explain the implementation and evaluation of a gaze controller that was first presented by Mohammad and Nishida (2014b) based on the architecture proposed in Sect. 9.4, and learned through imitation based on the three stages developmental approach presented in Sect. 10.1.5.

The following section briefly presents the reactive gaze controller (as developed by Mohammad and Nishida (2010a)) which is the system we compare with in this work. The book then presents how can SILI be used to develop a gaze controller autonomously. After that, we report an experiment to evaluate the proposed approach by comparing the proposed gaze controller with the reactive gaze controller. The work presented in this section is based on Mohammad et al. (2010) and Mohammad and Nishida (2014b).

10.1.1 Reactive Gaze Controller

The Embodied Interactive Control Architecture (EICA) is a robotic architecture designed to achieve natural interactive behavior. EICA was designed to be a generic base architecture over which more specific computation models (called *levels of specification* by Mohammad and Nishida (2009f)) are built. The architec-

ture is massively parallel and provides custom control of the computational resources assigned to each running process as well as the level of effect it can have over other processes running in the system (Mohammad and Nishida 2009f). This section gives a brief introduction to the architecture and its utilization for gaze control. The basic components of this architecture and their relation are shown in Fig. 10.1.

Any control system implemented in EICA consists of a set of running processes without a predefined hierarchical structure (contrast that with SILI) which gives the designer maximum control over the assignment of resources to these processes as well as complete control over their relative relation. This extreme flexibility comes at the price of giving very little guidelines in how to implement some behavior using this architecture.

The heart of EICA is an action integration mechanism that takes the actions *proposed* by the running processes and combines them using two stages. The first stage is a distributed action integration stage that is implemented implicitly by *effect channels*. These effect channels are constructs that allow processes to influence the activation level of other processes. The system provides predefined effect channels and allows the controller designer to extend them using any mathematical mapping between the outputs or activation levels of source processes and the activity level of the target process. The second stage of action integration is achieved via a central action integrator process that ultimately controls the final commands sent to the actuators of the controller.

Activity level of processes is controlled via two distinct attributes (that are dynamically calculated based on the topology of the controller and the current attributes of other processes). The first of these attributes is the *attentionality* which directly controls the computational resources that can be devoted to a process. This translates

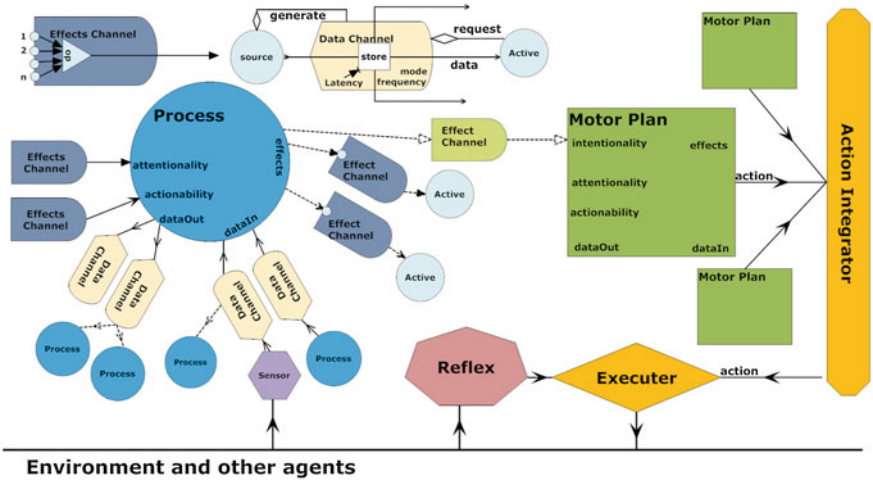


Fig. 10.1 EICA architecture used for the reactive gaze controller (Mohammad and Nishida 2009f). See the main text for more details. © 2009, Springer. Reproduced with permission

to a mechanism for distributed implementation of attention focusing. The second attribute is *actionability* which controls how much can this process affect the activity level of other processes through effect channels.

There are two distinctive topologies associated with any EICA controller. The first one defines the connections between input and output ports of different processes and this topology determines the directions of data flow in the system. The links of this topology are called data channels in the jargon of EICA. The second topology defines the connections and types between activity level attributes of different processes and this topology determines the path of control in the system which can be very different from the data flow graph. As described earlier the links in this topology are called effect channels.

EICA uses three main types of active components: *Motor Plans* that can affect the behavior of the robot directly, *Processes* that only affect the activation levels of other *Active* components, and *Reflexes* that directly access the actuators of the robot and are used for safety routines.

Mohammad and Nishida (2010a) used EICA to implement and evaluate a series of gaze controllers that aimed at achieving natural human-like gaze behavior. The best performance was achieved by a gaze controller that combined careful design of the data flow and control topologies with learning of process parameters using a floating point genetic algorithm. This gaze controller was called the *Dynamic Structure Gaze Controller* by Mohammad and Nishida (2010a) and will be referred to as the EICA controller in this chapter. The structure of the controller and the types of processes involved was based on an approach-avoid mechanism inspired by spatial behavior of humans in face-to-face close encounters. FPGA was only used to learn the parameters of the interacting processes and not the effect channels. This is a clear contrast with traditional connectionist approaches that focus on learning the *weights* associated with the connection links and use a predefined fixed set of computational processes (neurons).

The scenario used for evaluating this controller was a natural listening scenario in which a human *instructor* explains to a humanoid robot, Robovie II (Kanda et al. 2002) how to assemble and use a device that is disassembled over a table between the listener and the instructor. The performance of this controller was evaluated in two ways: Firstly, the percentage of the interaction time spent in mutual attention, mutual gaze, and gaze toward instructor where compared with values obtained from human–human interaction and showed remarkable resemblance to the human behavior. For example, 26.87 % of the time was spent in mutual gaze compared with 28.15 % for the human–human interaction case and the difference was not statistically significant. Similar results held true for mutual attention and gaze toward instructor behaviors. The second evaluation technique involved using the edit distance between the behavior of the listener robot and a human listener to quantitatively assess the human-likeness of robot's behavior. It was shown also that the *Dynamic Structure Gaze Controller* achieves high resemblance to actual human's gaze behavior in the same situation. For more details about this evaluation procedure and other controllers that were shown to be inferior to the EICA controller please refer to Mohammad and Nishida (2010a).

10.1.2 SILI Controller

An early version of SILI was used to develop a gaze controller by Mohammad et al. (2010) and Mohammad and Nishida (2014b). In this section, we report the general approach to controller design using SILI using this work as a core example and focus on the practical design decisions involved in SILI based controller design. More technical aspects of the study can be found in the original publications.

SILI is concerned with the development of interaction capabilities of the agent. Interaction implies being able to perceive the environment and behavior of other agents as well as being able to affect them. This is why the *designer* is expected to pick up a set of representative interaction dimensions that allow the agent to perceive important aspects of its surroundings and other agents'/humans' behavior as well as a set of behavior primitives (verbal or nonverbal) that has the potential of modifying these perceived interaction dimensions.

The specific kinds of interaction dimensions relevant to any interactive situation vary considerably. For example, learning natural proximities behavior requires the agent to perceive its distance and orientation to humans as well as the distribution of power in the situation which can affect spatial behavior significantly. Another agent designed to achieve natural turn taking will need to perceive nonverbal aspects of the speech signal generated by other partners, and may be some features of the content of this speech signal. As this last example shows, it may not be obvious what are the interaction dimensions relevant to a specific interactional or conversational task. The nonverbal content of speech may not appear as a very important aspect of turn taking in the first glance but research in dialog based interaction shows that this kind of signal is very important to turn taking. People seem to be able to detect appropriate turn-taking points in a conversation even if understand nothing of the speech content. Acoustic features are not the only nonverbal features related to turn-taking in conversation. Gaze behavior and other nonverbal cues are also taken into account by the speakers in natural human–human conversations. These kinds of behaviors would be of value for the SILI controller as well. Even cultural aspects (Stivers et al. 2009) were found to affect turn-taking behavior to some level.

Another challenge with selection of interaction dimensions is to make them not only comprehensive but also as orthogonal as possible. This will make it much easier to specify the primitive behaviors that can modify each one of these interaction dimensions with minimal impact on the rest of them. As we will see, this is important to make the interaction-babbling stage converge to a useful set of basic interactive acts (BIAs) to be used by the interaction structure learner as discussed in Chap. 9.

This discussion shows that it may not be trivial to decide which aspects of the situation are better representations of interaction dimensions and one of the most challenging aspect of controller design in SILI is the selection of these interaction dimensions. A least challenging problem is to provide a set of motor primitives to change the values of these interaction dimensions. These are the only manual steps required by the designer as will be cleared later in this chapter.

Given these two sets of perceptual and motor primitives and training data from human–human interactions, SILI learns a hierarchical representation of the interaction protocol.

As discussed in Chap. 9, this learning process is done in three stages. The first stage involves the discovery of recurrent patterns in the interaction dimensions that represent basic interactive acts. These BIAs are learned in the input interaction dimensions by employing a constrained motif discovery algorithm (see Sect. 4.5 where the first guess of motif locations (e.g., the constraint) is found using a change point discovery algorithm (e.g., RSST as explained in Sect. 9.5.1). The importance of these BIAs is two folds: They represent important aspects of the interaction (otherwise, why are they so ubiquitous in the training data) that allow the agent to reduce the dimensionality of the learning problem. Rather than having to learn the interaction protocol based on millions of data points that have little meaning in themselves, the agent can now learn the interaction in terms of the activation/deactivation of these BIAs which is generally a much smaller and more representative dataset. The second advantage on basing the interaction on BIAs relates to behavior generation. Each learned BIA can be thought of as a basic component of the interaction and a controller is then generated that allows the agent to produce this BIA in its behavior. This simplifies the problem of controller generation as the agent will learn to generate controllers that can achieve only these short (and in most cases smoothly varying) BIAs. This divide-and-conquer approach is indispensable for any form of engineering and learning to interact naturally is not any different in this aspect.

The second stage (see Sect. 9.5.2) involves learning higher level representations of the interaction protocol by progressively finding patterns in the activation levels of interactive processes in lower levels of the hierarchy.

These two main stages of SILI are offline stages that require no interaction on the part of the robot. It is not surprising that artificial agents may need a lot of offline computation before being able to develop natural interactive capabilities. After all, human infants require much longer time (in the order of years not minutes) to be able to engage in adult-like natural interactions as children. These two stages can be thought of as the infancy period of SILI agents. This analogy though should not be taken too far because infants are not as passive as previously thought as we discussed in the previous chapter. Nevertheless, the analogy is useful in highlighting that the behavior learned by SILI should not be expected to be perfect which means that subsequent adaptation of this behavior may be necessary during actual interaction with human partners. In previous research, we proposed an adaptation algorithm that can modify the learned hierarchy of interactive processes but we will ignore this possibility for the rest of this case-study report for the sake of brevity and to focus the attention on the learning process of SILI which is the heart of our approach. Interesting readers can refer to Mohammad and Nishida (2009f).

Assuming that the designer has decided the set of interaction dimensions and basic motor primitives, and in order to apply SILI's two-stages learning algorithm, she will still need to have a training set of human–human interactions and after SILI learns the interaction protocol, she will need to evaluate the learned behavior using human–agent interactions. This following subsections discuss the methodological

issues and choices available to the designer during each of these steps in the context of gaze control.

10.1.3 Interaction Dimensions

As we just discussed, selection of interaction dimensions is the first (and only manual) step in the process of using SILI. Human's gaze behavior depends on several factors including information seeking, backchanneling during conversation, behavior entrainment among many others and it is not trivial to decide upon a set of basic interaction dimensions to use. One possible set of interaction dimensions are just the skeletal positions of all tracked joints of interacting partners during the interaction as well as gaze direction and speech signals. After all, nonverbal behavior of people is fully represented by this set of features. It is instructive to understand why each one of these candidates is not a good interaction dimensions in itself.

Consider skeletal data of the interacting partners in the global frame of reference. These signals provide poor interaction dimensions for several reasons. Firstly, they are sensitive to the location of interacting partners in the environment. This makes similar behaviors appear differently in the timeseries data preventing BIA discovery. This problem can easily be handled by using an ego-centric frame of reference but which *ego* should we center the frame around. The first option that comes to mind is the frame of reference of the human whose behavior we are trying to learn but SILI utilizes indistinguishability of roles in learning and learns the behaviors of all participating humans in a single run. The second option is to project the data into multiple frames of reference representing the ego-centric view of each human involved in the interaction. Other than the obvious problem of increasing significantly the number of dimensions used for learning, this representation is redundant and violates the recommendation of having orthogonal interaction dimensions in order to make it easier (or even possible) to provide the required motor primitives that can selectively change the values of interaction dimensions.

These considerations show that raw skeletal data is not a useful set of interaction dimensions for SILI. Consider on the other hand, audio information representing the speech signal of each person in the interaction. This time it seems obvious that this signal is important for gaze control and it does not suffer from the ambiguity of the frame of reference associated with skeletal tracking. For example, if one human participant in the interaction hears the statement "*look here*", she will likely look in the direction pointed to by her partner. This suggests that audio information is very important for adequate gaze control. We do not argue that audio information has no value for learning gaze control but we will argue that it is not essential—at least—for an initial exploration. Even in the example we just mentioned, the information about the location of the object to be looked at is redundantly encoded in the gaze direction of the partner which can easily be inferred from the skeletal data alone. Moreover, decoding this statement using current state of speech recognition and understanding it using natural language processing without imposing limitations on the kind of

speech that can accompany the behavior (we are after *natural* behavior after all) is a challenging task that may introduce more noise and errors into the data than SILI can handle. For these reasons, it seems that audio information may not be on the same level of important compared with features extracted from skeletal data.

After showing that raw skeletal data are not adequate and audio data is not essential, we turn to the criteria for selecting interaction dimensions and try to tease out a methodology for selecting interaction dimensions. The goal is to minimizing the effort required to select these dimensions (after all this is one of the main reasons we are considering the data-intensive SILI approach in the first place) while making sure that provided dimensions are adequate for SILI learning.

The four major criteria for interaction dimensions selection are orthogonality, relevance, frame-independence, and role-independence. It should be clear that these three criteria provide *guidelines* and not stone-hard *requirements* for interaction dimensions selection. Moreover, in our experience, SILI can still work and achieve adequate performance even if some of these criteria were violated.

Orthogonality refers to the need of having interaction dimensions as independent as possible. This criterion stems from the need to provide basic motor primitives that can modify these interaction dimensions independently. This requirement may be difficult to achieve completely and some correlation between interaction dimensions may always exist in real world situations but this correlation needs to be minimized. If multiple dimensions have high correlations, they should be considered as components of a single interaction dimension. This was one of the reasons that we rejected raw skeletal data as interaction dimensions for gaze control. If such skeletal data were in themselves relevant to the interaction protocol being learned (e.g., a collaborative sport), then most likely the dimensions encoding the motion of a single joint should be encoded together as a single interaction dimension.

Relevance refers to the guideline that each interaction dimension needs to be relevant the interaction protocol being learned at least in some stage of it. SILI's basic interactive act discovery stage can in many cases reject irrelevant dimensions. Moreover, the interaction structure learner can in many cases provide a second line of defense against irrelevance of some interaction dimensions to the behavior being learned during the whole interaction or parts of it. Nevertheless, spending some time in selecting relevant dimensions may prove useful in reducing the time needed to learn the interaction protocol specially that the complexity of the problem increases exponentially with the number of BIAs used.

Frame-independence refers to the guideline that interaction dimensions should not depend on the exact frame-of-reference used to collect the data. This is of special importance if the collected dataset (as was the case in our gaze controller) were collected in a global frame of reference. Learning a frame-dependent behavior may make it impossible (or at least hard) to generalize the behavior to other situations that may involve a change in sensor modalities or arrangements.

Role-independence refers to the slightly more subtle guideline that interaction dimensions associated with different roles in the interaction should be similar (ideally the same). Consider an explanation situation in which you use the speech of the speaker/teacher and ignore the audio signal of the listener. SILI may be able to learn

an interaction protocol but it would not be able to discover in its own this difference between the two roles. This means that when using SILI for future interactions with people, you will have to tell it explicitly who is the speaker and who is the listener. This is obviously a limitation that could be avoided by just providing the audio signal of the listener even if it was not relevant to the interaction protocol directly. This case shows that the four criteria we discussed for selecting interaction dimensions can some time be conflicting and—as in any design task—it is the role of the designer to best resolve these conflicts.

Let's see how did the interaction dimensions we used in the gaze controller being discussed fair under these four guidelines. Mohammad et al. (2010) used the following interaction dimensions:

1. Three absolute angles of each partner's head (θ_y for yaw, θ_p for pitch, θ_r for roll).
2. Head alignment angle (θ_h) defined as the angle between the line connecting the forehead and back of the head sensor of the two partners.
3. Body alignment angle (θ_b) defined similar to the head alignment angle.
4. Distance between two interacting partners (d).
5. Difference between center of body coordinates in the three spatial dimensions (X, Y and Z) (d_x, d_y, d_z).
6. Salient-object alignment angle (θ_s) defined as the angle between the line connecting back of the head sensor and forehead sensor and the line connecting back of the head sensor and the location of maximum saliency according to the gaze map maintained using the algorithm described by Mohammad and Nishida (2010a).

Orthogonality is clear for most of these dimensions but there are some exceptions. The distance between the interacting partners (d) is easily discoverable from the difference between center of body coordinates (d_x, d_y, d_z). Why did we provide these two related interaction dimensions? This shows one strength of SILI that may be obscured by the detailed discussion for interaction dimensions guidelines. SILI can in most cases ignore irrelevant dimensions and when the designer is not clear about which *representation* of the interaction dimension should be used, she may just include several representations with the only price of increased processing time in many cases. In this case, it was not clear for us whether the distance of the centers of body coordinates which encodes to some level the difference in body height is essential or the difference in the horizontal coordinates that do not encode this difference. As it turned out SILI learned BIAs in both dimensions signaling that both were important.

Most of these dimensions is obviously relevant to the interaction protocol being learned (gaze control). Nevertheless, some potentially relevant dimensions were not included. We discussed earlier the reasons for not including any audio information in the processing. Another seemingly missing dimension is the location of objects in the environment. The inclusion of this dimension may seem important given that people gaze to different objects in the environment at different times to signal shared attention, request the establishment of mutual attention, or just for information collection. Nevertheless, including the locations of objects in the environment would have violated the frame-independence guideline. Just changing these locations in a

new situation may have lead to very strange behavior by the learned SILI controller. Still, we needed some form of data about locations of objects to be passed to SILI because of the importance of this information in the interaction protocol. To resolve this problem we utilized a learning approach that would incrementally build a gaze-map representing the relative importance of different *locations* in the interaction space during the interaction (Mohammad and Nishida 2010a). These locations usually correspond to locations of salient objects in the environment but without being hard-coded into the training data. Moreover, this approach provided a simplified form of attention focusing by providing SILI with data about the most *salient* object in the environment only which reduced the processing required by the system.

Looking again at the set of interaction dimensions, it is clear that all of them depend on no global frames because they all are ego-centric. Most of them are not even dependent on the form factor of the learner (e.g., all angles) but some are implicitly dependent on the different of this form factor (e.g., d) which is still independent from the specifics of any single partner in the interaction. What about role-independence? This is a feature of the whole set of interaction dimensions not any single one of them. It can be examined by considering every dimension related to a role and finding if a corresponding one exists for every other role in the interaction. In our case, we had two roles and examining the aforementioned set of interaction dimensions show that all of them either come in pair corresponding to both roles (e.g., absolute angles of the heads), or are independent from that distinction (e.g., d).

10.1.4 Training Data Collection

Training data collection is a critical step in any machine-learning based approach because the quality of learning depends on the quality of the dataset provided for the learner.

Human–human interactions are complicated and situation dependent and collecting useful training examples of them is not an easy task as discussed in Chap. 8. The first decision is whether to depend on existing human–human interaction datasets or relying on newly collected training examples. This decision is mostly determined by the adequacy of existing datasets in terms of the availability of the decided upon interaction dimensions (or being easy to extract from the data) and other normal dataset quality issues including accuracy of measurements. For human–human interactions, accurate timing of the behavior of different partners and synchronization between different modalities are two other factors that must be taken into account.

One good news for the designer is that SILI uses offline processing which allows some margin of error. If the selected interaction dimensions are not adequate we can just run it again using a different set of interaction dimensions. In many cases, it is easy to spot failures in SILI's learning (e.g., when no BIAs are learned, when higher interaction structure learning fails to converge, when learned BIAs correspond to no meaningful episodes of the interaction or when the causal relations expected between different roles are not obeyed by the learned protocol). Most of these failures are easy

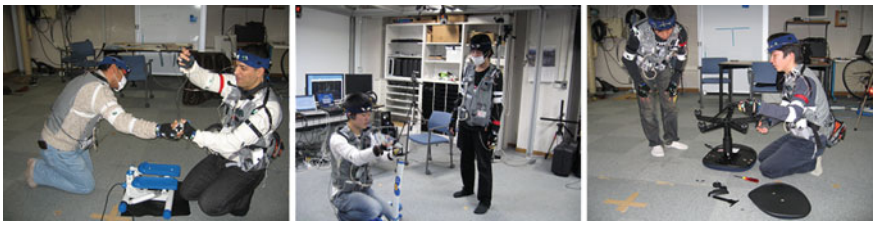


Fig. 10.2 Few frames from the training data collection experiment showing the speaker and the listener

to spot without even the need for running a real-world evaluation experiment. To utilize this flexibility, we have to be sure that the interaction date we collect contains many potential interaction dimensions that can be reliably derived from it. This is the reason that Mohammad et al. (2010) and Mohammad and Nishida (2014b) used the H^3R interaction corpus which contains not only skeletal data but synchronized audio data and locations of all objects in the environment.

For the purpose of learning gaze behavior, we used a subset of the interaction records provided by the H^3R interaction corpus (Mohammad et al. 2008). The part of the corpus used for gaze control learning by Mohammad et al. (2010) and Mohammad and Nishida (2014b) consisted of 22 sessions in which a listener is attending to an instructor while (s)he explains the operation of assembling and disassembling one of two comparably complex objects. Figure 10.2 shows three frames from three of these sessions. Participants could stand, sit or move freely in the environment. This freedom was somewhat limited by the task because both partners needed to be near the explanation object and the listener needed to be able to see the behavior of the instructor. Skeletal data provided by a PhaseSpaced motion capture system was used to find the interaction dimensions introduced in the previous section.

10.1.5 Learning Through Imitation

After collecting the training data from human–human interactions, a SILI controller was developed using the approach proposed in Chap. 9.

It is instructive to examine the basic interactive acts (BIAs) learned using SILI from this dataset. This will give the reader a sense about the level of complexity in behavior expected at this level of abstraction in the interaction protocol. Five BIAs were reported to be learned in this study. Two of them represented looking at the partner or at the salient object. Two of them represented body alignment with the partner or against her and the fifth corresponded to nodding.

It is apparent from this list that BIAs do not represent complex behaviors. All of these learned behaviors with the exception of nodding appeared as linear increase or decrease in a single interaction dimension. This simplicity of BIAs is a major

advantage for practical application of SILI and it can only be achieved by careful selection of the interaction dimensions according to the guidelines mentioned earlier.

It is curious that the only nonlinear motion that were learned (i.e., nodding) represented an oscillation up and down in a single interaction dimension as well (θ_p). This is not limitation of SILI as other applications of the system shows. For example Mohammad et al. (2009) showed that SILI can be used to learn complex gestures and navigation actions in a guided navigation task even when limited to a single interaction layer above the BIA layer. The reason for the special simplicity of BIAs learned in this gaze control experiment may be attributed to the selection of orthogonal interaction dimensions which simplified the situation tremendously.

SILI was not perfect in learning BIAs of gaze control. Mohammad and Nishida (2014b) reported two meaningless behaviors that were learned from the dataset as BIAs even though they did not correspond to any consistent interactive behavior. Fortunately, the interaction structure learner (see Sect. 9.5.2) was able to effectively remove these two BIAs by connecting them to no interaction processes higher level protocol layers. This ability to marginalize incorrectly discovered BIAs (to the level of having them consume no computational resources and affect no other part of the system) is an important advantage of SILI and reduces the pressure for selecting relevant interaction dimensions. This reveals an important aspect of SILI: it can without prior knowledge other than that which is implicit in its design learn meaningful human-like interaction capabilities that can be essential for agents that can provide a behavioral basis for attributing a ToM to them.

SILI cannot learn anything not available in the training set. For example, gaze aversion is a known human gazing behavior but this behavior was not learned because examples of it were not available in the training set. Nevertheless, SILI allows learning of protocols to advance incrementally. After learning basic gaze control using one training set, it is possible to enhance the learned protocol by combining it with another protocol learned from a different set. The dynamic integration of action commands implemented in SILI makes this combination possible.

Mohammad and Nishida (2014b) showed also that the Interaction Structure Learning (ISL) algorithm was able to learn two higher layers of control for the listener. The second layer of control consists of four processes. These processes corresponded directly to known human gaze behaviors during face to face interactions: gaze toward instructor, mutual gaze, mutual attention and a process that starts nodding when the instructor looks at the listener for more than 10 s. The final layer of control consisted of a single process that starts gaze toward instructor and alternates between it and the other processes based on the focus of attention of the instructor (Mohammad and Nishida 2014b).

10.1.6 Simulation Based Interaction

This section reports one of our experiments to compare the performance of SILI based gaze controller with the dynamic structure gaze controller developed using EICA (see Sect. 10.1.1).

The experiment was designed to see whether the proposed SILI controller will get higher subjective scores from third-party viewers of its behavior compared with both EICA and initial expectations of these viewers in terms of behavior naturalness and human-likeness and was reported by Mohammad and Nishida (2014b).

Evaluating SILI can be conducted at different levels. We can just look at the learned protocol in terms of connections between BIAs and try to elicit the underlying protocol in a rule-based human friendly format that can be compared with knowledge of human–human interaction protocols. This approach can give us insights on what is actually learned by the system but it is of little value for the actual goal of having a gaze controller in the first place: interacting with people.

Mohammad and Nishida (2014b) used a more empirical approach and implemented the learned protocol directly into a humanoid robot (Robovie II) then evaluated subjective perception of the behavior of this robot in a similar—yet not identical—task to the task used for learning the protocol. An important design decision for this kind of empirical evaluation is how similar should the task used for evaluation be compared to the task used for learning? Appropriately answering this question depends on the goal of the learning algorithm. SILI is not designed to learn a general interaction protocol that can be used in any context. We believe no such system can be built given the variability of human interactive behavior in different contexts. SILI on the other hand is designed to learn the interaction protocol of a specific situation and should be able to generalize the protocol over sufficiently similar situations but not very different ones.

The evaluation situation was designed to be similar to the training situation but not too similar that just a replay of motion or a simple modification of it could be sufficient. The evaluation situation consisted of an explanation session. In that it was similar to the training example. Nevertheless, the situation was made different by using a different explanation object (a Polymate device) that has a different number of components that are assembled in a very different way compared with the objects used for the training set and using different tools. Moreover, the disassembled parts

Fig. 10.3 Snapshots of explanation scenario evaluation experiment

of the device were put on a table between the instructor and the robot which was not available in the training set. This led to very different motions for the instructor. Moreover, the instructor did not actually assemble the device during the explanation but just pointed out how to do the task. Again this is in contrast with the training set (Fig. 10.3).

Another important decision concerning empirical evaluation of the interactive performance of agents and robots is the baseline performance we compare with. A simple approach is to compare the performance of the system with users' expectation. This is a useful comparison because—in the end of the day—conversational agents are built to interact with people and subjective evaluation of their performance is the most important factor in determining their success or failure. Being able to meet or go beyond users' expectations is a good sign for any conversational system. Nevertheless, this evaluation metric is highly dependent on many factors including the history of interaction of the evaluator human with conversational technology, novelty of the situation and many other factors.

A more useful statistic for the performance of the system can be found by comparing it with other successful conversational systems in the same situation. It is important to emphasize that comparison must be with a *successful* system in terms of the design goals of the conversational artifact or system being tested. SILI was designed to achieve human-like interactive behavior and the specific controller we are testing in this case study targeted gaze control. This means that we need compare with a successful gaze controller that was shown to achieve human-like behavior in an independent study. The dynamic structure EICA based controller introduced in Sect. 10.1.1 satisfied these conditions and was used as our base system.

Evaluation was conducted by an internet poll. Data from 49 subjects were used in the evaluation according to Mohammad and Nishida (2014b). The design was a within-subject design where each participant watched a video of the SILI controller performing on the Robovie II robot in the aforementioned task and another video of the EICA controller on the same robot and interacting with the same person. Details of the evaluation were given by Mohammad and Nishida (2014b). What is important for our purposes is that both EICA and SILI could outperform subject expectations. Moreover, SILI could achieve higher average scores (4.94) compared with EICA (4.49) in six dimensions of evaluation: attention, naturalness, understanding instructor's explanation, human-likeness, instructor's comfort, and complexity of underlying algorithms. Average score comparison between SILI and EICA controllers is shown in Fig. 10.4. For detailed analysis of the difference between the two controllers in specific evaluation dimensions and the correlations between these dimensions, refer to Mohammad and Nishida (2014b).

It is important to notice that the EICA controller was carefully designed using analysis of human–human gaze behavior while the design of the SILI controller outlined earlier relied mostly on unsupervised techniques and the only mental labor of the designer involved the selection of interaction dimensions and corresponding basic motor primitives. Despite this unsupervised nature, SILI was able to outperform EICA on three evaluation dimensions. This shows that SILI was able to capture the underlying interaction protocol at the same level—or slightly better—than what our

long analysis of human–human face to face interactions implemented in EICA was capable of achieving. The improvement of SILI over EICA is most significant in naturalness and comfort.

In the study reported by Mohammad et al. (2010), 30 subjects selected SILI controller as their preferred controller compared with 19 subjects for EICA. This again supports the superiority of SILI controller over EICA controller in terms of total score.

The dependence of gaze behavior (and interaction protocols in general) on the agent's theory-of-mind (ToM) is a major motivation behind the proposed approach which utilizes insights from the theory of simulation to organize the computational structure of the agent as shown in Fig. 9.3 and discussed in Chap. 9. Staudte and Crocker (2010) provided a supporting evidence for the importance of ToM for successful implementation of natural gaze behavior in robots. Two possible explanations were examined for the fact that listeners exploit speaker gaze to objects in a shared scene to ground referring expressions, not only during human–human interaction, but also in human–robot interaction. They compared an intentional account of this phenomenon to an attentional account and found that in a human–robot interaction settings similar to the one we use in this book the intentional account is more supported by the data. This means that gaze is not only utilized to convey information about the attentional state of the agent (robot/human) but is also interpreted by partners to infer the intentional state of the agent which obviously requires a form of ToM.

In summary, SILI was successful in learning a gaze controller for listening behavior in a human–robot dyadic interaction. We believe that this approach is general enough to extend to other nonverbal communication, to other gaze behaviors than listening, and to more than one interaction partner, but the demonstration of these is left for the future.

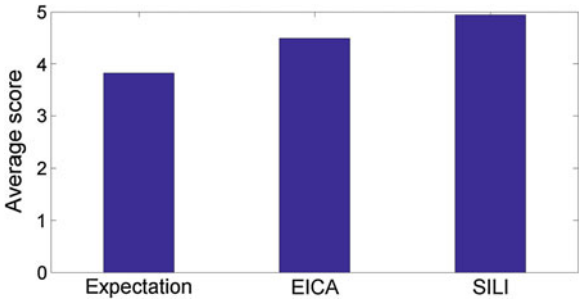


Fig. 10.4 Comparison between SILI and EICA controllers in terms of total score

10.2 Fluid Imitation: Imitation in Social Context

The main idea behind SILI is to utilize imitation for learning the computational processes of agents and a simulation-like behavior generation mechanism in order to achieve natural interaction and conversational behavior. A hidden assumption in the proposal we put forward until now in this chapter, is that the agent has a set of training interactions between humans that it can utilize to train SILI as was shown in the case study reported in the previous section. But we do not—usually—give our children markers specifying the beginning and ending of learning *episodes*; nor do they require much attention from us to *demonstrate* for them carefully how to interact with other people. Children seem to be able to just learn—imitatively—through *unintended demonstrations* that we provide for them continuously through our daily life. This is not to deny the importance of conscious deliberate demonstrations that we provide for them with marked episodic boundaries. We are just arguing that imitation in children is more *fluid* and can utilize information from unintended as well as intended demonstrations. Children can actually surprise us by how much they can learn without active involvement from us.

As it stands now; SILI uses a very *rigid* two-phases algorithm that we highlighted in Sect. 9.5. In this final section of the chapter, we outline a proposal for a *fluid imitation engine* that can enhance the ability of conversational agents to stimulate our anthropomorphic responses by having a less rigid child-like imitation capacity. Notice that the simulation aspect of SILI's behavior generation needs not be affected or changed in order to achieve that. The case-study presented in this section is based on the work presented earlier by Mohammad and Nishida (2012a).

Section 9.2 discussed the main challenges facing the realization of effective imitative robots that can learn through human demonstrations. One of the major differences between imitation based learning and traditional supervised learning, is the availability of a limited number of training examples for the learner. This limits the applicability of traditional machine learning approaches like SVMs and BNs. Another major difference—that is usually ignored in LfD research—is that in real world LfD situations, the learner may have to detect for itself what behaviors it needs to learn as the imitatee may not be always explicit in marking the boundaries of these behaviors or the dimensions of the input space that are of interest for learning.

Most of the research in imitation learning has focused on the perspective taking, imitatee modeling and the correspondence problems above (Argall et al. 2009). A problem that is usually ignored in most cases is discovery what actions to imitate in the first place. These are usually selected by the researcher or the user and the robot is assumed to be in some *imitation* mode during learning then is explicitly put outside this mode during production of learned behavior. This is not how infants or people in general learn. We decide what to imitate as much as we decide how to do the imitation. Solving this action segmentation problem is the key to achieve fluid imitation. This means that the conversational agent should be able to discover the significance of each behavior it perceives in order to decide which behaviors to imitate (Mohammad and Nishida 2012a).

There are many factors that affect the importance of a behavior for the learner. If we some behavior happening again and again, we tend to assume that it has some value. Learning agents can also rely on behavior repetition for judging the importance of different behaviors. On the other hand, novelty can also bias the agent toward learning the behavior in some contexts. The importance of any behavior does not stem only from its ubiquity, novelty, or any other factors intrinsic to the behavior. Behavior importance for the learner depends on many other factors as well. One such factor is whether or not the behavior can be replicated by the learner in the first place. Goals of the learner also dictate to some degree the importance of behaviors for that learner. For example, a robot designed to act as a cook should not be very interested in imitating aerobic exercises it sees some people performing but on activities related to cooking. The place and timing of behavior may also influence its importance for the learner. Again a cook robot may be interested in activities in the kitchen or just before a meal is served much more than other places and times.

Fluid imitation is possible only if all of these factors affecting importance can be combined easily in deciding the boundaries of actions to be imitated.

Notice that—in a sense—fluid imitation is a general problem that is not restricted to conversational agents which is the focus of this book. This is good news because it allows us to study it without requiring a full implementation of a conversational agent. Moreover—given that it is a novel problem (Mohammad and Nishida 2012a)—it is even beneficial to start its treatment in a domain that is much simpler than conversational informatics and for the sake of this chapter we will consider the much easier navigation domain. This allows us to focus on the fluidity of imitation without being lost in the details of conversations or even object manipulations.

The main idea we explore in this very preliminary treatment is to define the fluid imitation problem as a special case of the well-defined constrained motif discovery problem in data mining (see Sect. 4.5). This approach allows us to naturally integrate all factors affecting the importance of some behavior for the learner.

As this is a young problem that is defined—for our best knowledge—in 2012, we do not have a lot of history to explain here. We will then focus our attention in providing a brief treatment of a recently proposed fluid imitation engine that was tested by a series of experiments in learning navigation behaviors (e.g., object triggered motion patterns, and intrinsically triggered motion patterns) (Mohammad and Nishida 2012a).

The main problem in fluid imitation can be stated as follows: *Given a continuous stream of actions from the imitatee, find the boundaries of important behaviors for imitation subject to a predefined set of goals and abilities for the learner.*

This characterization covers both learning motion primitives from trajectories of motion (or other behavioral signals) as well as learning more complex plans of motor primitives from the activation levels of these primitives as perceived by the learner (Mohammad and Nishida 2012a).

Mohammad and Nishida (2012a) used navigation in a 2D space as a running example in order to test a concrete implementation of the idea. In this section, we will utilize a hypothetical example involving a cook robot watching behaviors of different people inside and outside a kitchen. This discussion is strictly theoretical

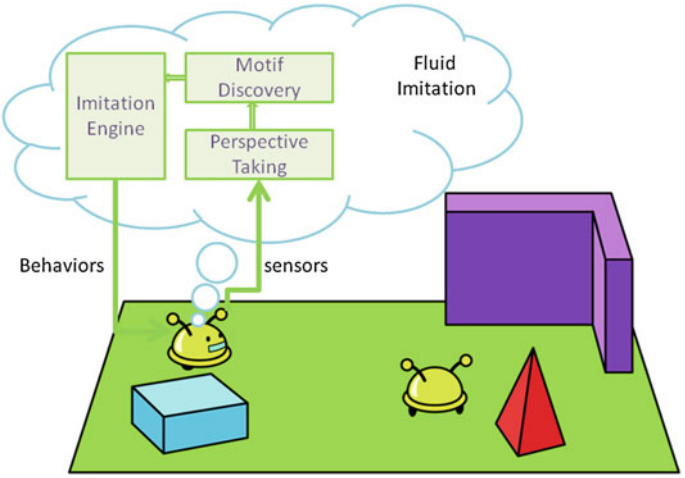


Fig. 10.5 The simplest fluid imitation system. The learner robot observes the behavior of other robots interacting with the environment (or each other), passes this information after perspective taking to a motif discovery system that generates sets of behaviors to be learned. These behaviors can then be learned using a conventional imitation engine and used to control the robot. © 2014, Yasser Mohammad and At, Inc. 2014. Reproduced with permission

but we will follow it by some results given by Mohammad and Nishida (2012a) to show that these ideas are in face applicable to real world—albeit simpler—situations.

The simplest possible fluid imitation engine is shown in Fig. 10.5. The learner robot observes the behavior of other agents (either humans or robots) interacting with the environment (or each other), passes this information after perspective taking to a motif discovery system that generates sets of behaviors to be learned. These behaviors can then be learned using a conventional imitation engine and used to control the robot. Motif discovery was first discussed in Sect. 4.5 and was utilized in the design of SILI in Chap. 9. This system can be used in simple cases as in navigation learning to learn basic concepts related to the target behavior and basic motions that can be used to navigate the environment safely (Mohammad and Nishida 2013a).

What is missing from this system? A principle for structuring the learning process. The system learns too much in a sense. It tries to find patterns in everything around it which is a good idea for a beginning but it is not constructive in the sense that there is no goals that structure the exploration of others' behavior. Consider a robot watching what is happening in the kitchen in order to learn how to cook scrambled eggs. It would not be very effective to use the simple system of Fig. 10.5 in this case because many actions that are performed by people in the kitchen vicinity will not be even related to this task.

There are several ways to structure the learning process of the system. One particularly simple way is to rely on object saliency. Consider a robot trying to learn nonverbal conversational behavior (the main focus of this book). The robot needs not imitate all kinds of behaviors executed by humans in its vicinity. It can look

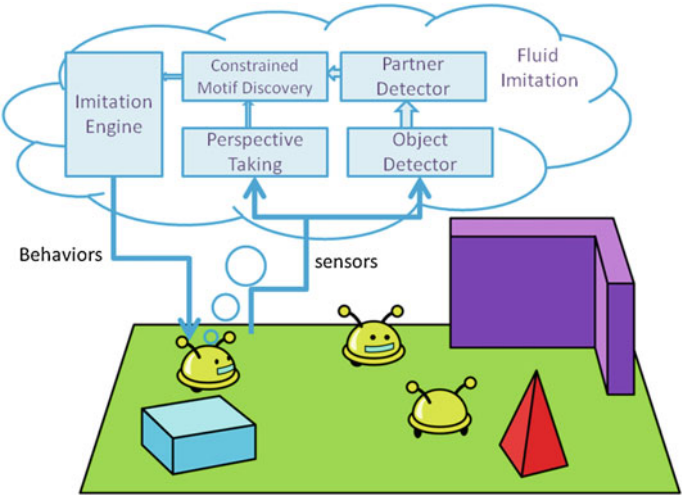


Fig. 10.6 An interaction oriented fluid imitation engine. The robot detects objects of interest in the environment and uses their saliency as a constraint for its segmentation/motif discovery subsystem. © 2014, Yasser Mohammad and At, Inc. 2014. Reproduced with permission

for salient objects (humans in this case) and only learn behavior near them (interactive behavior mostly). Figure 10.6 shows a simplified version of the fluid imitation engine of such a robot. Here we take advantage of the constrained motif discovery algorithms' ability to utilize constraints on motif locations (see Sect. 4.5) and provide these constrains from the saliency detector. Saliency is a very important cue for human learning, for example, it is well-known that care-givers use very specific types of motion when teaching their children that they do not use when interacting with adults. These special motion patterns (called motionese) where shown to increase the saliency of objects and behavior by moving the first and exaggerating the later (Nagai and Rohlfing 2007).

We can organize our discussion of more complex behaviors by dividing the outputs of the learner's sensors into different *streams* that correspond to different important features of the perceived scene. We use the typology we developed in previous work (Mohammad and Nishida 2012a). We assume that the engine receives three streams of information (all in the form of multidimensional time-series): imitatee-action, imitatee-perception, and objects streams. The *imitatee-action* stream represents the behavior of the imitatee. The *imitatee-perception* stream represents the *objects* stream as would be perceived if the learner was in the location and external configuration of the imitatee (this stream is generated from perspective taking processes). The *objects* stream represents the state of all objects in the environment (we assume that objects are already separated from the background either by pre-determining them or using object recognizers and saliency maps (Mohammad et al. 2009)). The engine generates a single output which is the set of behaviors from the *imitatee-action* stream that are to be imitated by the robot.

If we increase the complexity of the situation slightly by having some interaction between the imitatee we are learning from and other objects (or other people) in the environment. *Objects* and *imitatee-perception* streams start to play a more significant role in the system. In this case, *behavior-significance* depends not only on change in the *imitatee-action* stream but also on changes in the *objects* and *imitatee-perception* streams that are *causally* related to the *imitatee-action* stream. It is possible in this case to utilize a causality graph induction method that allows us to detect these parts of the *imitatee-action* stream that are causally related to the changes in the other two streams (Mohammad and Nishida 2010b).

In the general case, the robot will have a high-level cognitive component that is responsible of setting robot's goals (which directs the relevance calculation of objects and behaviors) and specific saliency features (e.g., specific colors, textures or change patterns) for the three input streams. For example, a robot learning cooking will assign high relevance to any behavior happening in the kitchen or any behavior that changes the state of the oven to *on*. In this case, the complete system depicted in Fig. 10.7 will be utilized with saliency and significance estimators that are now utilizing this input from higher cognition.

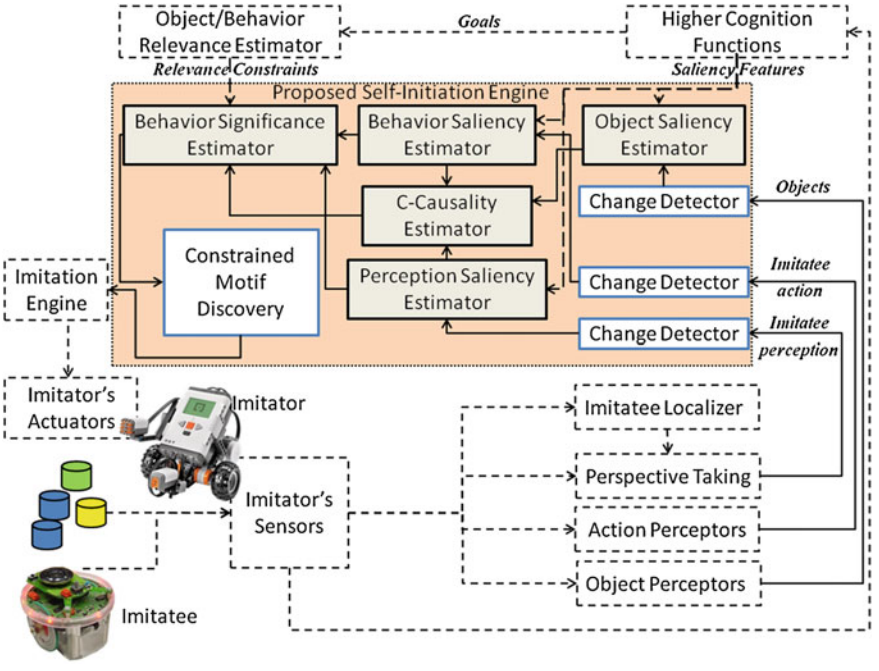


Fig. 10.7 Overview of the fluid imitation engine and its relation to other components of the robot. The components and signals that interact with the engine but are not a part of it are shown in *dash-lines*. The optional parts of the engine are shown with *shaded background* (Mohammad and Nishida 2012a). © 2012, Springer. Reproduced with permission

The main advantage of this proposed system is that it can merge information from low level signal processing (change detection) and high level cognition (relevance) easily in the form of constraints that are passed to the constrained motif discovery algorithm to decide what actions to imitate. This makes it easier to extend the system to take into account any form of saliency features or any complex relevance measure while still being grounded in the low level sensor-driven motif discovery process.

Consider a robot that tries to learn how to open a box (similar to the jar example used by Breazeal and Scassellati (2002)). The robot watches a continuous stream of actions from the human subject but it should not care about most of them. Only few of these actions are interesting for its goals and these are the actions that *cause a change in the state of the box*. The purpose of change causality estimation in our system is to find what are the salient parts of the *imitatee-action* stream (discovered by the change detector and probably augmented by high level saliency features) that are *causing* relevant or salient changes in the *objects* and *imitatee-perception*) streams. There are several causality tests in the literature that can be used here and we just outline one possible alternative here called the change-causality test (Mohammad and Nishida 2011a).

The change causality test utilizes the outputs of the change detectors (optionally adjusted using high level saliency features) as its inputs. The time series corresponding to the *objects* and *imitatee-perception* streams are called P_j while the ones corresponding to the *imitatee-action* stream are called A_i for some positive integers i and j. The main assumption is that if A_i causes P_j then P_j will *most of the time if not always* has major changes near τ_{ij} time-steps after A_i where τ_{ij} is a constant representing the delay of the causation. Because in the real world, many factors will affect the actual delay (add to this inaccuracies in the change point detection algorithm), it is expected that in reality the delays between these change points will be well approximated with a Gaussian distribution with a mean of $\hat{\tau_{ij}}$ where $\left|\hat{\tau_{ij}} - \tau_{ij}\right| < \varepsilon$ for some small value ε. The main idea of the algorithm is to check for the consistency of the delays and to use the resulting statistic as a measure of causality between the two processes.

To make our discussion more concrete; Mohammad and Nishida (2012a) report a series of proof-of-concept experiments with increasing complexity aiming at showing the applicability of this approach. The task used for all of the experiments was robot navigation. All simulations were done using the V-REP simulator[1] with realistic physics based on the Bullet Physics engine.

10.2.1 Self-Initiated Behavior

In the first scenario we consider, the imitatee is a simulated e-puck robot that moves around an empty arena (no *objects* or *imitatee-perception* streams). There is no high-level cognition component involved and no goals preset for the learner. In such

[1] http://www.v-rep.eu/.

a scenario, there is no clear right or wrong answer for the question: *what should I imitate?* Nevertheless, analyzing the input streams (consisting of 3D state information of the imitatee), would allow the learner to separate recurring patterns embedded in the behavior of the imitatee.

The imitatee moves randomly around the arena but every few seconds it selects one of three predefined models (square, triangle, circle) and executes it. As no high-level cognition is available in this problem, all parts of the system affected by this module where inactive. The same is true for the *objects* and *imitatee-perception* streams. This leaves us with only five active components that we will explain in the following paragraphs.

The *action-perceptors* received a 3D time-series representing the location and orientation of the imitatee over time $(x(t), y(t), \theta(t))$ and outputs a one dimensional time series $(\hat{l}(t))$ representing the location of the imitatee in the arena. To calculate this representation, the first order difference of the inputs $(\hat{x}(t), \hat{y}(t), \hat{\theta}(t))$ is calculated where for any signal $(z(t))$, $\hat{z}(t) = z(t) - (t-1)$. This first order difference represents the local behavior of the robot independent of its location in the global frame of reference.

The second active component is change detection. The input to this component is the time series $(\hat{l}(t))$ and it applies the change detection algorithm discussed in Sect. 9.5.1 to them resulting in one output time series representing change scores $(\hat{c}_l(t))$.

The behavior saliency estimator just passes its two inputs from the change detectors to its outputs as it has no other sources of information about saliency. The behavior significance estimator simply applies either a summation or multiplication operation on all its inputs at every time step under the control of the higher level cognition module. In this experiment because it has a single input it just passes its single input to the output and the one dimensional signal $(\hat{c}_l(t))$ is passed as the constraints to the CMD component.

The CMD component, receives the signal $(\hat{l}(t))$ as well as its corresponding constraint $(\hat{c}_l(t))$ and generates a list of motifs and motif occurrences using the CMD algorithm which—as the reader may already notice—lies at the heart of our approach to autonomous learning in general. The motif occurrence locations are then passed to the simulation engine as the locations at which imitation should occur. The details of the imitation engine itself will vary from an agent to another and is not crucial for our discussion here. For this reason we just skip it here and commit ourselves to no specific imitation approach.

The algorithm required on average 3.246 ms/point of input with a standard deviation of 1.30499 ms/point in a Core 2 Due T9600 machine with 4 GB of memory and with a MATLAB implementation. This means that the learning time required by the learner is around one tenth of the imitatee's execution time. Figure 10.8 shows an example motion session with the patterns discovered by the system.

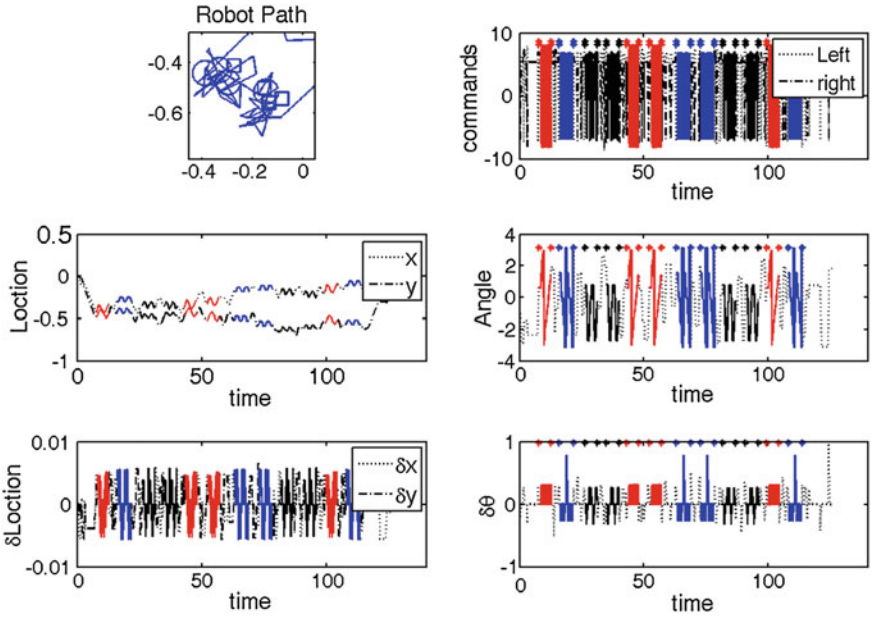


Fig. 10.8 Example motion with the corresponding patterns in different dimensions. In this session, the learned model was exactly matching the ground truth (Mohammad and Nishida 2012a). © 2012, Springer. Reproduced with permission

10.2.2 Object-Caused Behavior

A more interesting situation arises when there are objects in the environment that modify the behavior of the imitatee. Figure 10.9 shows the simulation environment after adding two types of object: cylinders about which the imitatee moves in squares and *K*-Junior robots around which it moves in circles. Proximity is measured using simulated robot's infrared sensors. The role of the learner now is not only to find recurrent patterns in the motion of the imitatee (as was the case in the first experiment) but to know the relation between these patterns and different types of objects in the environment.

Four new components are now active because of the activation of the input *imitatee-perception* stream.

The most important component introduced in this settings is the change causality estimator which detects possible causal relations between the two input streams.

Here, the object significance estimator becomes a little bit more complicated as it does not add all its inputs without discrimination but it now multiplies the constraints from the *imitatee-perception* stream that are shown to be causally connected to changes in the *imitatee-action* stream by the constraints of the later stream after appropriate delay (found by the change-causality estimator). This way,

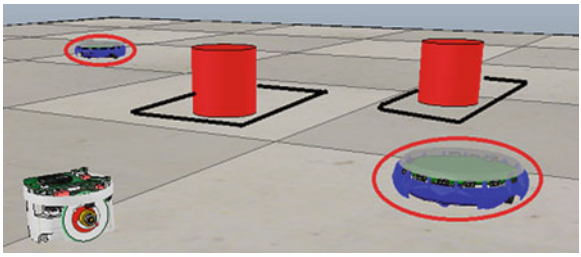


Fig. 10.9 The robot and its environment during one fluid-imitation evaluation experiment. Two types of objects exist in the enjoinment: *K*-Junior robots and cylinders. The robot should learn to move in *squares* around the *cylinders* and *circles* around the *K*-Junior robots (Mohammad and Nishida 2012a). © 2012, Springer. Reproduced with permission

the learner will only care about behaviors that happen when the perception of the imitatee changes (i.e., near objects in the environment in this experiment).

10.2.3 Relevance-Informed Learning

The last experiment reported by Mohammad and Nishida (2012a) involved making the *K*-Junior robots rotate continuously in place until the imitatee rotates around them.

A very simple high-level component was added to the system that gave the imitatee a single goal of stopping *K*-Junior robots from rotating. This was translated into two signals: A behavior saliency feature that increased the saliency score of the behavior based on the distance between the imitatee and its nearest *K*-Junior robot and a relevance feature that increased the significance score of the imitatee's behavior if within 1 min of it any *K*-Junior robot stopped rotating. The *objects* stream is now active giving the rotational state of the *K*-Junior robots and their locations. The complete system of Fig. 10.7 is now utilized.

10.3 Summary

This chapter reported two case studies that focus on the two main concepts of Chap. 9. The first case study concerned learning to control the eye-gaze using the SILI architecture proposed in the previous chapter and showed that this approach can outperform a carefully designed gaze controllers and user expectations. These encouraging results suggest that utilizing simulation theory for behavior generation combined with imitation for computational process learning (as in SILI) may provide future artificial conversational agents with the ability to autonomously learn subtle nonverbal behaviors that are crucial for natural interaction with humans. These competencies may

also lead to more social predictability for these agents which in turn will enhance their ability to engage in artificial empathy with human partners.

The second case study provided a preliminary treatment of the fluid imitation problem which was recently proposed as a more human-oriented form of learning from demonstration. Robots that can solve the fluid imitation problem will be able to learn throughout their lives not only from explicit teaching but also be just watching what humans do in their environment more like children. Fluid imitation uses the same building blocks of SILI (namely, motif discovery, change point discovery and causality analysis) but it does not utilize the simulation theoretic behavior generation mechanism because it is not directly targeting learning interaction protocols but it tries to leverage the ability to interact with people which SILI and similar architectures can provide in order to improve the robot's capacities to act in its own environment.

Chapter 11
Cognitive Design for Discussion Support

Abstract Conversation is not only a joint activity in itself, but also a means for joint activity. Discussions can benefit from augmented conversation in stimulating, editing, disseminating, and reusing conversations. Conversational intelligence empowered by conversational agents allows wisdom exchanged in conversation to be shared and evolved in a community. Cognitive design allows for deeper support for discussions by allowing conversational agents to sense social signals, estimating the tacit intentions of discussants, and even leading discussions in a direction potentially preferred by the discussants. In this chapter, we present some pilot studies on grounding conversational support at the cognitive level.

Keywords Decision making · Human-agent interaction · Cognitive design · Facilitation · Joint intention · Interaction design for ECA · Interaction analysis in HAI

11.1 Cognitive Framework for Cognitive Support

A large collection of previous work exists on group facilitation and supporting group decision making using information systems. Among others, Stefik et al. (1987) considered how a computer system can contribute to human discussion for problem-solving, while Niederman et al. (1993) investigated supporting remote discussions using an information system. Clawson and Bostrom (1993) proposed a group support system with automated facilitation. Limayem (2006) investigated the tradeoffs for both human and automated facilitation. However, all these works focused on the decision making phase of group discussion and did not address the individual behaviors of participants or the understanding phase.

In contrast, far less work on facilitation has been published (Reagan-Cirincione 1994). As it has been suggested that facilitation highlights the group's social and cognitive processes allowing participants to focus on more substantive issues in the decision making process (Schuman 1996) and ultimately find the most appropriate

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_11

solution to a problem (Khalifa et al. 2002), shedding light on facilitation would be a worthwhile research topic.

Our research study set up to unveil the facilitation process, involves detailed observation of how an experienced facilitator guides the discussion. It appears that experienced facilitators appropriately interpose based on the most important arguments of the participants, as inferred from their verbal and nonverbal behavior, although it is not known how a good facilitator determines the facilitating behavior or on what kind of information this is based. We focused on the nonverbal and paralinguistic behavior of participants in the context of a face-to-face discussion to which a good facilitator pays close attention during the discussion.

In the rest of this chapter, we assume that the facilitator is skilled in detecting and producing social signals that are key to guiding the discussion. We discuss a method for singling out the most important facilitating actions that expert facilitators apply to guide the discussion, identifying key social signals that they appear to leverage in facilitation, building a model describing how facilitating actions are determined from key social signals, and building a facilitative agent that can reproduce facilitating behavior based on the facilitation model (Fig. 11.1).

We assume that each participant, either a discussant or the facilitator, has a particular intention, which is often vague or even inconsistent particularly at the beginning of a discussion. The intention is dynamic in nature in the sense that the content and structure thereof changes, often inconsistently, during the period of the discussion. It is reasonable that both the individual intention of a participant and the joint intention shared by the group may appear from time to time during the discussion and gradually be shaped in a rather heuristic fashion of building and scrapping (Fig. 11.2). Our facilitation model should be able to deal with the vagueness and dynamism of intentions in nature.

We employ a preference structure to characterize how a decision is associated with various aspects and factors. A decision is the choice that a group of discussant may make as a result of the discussion. For simplicity, we assume that decisions are discretized and that the decision space comprising the possible decision is finite, for example, consisting of 100 entities. The aspect is an abstract conceptual feature that participants share in characterizing each decision, e.g., “being savvy”. We assume

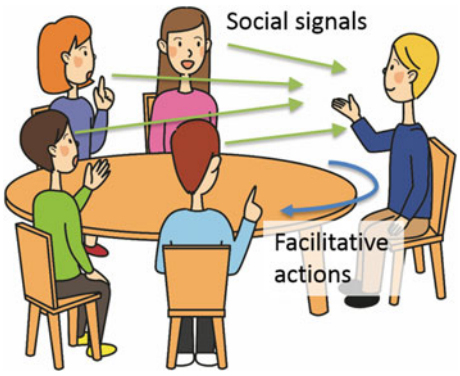


Fig. 11.1 The facilitating model. © 2014, Yoshimasa Ohmoto and At, Inc. Reproduced with permission

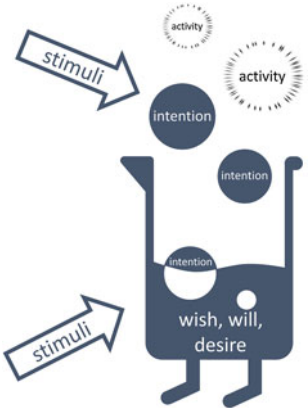


Fig. 11.2 Bubbling intention. © 2014, Yoshimasa Ohmoto. Reproduced with permission

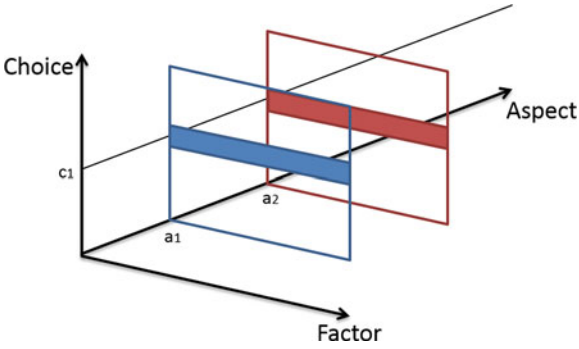


Fig. 11.3 Preference structure. © 2014, Yoshimasa Ohmoto. Reproduced with permission

that participants mostly, but not completely, share the definition of aspect. In contrast, a factor is a concrete feature specifying a property that can be judged in a rather objective fashion, and hence can be assumed to be shared by all participants.

We represent the structure of potential decisions as a three-dimensional array as shown in Fig. 11.3. For each choice c_i, the corresponding aspect a_j is represented as a vector of factors $(f_{ij1}, f_{ij2}, \ldots, f_{ijn})$, where f_{ijk} denotes how the k-th factor contributes to a_j in choice c_i.

11.2 Analysis of Facilitating Behavior

What are the key facilitating actions that expert facilitators apply in guiding the discussion to obtain consensus among individuals in our social activities? To answer this question, we collected relevant data, which were analyzed and interpreted to understand and model the behavior of an experienced facilitator.

11.2.1 Data Collection

In order not to miss subtle social signals in facilitation, we conducted four preliminary experiments under different discussion conditions, for example, varying the number of participants, the subject of the discussion, and the facilitators. From the results of the preliminary experiments, we hypothesized four basic facilitating behaviors:

Diverging: the facilitator interposed to encourage divergent discussion. For example, the facilitator asked the whole group for new opinions.

Converging: the facilitator interposed to encourage convergent discussion. For example, the facilitator listed the opinions expressed up to that point, and then confirmed which participants were in agreement and which had expressed conflicting opinions.

Conflicting person: the facilitator asked a member of the group with a conflicting opinion to speak. We did not consider whether the opinion was divergent or convergent. There were two reasons for this: one was that we could not distinguish whether the facilitator requested a divergent or convergent opinion, the other was that the facilitator told us that he did not consider whether it was a divergent or convergent opinion.

Objectifying: the facilitator asked the last speaker to objectify his/her opinion. In the experiment, this facilitating behavior occurred infrequently.

Here a more comprehensive explanation of the data collection process is given. Five discussion groups or *D-groups* were formed with each D-group consisting of four undergraduate students as discussants and one experienced facilitator. Although we selected the discussants so that no-one participated in more than one D-group,

Fig. 11.4 Experimental setting to observe the discussion. © 2014, Yoshimasa Ohmoto. Reproduced with permission

Fig. 11.5 Examples of the recorded video in the discussion. **a** The whole of the group, **b** the right side of the facilitator, **c** the left side of the facilitator, **d** the facilitator face. © 2014, Yoshimasa Ohmoto. Reproduced with permission

we employed one experienced facilitator for all five D-groups. We selected the best facilitator with sufficient practical experience of the three available candidates.

Figure 11.4 illustrates the experimental setting we analyzed (Ohmoto et al. 2011). In this section, we introduce the investigation briefly. There was no table and the discussants could not take notes during the discussion. As the topic of the discussion, we asked participants of each D-group to agree on a plan for an overnight trip for the next summer. This particular subject was chosen, as all participating undergraduate students had experience in the task and the subject involved some interrelated topics that were expected to cause multiple divergent and convergent changes in the discussion and varying states of confliction and agreement among the discussants. In fact, the subject had many interrelated topics, e.g., "whether the purpose of the trip was to engage in energetic activities, such as visiting many tourist spots, or relaxation, such as sunbathing", and "whether the destination of the trip should be urban or rural".

We collected head orientation, voice, and video data during the experiment. We did not, however, measure facial expressions and postures, as these differed greatly among the individuals without any consistency in the variations. We used a motion capture system (Motion Analysis Corp., MAC3D) to measure the head orientation as an approximation of the gaze direction. We did not measure gaze direction, as we

Table 11.1 Average ratings of facilitators based on questionnaires

	Good	No-good
The facilitator controlled chances to talk	7.0	4.0
The facilitator advised appropriate points of the discussion	6.6	5.8
The facilitator interposed in the discussion at the right time	7.1	4.4
The facilitator was good for the discussion	7.1	5.1

did not wish to disrupt discussants by using appropriate devices for measuring gaze direction. We recorded audio by means of throat microphones worn by each D-group member. Four video cameras were used to record the facilitator, two discussants on the right side of the facilitator, the other two discussants on the left side of the facilitator, and the whole D-group, as illustrated in Fig. 11.5.

Participants were allowed up to 30 min for their discussion (mean = 22 min). The facilitator was given instructions to conclude within 30 min and to encourage the discussants as much as possible. He was given no other instructions and allowed free facilitation. Facilitating behavior was employed 154 times throughout all the discussions (mean = 30.8).

After the discussion, the discussants were asked to complete a questionnaire to confirm whether the facilitator's behavior was good. The questionnaire focused on the following four behaviors: "the facilitator allowed the participants a chance to talk," "the facilitator offered advice at appropriate points in the discussion", "the facilitator interposed in the discussion at the right time," and "the facilitator served to enhance the discussion". The discussants (20 students) rated each behavior of the "good" facilitator on a scale from one to eight. In the preliminary experiments which were conducted under the same experimental settings, the discussants (8 students) also rated the behavior of the facilitator, who was not that good ("no-good"). Table 11.1 shows the results of both experiments.

The average scores of the facilitator who participated in the actual experiment are higher than those of the facilitator in the preliminary experiment, who was not regarded as being very good. Therefore, the facilitator who participated in the actual experiment was a good facilitator in this situation and we considered his facilitating behavior to be one of the best.

11.2.2 Data Analysis

Data analysis was done in three stages. We first defined facilitating behaviors in terms of the behavior of the facilitator in conducting a smooth discussion, such as encouraging discussion, organizing the discussion topics, confirming opinions between participants, and so on. We then segmented the videos into separate clips containing one complete session of a single facilitating behavior.

Table 11.2 Concordance rates of the two coders

	The number	Concordance rate (%)
Which types of the facilitating behaviors	142	92.2
Whether the discussion was diverging or converging	152	98.7
Which participants were conflicting or agreeing	148	96.1

Table 11.3 Frequency of facilitating behaviors in the collected data

Diverging	59
Converging	37
Conflicting person	40
Objectifying	18
Total	154

In the second stage, we labeled each video clip with the type of facilitating behavior, whether the discussion was divergent or convergent, and which participants were in conflict or in agreement. We chose diverging/converging as an explanatory variable, as the discussion condition generally alternated between diverging and converging in reaching a conclusion. We chose conflicting/agreeing as another explanatory variable, since the facilitative behavior appears to depend on identifying which participants are in conflict and which are in agreement in reaching consensus. The two variables denote the context of the discussion and arguments of each participant.

We employed two coders, the researcher and a person very familiar with video annotation (hereafter referred to as the "reference coder") for annotation. We compared the annotations of the two coders to confirm the reliability thereof. Table 11.2 gives the results where 92 % of the facilitating behavior labels, 99 % of the diverging/converging labels, and 96 % of the conflicting/agreeing labels matched. This confirmed the reliability of the annotations. Whenever the two coders disagreed, they resolved their differences through discussion.

Table 11.3 shows the frequency of the facilitating behaviors in the collected data. Linear discriminant analysis was applied to the data, the following explanatory variables of which were elicited from the annotated data and the data for head orientations and voices in the experiment in the ten seconds immediately prior to the facilitating behavior

Annotated data for "diverging or converging": The variable for "whether the discussion was diverging or converging" was set to 1 if the corresponding annotation was "diverging". Otherwise, it was set to 0.

Total time paying attention to the facilitator: The discussants paid attention to the facilitator when they wanted to say something in the discussion. This variable denotes the total time that the four discussants watched the facilitator.

Total speaking time: The amount of speaking time indicates how active the discussion was. This variable denotes the total speaking time of the four discussants.

Difference in attention between discussants: This variable, which indicates which discussant was leading the discussion, is calculated as follows. First, we calculate the amount of time each discussant was watched by the facilitator. Then, we subtract the average time from the longest time, and assign the result to the variable. The value of this variable increases if a single participant attracts more attention.

Difference in speaking time between the discussants: This variable is key in selecting the person to whom the facilitator should speak and is calculated as follows. We sum the amount of time each discussant spoke, subtract the average time from the longest time, and assign the result to the variable. The value of this variable increases if a single participant does most of the talking.

Difference in attention between groups in agreement: The difference in attention between agreeing groups with a similar opinion, indicates which of these groups was leading the discussion. This variable is calculated as follows. We calculate the amount of time each group was watched, subtract the average time from the longest time, and assign the result to the variable. The value of this variable increases if one group attracts more attention.

Difference in speaking time between groups in agreement: This metric also indicates which of the agreeing groups was leading the discussion. This variable is calculated as follows. We calculate the total speaking time for each group, subtract the average time from the longest time, and assign the result to the variable. The value of this variable increases if one group does most of the talking.

The aim of linear discriminant analysis is to use explanatory variables to classify the facilitating behaviors into two groups: those matching the facilitating behavior types and the others. The results of the discriminant analysis are given in Table 11.4. The numbers of facilitating behaviors correctly classified by the discriminant analysis are listed in the "Number of correct classified F.B." row, while the results of the discriminant analysis are listed in the "Discriminant ratio" row. The numbers of facilitating behaviors correctly classified by cross validation are listed in the "Number of C.V." row with the ratios of the cross validation listed in the "Ratio of C.V." row.

We also applied each discriminant function to the explanatory variables and then selected the facilitating behavior type for which the discriminant function returned the highest value as the appropriate behavior. The concordance rate between the selected behavior and the actual behavior is 78.6 %. Thus, the explanatory variables can correctly classify facilitating behavior.

We analyzed the contribution of the explanatory variables in classifying the facilitating behavior using structure matrices. The higher the value of the structure matrix is, the more likely it is that the target facilitating behavior is classified as the facilitating behavior.

11.2.2.1 Diverging

Table 11.5 gives the values of the structure matrix for the discriminant function of the facilitating behavior "diverging". The highest value is the explanatory variable for the annotated data of "diverging or converging". This means that the facilitating behavior of "diverging" is appropriate when the discussion is diverging. We confirmed that the facilitator encouraged divergent discussion in many cases from the recorded video of the experiment. In addition, the total speaking time is low. This means that the facilitator often encouraged divergent discussion if the discussants did not volunteer opinions. Moreover, the values of the variables related to the differences between the agreeing groups are low. This means that the facilitator often elicited a variety of options before agreeing groups had formed.

11.2.2.2 Converging

Table 11.6 gives the values of the structure matrix for the discriminant function of the facilitating behavior "converging". The lowest value is the explanatory variable for the annotated data of "diverging or converging". This indicates that the facilitating behavior of "converging" is appropriate when the discussion is converging. In addition, the total speaking time is high, but the difference in speaking time between the agreeing groups is low. This means that the facilitator often encouraged convergent discussion when the discussants were engaged in active discussion.

Table 11.4 Results of four discriminant analysis of each facilitating behavior

	Diverging	Converging	Conflicting	Objectifying
Number of correct classified F.B.	133	129	122	133
Discriminant ratio (%)	86.4	83.8	79.2	86.4
Number of C.V.	132	125	119	132
Ratio of C.V. (%)	85.7	81.2	77.3	85.7

Table 11.5 Structure matrix for discriminant function of facilitating behavior "diverging"

Annotated data of "diverging or converging"	−0.831
The total time to pay attention to the facilitator	−0.121
The total time of speaking	−0.240
The difference of attention between the discussers	−0.076
The difference of speaking time between the discussers	0.080
The difference of attention between the agreeing groups	−0.335
The difference of speaking time between the agreeing groups	−0.313

11.2.2.3 Conflicting Person

Table 11.7 gives the values of the structure matrix for the discriminant function of the facilitating behavior "conflicting person". The highest values are for variables related to the differences between agreeing groups. This means that the facilitator often asked a discussant with a conflicting opinion to speak when one group was leading the discussion. In addition, the difference in speaking time between the discussants is low. This means that the discussion was actively conducted between the groups. The total time the discussants paid attention to the facilitator is also high. This means that the discussants often watched the facilitator in order to say something in the discussion when one group was leading the discussion.

11.2.2.4 Objectifying

Table 11.8 gives the values of the structure matrix for the discriminant function of the facilitating behavior "Objectifying". The highest values are those for variables related to the differences between agreeing groups. This means that the facilitator asked the last speaker to objectify his/her opinion when one group was leading the discussion. In addition, the difference in speaking time between the discussants is high whereas the total speaking time is low. This means that a particular discussant did most of the talking and there were no other opinions.

Table 11.6 Structure matrix for discriminant function of facilitating behavior "converging"

Annotated data of "diverging or converging"	−0.642
The total time to pay attention to the facilitator	−0.044
The total time of speaking	0.322
The difference of attention between the discussers	−0.031
The difference of speaking time between the discussers	0.050
The difference of attention between the agreeing groups	−0.160
The difference of speaking time between the agreeing groups	−0.372

Table 11.7 Structure matrix for discriminant function of facilitating behavior "conflicting person"

Annotated data of "diverging or converging"	−0.224
The total time to pay attention to the facilitator	0.293
The total time of speaking	0.254
The difference of attention between the discussers	−0.065
The difference of speaking time between the discussers	−0.338
The difference of attention between the agreeing groups	−0.475
The difference of speaking time between the agreeing groups	0.619

11.2.3 Insights Obtained

The facilitator who participated in the experiment tended to encourage divergent discussion when the discussion was diverging and to encourage convergent discussion when the discussion was converging. In addition, he tended to encourage divergent discussion when the total speaking time was short and to encourage convergent discussion when the total speaking time was long. These interpositions played a role in controlling the number of opinions so that all participants could comprehend most of the opinions in the discussion.

On the other hand, the facilitator asked a discussant who had a conflicting opinion or the last speaker to objectify his/her opinion if the discussion was converging. The facilitator told us that he did not consider whether the discussion was diverging or converging. In addition, the facilitator often asked a discussant with a conflicting opinion in an active discussion between the groups or the last speaker in a discussion that was reaching consensus to objectify his/her opinion. This means that the good facilitator arranged the discussion as follows. An opinion was not accepted as concrete content while the discussants were producing new or conflicting opinions. After most of the opinions had been placed on the table, each opinion was carefully examined. Therefore, we suggest that the facilitator controlled the discussions effectively and smoothly through appropriate facilitating behaviors.

Moreover, the total time discussants paid attention to the facilitator was long when a single group led the discussion and the discussants engaged in active discussion. In other words, by watching the facilitator, discussants sent him a message stating that they wished to say something. This is one of the nonverbal behaviors that is useful in encouraging effective and smooth discourse.

An open problem for future research is to analyze the facilitating behaviors of other good facilitators in other types of discourse to reveal to which discussants' behavior the facilitator pays attention, when the facilitator interposes, and how the facilitator behaves. For the investigation, we will have to focus on other verbal and nonverbal behavior, such as actual gaze directions, postures of the discussants, facial expressions, and pitch and power of the voices, none of which were analyzed in this study.

We noticed that the facilitator often nodded when a discussant talked to the other discussants. This may be an indication that he was listening to the opinions of all

Table 11.8 Structure matrix for discriminant function of facilitating behavior "objectifying"

Annotated data of "diverging or converging"	−0.346
The total time to pay attention to the facilitator	0.002
The total time of speaking	−0.251
The difference of attention between the discussers	0.240
The difference of speaking time between the discussers	0.330
The difference of attention between the agreeing groups	0.506
The difference of speaking time between the agreeing groups	0.650

discussants. The facilitator sometimes looked around when a discussant talked to him. This means that he was also paying attention to the responses of other discussants. The facilitator laughed with the discussants even if he did not understand the reason why they were laughing. Such communication behavior is often observed in general communication. We expect that we will be able to reveal the behavior of good listeners by analyzing good facilitators.

11.3 Dynamic Estimation of Emphasizing Points

A key to understanding facilitation behavior is the joint intention shared by the participants in the discussion. Some previous works have focused on helping the discussants to form joint intentions. For example, the systems proposed by Kitamura et al. (2008), Aydogan and Yolum (2007), Kurata (2010) assist in satisfying user's demands by gradually estimating these demands throughout the interaction. They suggest that user demands and needs can be estimated through repetitive interactions. The limitation of these systems is that they do not consider that user's demands and needs may change during the interaction.

However, it is not easy to estimate human internal states through nonverbal information, especially when passively interacting with others. Most previous work estimated user demands and needs over time through repetitive interactions that required active demands from users. There are various studies on estimating human internal states by measuring physiological indices (e.g., Mandryk and Inkpen 2004; Iwaki et al. 2008). There are also several studies that use physiological indices for effective human-agent interaction (e.g., Bosma and André 2004; Prendinger and Ishizuka 2005).

Ohmoto et al. (2011, 2012) characterized intention as a preference structure called an emphasizing point and introduce the dynamic estimation of emphasizing points (DEEP) method, which uses physiological indices, namely, skin conductance response (SCR), electrocardiograms (low frequency (LF)/high frequency (HF) ratios), and the skin temperature of fingers, to detect mental stress such as pleasure, excitement, and tension. The method estimates emphasizing points using these physiological indices, as well as verbal expressions and nonverbal responses. In this study, we apply the proposed method to actual interactions and experimentally evaluate whether proposals that use physiological responses are useful in participants' decision-making and to achieve satisfactory results in the interaction.

Since individuals provide active demands and passive responses through verbal expressions, nonverbal reactions, proposal selection, and physiological state in interactive decision-making, we also consider the interaction process and verbal and nonverbal behavior during the interaction. In the rest of this section, we first introduce the DEEP method and then discuss the empirical evaluation thereof.

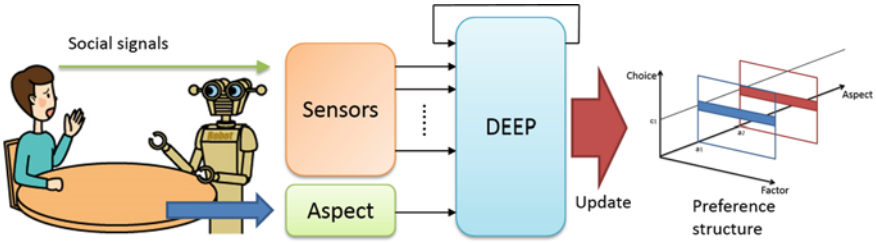


Fig. 11.6 Overview of DEEP. © 2014, Ohmoto and At, Inc. Reproduced with permission

11.3.1 Dynamically Estimating Emphasizing Points

DEEP, a method for estimating emphasizing points that change over time, takes as factors, physiological indices, to determine how aspects of the preference structure depend on these factors, as shown in Fig. 11.6.

We selected factors based on a preliminary analysis using videos and physiological indices of human-human interaction (Ohmoto et al. 2011).

DEEP is applied to situations where many factors, including unknowns, must be considered in decision-making. In such a situation, a user interacts with a system based on DEEP and the system provides some useful proposals for the user's decision-making. In the proposition process in an interaction, first, the two most appropriate proposals at that point are presented by the DEEP system. After the proposition, the system asks the user what his/her demands are and which proposal is better. The DEEP system considers the user's reactions and answers during the explanation and questions. The system then estimates emphasizing points. This process is repeated until one of the propositions satisfies the user's end goal.

In this study, the system provides two proposals simultaneously and the user selects the preferred one. However, the number of proposals does not strongly influence the accuracy of the estimation by DEEP because DEEP is an estimation method based on the user's reactions and active demands, which are independent of the number of proposals.

11.3.1.1 Overview of DEEP

DEEP estimates an emphasizing point from the following three factors:

Verbal Reactions: Classified as an emphasizing point if one of the following two reactions occurs:

(a) Listed words appear in the answers or demands; or
(b) The participant provides backchanneling phrases, which express acknowledgement, surprise, or understanding, such as "ah," "um," "oh," "aha," "I see," and "I understand".

Body movements: Classified as an emphasizing point if the participant repeatedly nods three or more times.
Physiological indices: Classified as an emphasizing point if one of the following responses occurs:

(a) SCR increases more than 10 % compared to the base levels; or
(b) The LF/HF value (electrocardiograph measurement) is greater than 6.0.

The degree of emphasis for an emphasizing point is rated on a scale from zero to five. DEEP repeats the cycle consisting of three phases: explanation, seeking-for-demands, and decision-for-termination, until a satisfactory proposal is found. The estimated emphasizing points are changed in each phase.

In the explanation phase, the estimated emphasizing points are updated according to the user's response, as follows:

Discovery of a new factor to be emphasized: Verbal reactions, body movements, and physiological indices are used as criteria for determining when a new factor is discovered and should be emphasized. When any one of the three criteria appears during an explanation, the system determines that the factor in the explanation should be slightly emphasized, and increases the degree of emphasis from zero to two. When any two or three criteria are present, the system increases the emphasis from zero to three.

Increasing or decreasing degree of emphasis: Verbal reactions, body movements, and physiological indices are used as criteria for determining when a user's degree of emphasis of a particular factor increases or decreases. When any one of the three criteria appears, the system determines that the factor should be emphasized, and increases the emphasis of the factor by one.

When there are physiological reactions, but no verbal reactions or body movements, the system determines that emphasis of the factor should be smaller, and correspondingly decreases the emphasis by one.

In the seeking-for-demands phase, the system asks whether a user has any demands. From the user's response, the system determines what the user's demands are and the corresponding changes to the emphasizing points. The system uses assumed keywords in the user's response to determine demands and changes to demands. Assumed keywords are words that express assumed emphasizing points, demands, and basic words necessary to capture demands. Words that are not expected to be included in answers are ignored. The estimated emphasizing points are updated according to the user's response, as follows:

Discovery of new factors to be emphasized: When the emphasis degree of the discovered factor is zero, the system increases the degree of emphasis from zero to three.

Increasing or decreasing degree of emphasis: When the emphasis of the discovered factor is greater than zero and the system determines that the factor should be increased, the system increases the degree by one. When the system determines that the emphasis of the factor should be decreased and the degree is greater than zero, the system decreases the degree by one.

In the decision-for-termination phases, the system asks if the user is satisfied with one of the proposals. If there is a proposal that satisfies the user's end goal, DEEP terminates the cycle. Otherwise, based on the answer, the system determines which proposal is better for the user or that either both proposals are equally satisfactory or that neither proposal is satisfactory. If the system determines that both proposals equally satisfy the user, the proposal in which the lowest skin temperature was recorded is regarded as the better proposal. If the system determines that neither proposal is satisfactory, the system does nothing.

11.3.1.2 Selecting the Next Step Based on DEEP Results

According to the criteria mentioned above, changes to user's emphasizing points are estimated after the proposals have been presented and data have been collected from the user's reactions and response. After the estimation, the next two proposals are selected based on the estimation results.

The next proposals are selected using a table of orthogonal arrays prepared in advance. Orthogonal arrays are a special set of Latin squares, which can be used to estimate main effects using only a few experimental runs. From the table, the two proposals that most satisfy the user's emphasizing points are selected. If several proposals in the table satisfy a user's emphasizing points, the two proposals closest to the user's preferred proposal are selected. If neither proposal satisfies the emphasizing points, the two proposals furthest from the previous proposition are selected. The distances between proposals are calculated by cosine similarity.

11.3.2 Evaluation of DEEP

Here we investigate whether the DEEP method can accurately estimate emphasizing points in which many factors, including unknown factors, must be considered in decision-making. In the experiment, we used human-like virtual agents (ECAs) to strictly control the verbal and nonverbal expressions of the agent, since these could affect a user's impressions of the proposals presented. A human operator controlled the ECAs using a WoZ (Wizard of Oz) interface.

We compared the DEEP method with a gradual method, which is a modified version of the work by Kurata (2010). The gradual method was deemed the most suitable previous work. Using only the user's selection of the two proposals, the method gradually approaches a satisfactory proposal. In the gradual method, the ECA provides the two proposals closest to the preferred proposal of the user. Should the user decide that neither of the proposals are suitable, the next two proposals presented are those furthest from the last two proposals. The method does not consider dynamic changes in the user's emphasizing points during the interaction.

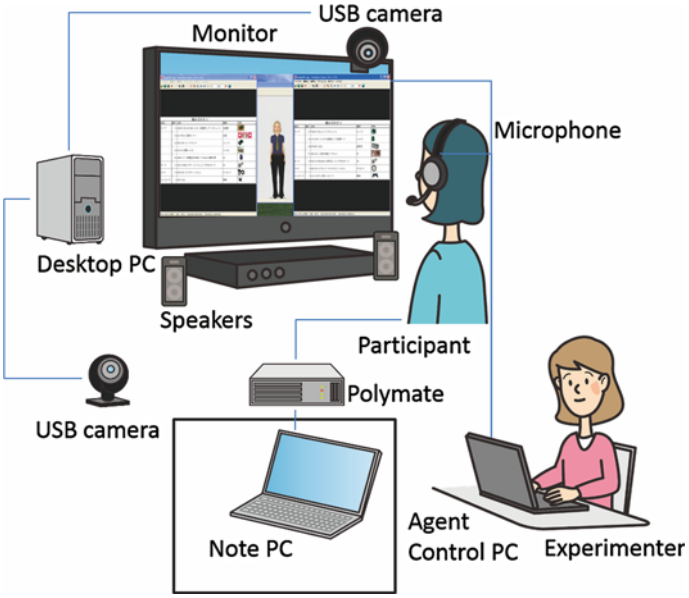


Fig. 11.7 Experimental settings to evaluate DEEP. © 2014, Ohmoto and At, Inc. Reproduced with permission

We set a mobile robot design task in which each group of participants was asked to design a mobile robot using a robot parts catalogue. We hired 27 students (20 males and 7 females) who were undergraduate students aged between 18 and 25 years (average of 20.6 years). These students were not familiar with mobile robots, but all were enrolled in a science course. We finally analyzed the data acquired from 26 participants, 20 males and 6 females; the data for one female participant were deleted because she slept during the experiment. All of the participants interacted with the ECA both via DEEP and without DEEP.

The experimental setting is shown in Fig. 11.7. The participant sat in front of a 100-inch screen displaying the ECA, while the investigator sat out of view of the participant and entered the stimuli via a WoZ interface. Two video cameras recorded the participant's behavior; one was attached to the screen to record the participant's behavior, and another was placed behind the participant to record the ECA's behavior. The participant's voice was recorded by microphones placed under the screen. A Polymate device was used to measure SCR, the electrocardiogram, and skin temperature of the fingers. The investigator instructed all participants to keep their left arms on the armrest.

After a brief explanation of the experiment, the investigator started the experiment, as well as the recording of the video and physiological indices. Two sessions were conducted during the experiment. The investigator randomly selected either the DEEP or the gradual method for the first session, with the other ECA used in the second session. Each participant repeatedly selected proposals provided by

the ECA until one of the proposals satisfied his/her end goal for the robot. At the end of each session, the participant completed a questionnaire aimed at evaluating the ECA.

Each participant interacted with the investigator over two sessions, in which they designed different robots to achieve different tasks. The participant could change the design concept of the robot during the session without informing the investigator. Each task included 23 criteria that the robot should satisfy and there were various ways to design the robots to achieve the same purpose. Example purposes in Situation A included "taking photos of beautiful scenery" and "introducing old temples and shrines," while in Situation B they included "a mountain climbing race" and "a city obstacle race".

The investigator input data into the system comprising verbal reactions, body movements, and physiological indices, because we could not robustly capture these data in real-time. Each ECA generated verbal and nonverbal behavior that had been previously designed by the investigator based on the expected reactions.

Both ECAs accepted the results of the user's choice. In addition, the ECA with DEEP accepted data as described in the previous section. Verbal reactions and body movements were determined through visual observation by the investigator. Physiological indices were automatically measured and the investigator annotated which words or explanations may have triggered the physiological responses. Each ECA used the input data to determine which proposals should be presented in the next proposition.

To evaluate the accuracy in estimating emphasizing points, we randomly selected seven participants before the experiment. These seven participants chose their top three emphasizing points out of 23 factors at the end of each session. The reason why we chose a limited number of participants is that the choice of emphasizing points was a very time consuming process because the participants had to understand the meanings of the 23 factors and reflect on their decision-making. Therefore, we could gather only a limited number of participants for the research. We then calculated concordance rates between the factors chosen by the users and the factors estimated by each ECA. We used a t-test to compare the concordance rates of DEEP and the gradual method. The results are given in Table 11.9, where the values reflect the average numbers of matched factors.

The results of the t-test confirm that DEEP can estimate emphasizing points more accurately than the gradual method. We suggest that DEEP has adequate performance in estimating emphasizing points because the average is high and the standard devia-

Table 11.9 Result of t-test for accuracy in estimating emphasizing points

	Proposed	Gradual
Average	2.1	1.0
Standard deviation	0.69	1.0
t	2.49	
p	0.029*	

*$p < 0.05$

Table 11.10 Result of Chi-squared test on the effect of the method on dynamic changes

	Changed	Not changed
Proposed	25	1
Gradual	22	4
p	0.158	

Table 11.11 Result of sign-test for the effect of ECA method on dynamic changes

	Score (proposed > gradual)
Average	1.0
Standard deviation	1.9
p	0.013*

*$p < 0.05$

tion is low. Therefore, by using verbal reactions, body movements, and physiological indices, DEEP can correctly estimate the emphasizing points of each participant.

To evaluate the ability to change emphasizing points and the purpose of the robot, the researcher asked the participants to answer five questions on the ECA's behavior by rating these using a seven-point scale. The scale was presented as seven ticks on a black line without numbers, which we scored from −3 to +3. Each of the five behaviors was evaluated in two ways, that is, how much the ECA affected the participant's thoughts ("how much" question), and which method affected the participant's thoughts more ("which" question).

We also asked each participant to answer whether s/he dynamically changed his/her emphasizing points and purpose of the robot during the interaction ("how much" question). We performed a Chi-squared test to confirm whether there was a significant difference between the results of DEEP and the gradual method. These results are listed in Table 11.10. Participants also indicated which method caused more dynamic changes ("which" question). We performed a sign-test to calculate the difference between the two methods, as depicted in Table 11.11. (A value of −3 denotes that the gradual method caused more changes, whereas +3 denotes that DEEP caused more changes.)

There was no significant difference for the "how much" scores, because both methods caused dynamic changes during the interaction. This means that humans easily change their emphasizing points even when simple algorithms provide the proposal and explanation. Meanwhile, DEEP caused significantly more changes than the gradual method. It is possible that the participants paid attention to broader factors than those contained in the mobile robot task because DEEP was sensitive to changes in emphasizing points and modified subsequent proposals accordingly.

To evaluate the participant's satisfaction with regard to the final proposal, the investigator asked participants to share their degree of satisfaction with the ECA's final proposal ("how much" questions). The results of a Wilcoxon signed-rank test are given in Table 11.12 (where −3 denotes "not at all", and +3 denotes "very much").

Participants also indicated which method provided the more satisfactory proposal ("which" question). The results of a sign-test are given in Table 11.13 (−3 denotes "satisfaction with the final proposal of the ECA with the gradual method", while +3 denotes "satisfaction with the final proposal of the ECA with DEEP").

Both Tables 11.12 and 11.13 show that the ECA with DEEP provided a significantly more satisfactory proposal than the ECA with the gradual method. However, it is important to note that the standard deviations for the results of the ECA with DEEP in Tables 11.12 and 11.13 are fairly large. We consider the implications of this result in the discussion.

To evaluate the naturalness of the proposal, the investigator asked participants to consider how natural the sequence of proposals was ("how much" question). We performed a Wilcoxon signed-rank test, the results of which are shown in Table 11.14 (−3 denotes "not at all", while +3 denotes "very much"). Participants were also asked which method provided more natural proposals ("which" question). The results of a sign-test are given in Table 11.15 (−3 denotes that the ECA with the gradual method provided more natural proposals, and +3 denotes that the ECA with DEEP provided more natural proposals).

Both Tables 11.14 and 11.15 show that DEEP provided significantly more natural proposals than those provided by the gradual method. The content of each of the proposals provided by DEEP and the gradual method was the same. Therefore, naturalness must be attributed to the presentation order and whether the proposals reflected the user's emphasizing points. DEEP most likely provided more natural proposals because it could quickly reflect changes in the user's emphasizing points.

To summarize, we confirmed that our proposed method has better accuracy in estimating emphasizing points, has more latitude in changing emphasizing points, is more natural, and ensures that participants are more satisfied with the final proposal. In addition, we found evidence that individuals often change their emphasizing points and the purpose of the task during an interactive decision-making process.

The standard deviation values of the results of our proposed method are all relatively large. This means that the effectiveness of the proposed method differs across individuals. One of the reasons for this is that some participants' demands could not be satisfied by the proposals. In these cases, the ECA did not provide any notification of the impossibility or alternatives. In many possible cases, the ECA with DEEP responded quickly to participants' demands, so, in the impossible cases, participants with impossible demands felt disappointed, as would be expected. Future work should include notification capabilities in the ECA.

Table 11.12 Result of Wilcoxon signed-rank test on user satisfaction of ECA's final proposal

	Proposed	Gradual
Average	1.8	0.81
Standard deviation	2.3	1.6
z	2.11	
p	0.035*	

*$p < 0.05$

Table 11.13 Result of sign-test on which ECA provided the best proposal

	Score (proposed > gradual)
Average	1.1
Standard deviation	2.3
p	0.038*

*$p < 0.05$

Table 11.14 Result of Wilcoxon signed-rank test on naturalness of ECA proposals

	Proposed	Gradual
Average	1.2	0.27
Standard deviation	1.8	1.6
z	2.4	
p	0.015*	

*$p < 0.05$

Table 11.15 Result of sign-test on which ECA provided more natural proposals

	Score (proposed > gradual)
Average	0.89
Standard deviation	1.7
p	0.027*

*$p < 0.05$

11.4 Dynamically Estimating Emphasizing Points for Group Decision-Making

In many cases of group decision-making, people often have conflicting opinions. In this situation, people consider not only their demands but also the effects on their relationships. There are many researches about conflict in conversation and conflict management. The typical reactions of people can be largely divided into two groups: avoiding conflict or persuading the partner. Extending our DEEP method to group decision-making, we need to respond to both conflict situations and human reactions. gDEEP uses DEEP to estimate the emphasizing points of groups, but it can separate avoiding conflict and persuading discussion partners. In the rest of this section, we first introduce gDEEP and then discuss the evaluation thereof.

11.4.1 Dynamic Estimation of Emphasizing Points Extended to Group Decision Making (gDEEP)

The gDEEP method uses DEEP to estimate emphasizing points in group discussions (Ohmoto et al. 2013). gDEEP is applied to situations where many factors, including unknowns, must be considered in decision-making.

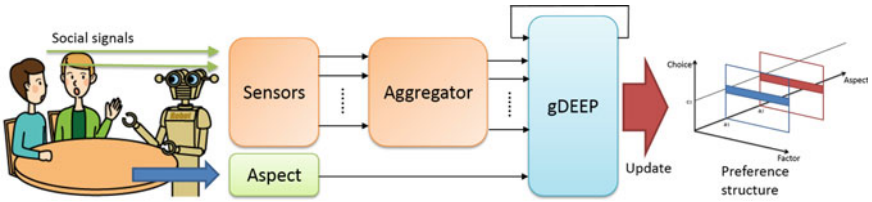


Fig. 11.8 Architecture of gDEEP. © 2014, Ohmoto and At, Inc. Reproduced with permission

The proposition process in an interaction can be described as follows. First, the two most appropriate proposals at that point are explained by the gDEEP system. After the presentation, the system asks the users to discuss within the group which proposal is better. After the discussion, the members state their preference and demands. The gDEEP system considers the users' nonverbal reactions, physiological indices, demands, and selection of the two proposals and then estimates the emphasizing points of the group. The two proposals that satisfy the group's emphasizing points more are picked from a prepared proposal list. If several proposals in the list can satisfy the group's emphasizing points, the two proposals closest to the best proposal are selected. If neither proposal can satisfy the emphasizing points, the two proposals furthest from the previous proposals are selected. The distances between the proposals are calculated by cosine similarity. This process is repeated until one of the propositions satisfies the group.

The architecture of gDEEP is shown in Fig. 11.8. Sensory inputs from discussants are aggregated and input into DEEP. We consider two types of aggregation: union-based and intersection-based. The difference between the two methods lies in their use of nonverbal reactions, physiological indices, demands, and choice between the two proposals in the group. The methods are explained below.

11.4.1.1 The Union-Based Method

The union-based method focuses on the responses to other members' opinions. We expect that this method will be used by the Avoiding Conflict group. The estimated emphasizing points contain as many emphasizing points as possible as identified by the members. The degree of emphasis for an emphasizing point is rated on a scale from 0 to 5 with the rating levels defined by the following rules.

- +3: The system increases the emphasis of the factor by 3 when a member clearly agrees with his/her partner's opinion through verbal reactions, body movements, and changes in physiological indices.
- +2: The system increases the emphasis of the factor by 2 when a member implicitly agrees with his/her partner's opinion via body movements or changes in physiological indices.

- +1: The system increases the emphasis of the factor by 1 when a member makes agreeable responses without any supporting body movements and changes in physiological indices.
- −1: The system decreases the emphasis of the factor by 1 when a reaction is recorded as a change in the physiological indices but there are no agreeable responses.
- −5: The system decreases the emphasis of the factor by 5 when a member clearly disagrees with his/her partner's opinion.

11.4.1.2 The Intersection-Base Method

The intersection-based method focuses on clearly accepted opinions. We expect that this method will be used by the Persuading the Partner group. Estimated emphasizing points only contain the emphasizing points shared by all members. The degree of emphasis for an emphasizing point is rated on a scale from 0 to 5, with the rating levels defined by the following rules.

- +3: The system increases the emphasis of the factor by 3 when a member expresses a positive opinion and other members clearly provide verbal reactions, body movements, or changes in physiological indices.
- +2: The system increases the emphasis of the factor by 2 when a member expresses a positive opinion but other members do not provide verbal reactions, body movements, or changes in physiological indices.
- +1: The system increases the emphasis of the factor by 1 when a member implicitly agrees with his/her partner's opinion through the use of body movements or changes in physiological indices.
- −1: The system decreases the emphasis of the factor by 1 when a reaction is recorded as a change in the physiological indices but a member does not make an agreeable response.
- −5: The system decreases the emphasis of the factor by 5 when a member clearly disagrees.

11.4.2 Evaluation Experiment

We conducted an experiment to investigate whether we should change the estimation method (that is, use either the union-based or intersection-based method) depending on the interaction style (avoiding conflict or persuading a partner). In the experiment, we used human-like virtual agents, ECAs, to strictly control the verbal and nonverbal expressions in the interaction, as these could affect a user's impressions of the proposals presented. The ECAs were controlled by a WoZ interface because accurate voice recognition can be difficult. We classified participants into two groups based on their interaction style of either avoiding conflict or persuading their partners. We

analyzed whether the accuracy of the ECA's final proposal differed for the union-based and intersection-based methods depending on interaction style, and whether satisfaction with the human-agent interaction differed according to the estimation method and interaction style.

The participants in this experiment were 16 Japanese college students (all female), aged between 20 and 28 years (average age was 22.5 years). The participants were divided into eight pairs with each pair representing a group. Members of a pair were acquainted with each other and they had a mutual friend, that is, the hypothetical target in the task. All groups interacted with the ECA using both union-based and intersection-based gDEEP. After the experiment, we divided the groups based on whether they avoided conflict or persuading their partners. Groups in which no conflict of opinion occurred were classified as the "avoiding conflict" group (AC group); the remainder were classified as the "persuading a partner" group (PP group). There were four groups in each category with a total of eight participants.

We formed two experimental groups. Each group interacted with two ECAs on two different tasks: choosing a wedding present and choosing a toy for a child. Thus, each group participated in two experimental sessions. The task of choosing a wedding present had 25 criteria while choosing a toy for a child had 23. Each ECA was implemented using a different gDEEP method, either union-based or intersection-based.

The investigator input into the system all data containing verbal reactions and body movements because these data could not be robustly captured in real-time. Each ECA generated verbal and nonverbal behavior that had been previously designed by the investigators based on the expected reactions.

Both ECAs accepted the users' demands, nonverbal reactions, physiological indices, and selection choice regarding the two proposals. Verbal reactions and body movements were determined via visual observations by the investigators. Physiological indices were automatically measured, and it was decided which words or

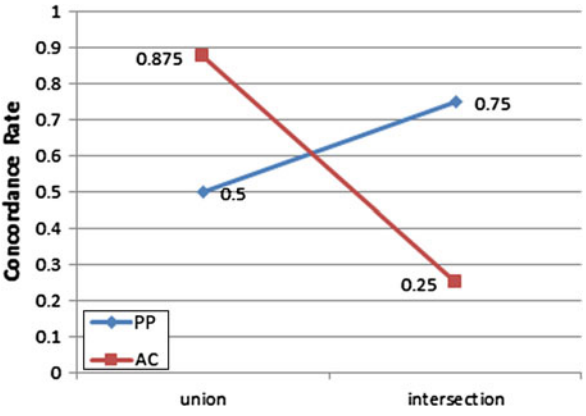


Fig. 11.9 Concordance rates between the proposals chosen by the participants and those estimated by each ECA

Table 11.16 Results of ANOVA for accuracy of ECA's final proposal

Source	SS	df	MS	F	p
Group	0.031	1	0.031	0.14	0.72
Error	3.2	14	0.23		
Method	0.28	1	0.28	1.5	0.25
Error	2.7	14	0.19		
Interaction	1.5	1	1.5	8.0	0.014*
Total	7.7	31			

*$p < 0.05$

explanations may have triggered the physiological responses. Each ECA used the input data to select the proposals presented in the next proposition.

The participants sat in front of a 100-inch screen displaying the ECA, while the investigators sat out of view of the participants and entered the stimuli via a WoZ interface. Two video cameras recorded both the participants' and the ECA's behavior: one was attached to the screen to record the participants' behavior, and another was placed behind the participants to record the ECA's behavior. We recorded the participants' voices using microphones placed under the screen. A Polymate device was used to measure SCR, the electrocardiogram, and skin temperature of the fingers. The participants were instructed to keep their left arms on an armrest.

After a brief explanation of the experiment, the investigator started the experiment, as well as the recording of the video and physiological indices. Two sessions were conducted during the experiment. The investigator randomly decided which ECA and which method, either union-based or intersection-based, were to be used in the first session; the other ECA was used in the second session. The group of participants repeatedly selected proposals provided by the ECA until one of the proposals satisfied their goal. At the end of each session, the participants completed a questionnaire to evaluate the ECAs.

All participants selected their best proposal out of 40 prepared proposals at the end of each session. We then calculated the concordance rates between the proposals chosen by each participant and the proposal estimated by each ECA.

Figure 11.9 shows the results. The value was set to 1 (truth) when the chosen proposal was the same as the estimated proposal, otherwise, it was set to 0 (false). The data were submitted to a 2 (interaction style) $\times$ 2 (method for gDEEP) analysis of variance (ANOVA) (Table 11.16).

Figure 11.9 shows the results. A value of 1 (true) indicates that the chosen proposal is the same as the estimated proposal, whereas 0 (false) indicates the contrary. The data were submitted to a 2 (interaction style) $\times$ 2 (method for gDEEP) analysis of variance (ANOVA) (see Table 11.16).

According to Table 11.16, the interaction effect is statistically significant, indicating that the average change scores for the two groups were different. The significant interaction effect was further assessed by simple main effect tests. In Table 11.17, a simple main effect test comparing the union-based and intersection-based methods

Table 11.17 Results of simple main effect for accuracy of ECA's final proposal

Source	SS	df	MS	F	p
Group (union-base)	0.56	1	0.56	2.7	0.11
Group (intersection-base)	1.0	1	1.0	4.8	0.038*
Error		28	0.21		
Method (AC group)	0.25	1	0.25	1.4	0.27
Method (PP group)	1.6	1	1.6	8.1	0.013
Error		14	0.19		

*$p < 0.05$

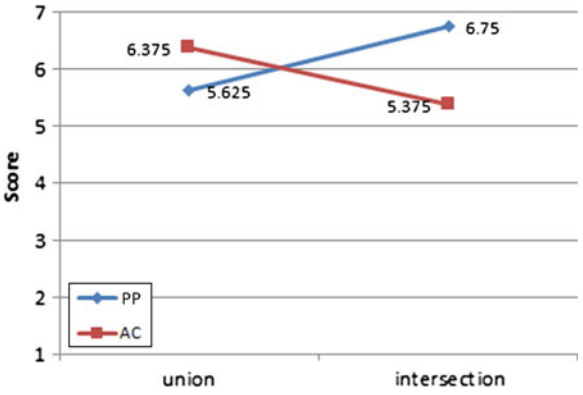


Fig. 11.10 Average scores for each group and each method

for the AC group shows that the concordance rate was significantly higher when the ECA used the union-based method. The intersection-based method greatly increased some degree of emphasis when participants persuaded their partners. As participants in the AC group established consensus by conforming to their partners' opinions, it was difficult to estimate the emphasizing points of the group when the ECA used the intersection-based method. In fact, in Table 11.17, a simple main effect test comparing the AC and PP groups when using the intersection-based method shows that the concordance rate was significantly higher when the ECA with the intersection-based method interacted with the PP group. Therefore, we recommend that the interaction style be considered when deciding on which estimation method to use to estimate proposals accurately.

As for the results of participant satisfaction with human-agent interaction (HAI), the participants rated questions probing the level of satisfaction with HAI using a seven-point scale. The scale was presented as seven ticks on a black line without numbers, which we scored from 1 to 7. We then calculated the averages for each group and each method. Figure 11.10 shows the results. The data were analyzed using a 2 (interaction style) $\times$ 2 (method for gDEEP) ANOVA (see Table 11.18).

As shown in Table 11.18, the interaction effect is statistically significant. The significant interaction effect was further assessed by simple main effect tests. In Table 11.19, a simple main effect test comparing the AC and PP groups with the

intersection-based method shows that the degree of satisfaction in the AC group was significantly lower. Participants in the AC group often retracted their opinions during the group discussion. However, the intersection-based method greatly increased the degree of emphasis when participants persuaded their partners. Thus, the retraction prevented the ECA from accurately estimating the emphasizing points of the group when using the intersection-based method. In contrast, a simple main effect test comparing the union-based and intersection-based methods on the AC group shows (see Table 11.19) a tendency for the PP group to be satisfied with the HAI during the experiment when the ECA used the intersection-based method. Therefore, we suggest that the two proposed methods for estimating the emphasizing points of a group have a different effect on satisfaction with HAI depending on the interaction style.

In this study, we confirmed that the interaction process to build consensus differed depending on the interaction style used, that is, either avoiding conflict or persuading a partner. We then proposed two methods to estimate the emphasizing points of a group using each of these interaction styles. Through experiments, we found that these methods estimated proposals accurately and satisfied participants in the corresponding group.

However, the proposed methods do have some limitations. First, the proposed methods and implemented system cannot detect which interaction style participants prefer. This is important to identify when deciding which estimation method to use. Although there was no significant difference for this experiment, speaking speed, response latency, and the rate of looking at a partner's face differed between the AC and PP groups. We expect that interaction behavior is a clue in detecting interaction style.

Table 11.18 Results of ANOVA for participant satisfaction with human-agent interaction

Source	SS	df	MS	F	p
Group	0.78	1	0.031	0.14	0.72
Error	28	14	0.23		
Method	0.031	1	0.28	1.5	0.25
Error	20	14	1.5		
Interaction	9.0	1	9.0	6.2	0.026*
Total	59	31			

$^{*}p < 0.05$

Table 11.19 Results of simple main effect for participant satisfaction with human-agent interaction

Source	SS	df	MS	F	p
Group (union-base)	2.3	1	2.3	1.3	0.27
Group (intersection-base)	7.6	1	7.6	4.3	0.047*
Error		28	0.21		
Method (AC group)	5.1	1	5.1	3.5	0.084^{+}
Method (PP group)	4.0	1	4.0	2.7	0.12
Error		14	1.5		

$^{+}p < 0.1$; $^{*}p < 0.05$

Second, we did not focus on mediating conflicts between group members. The reason for this is that we could not propose a compromise that would satisfy all members of the group in the preliminary experiment. It is difficult to mediate a conflict of opinions even when a human tries to do so, and even more so when the opinions of a virtual agent are often taken less seriously. To mediate conflicts between group members, we need to focus on how an agent can ensure that a user will value the agent's opinion.

Identification of preferred interaction styles is left for future research. If we could identify which style participants would prefer, we could dynamically switch to the appropriate estimation method during an interaction. A further issue is mediating conflicts between group members. Although conflicts are difficult to resolve, this is an important consideration in future studies.

11.5 Facilitative Agent

In previous sections, to support interactive decision-making, we proposed a method named DEEP which encouraged decision-making by awakening the intrinsic emphasizing points. Furthermore, extrinsic subjective interpretations, such as friend's opinion and word-of-mouth advertising, also encourage decision-making because they provide case examples to interpret the factors we have to consider and emphasize to reach an appropriate decision.

One of the methods to conduct smooth and effective decision-making using subjective opinions is "facilitation" as we discussed in Sect. 11.2 in this chapter. The facilitation process includes a divergent zone, groan zone, convergent zone, and closure zone (Kaner 2007). Especially in the convergent zone, the facilitator subjectively summarizes the discussers' opinions and limits the direction of the discussion. We assume that we can effectively support interactive decision-making based on extrinsic subjective information by applying the facilitation process to interactive decision-making. For example, a counselor in interactive decision-making provides subjective opinions, such as "I think that's good," to move into the convergent zone of the decision-making interaction.

In this section, we describe a "facilitative decision-making support agent" that provided proposals awakening the intrinsic emphasizing points and subjective opinions to realize divergent and convergent processes in decision-making. We then introduce the effect to encourage the decision-making using intrinsic and extrinsic factors by comparing an estimation agent that only provided proposals that reflected the emphasizing points by DEEP.

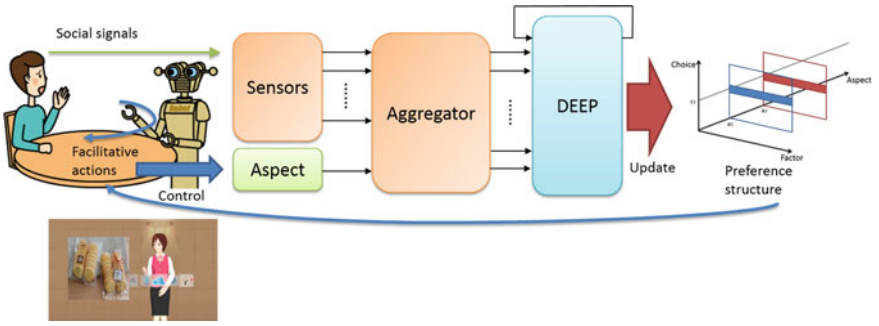


Fig. 11.11 Architecture of the facilitative agent and the displayed screen. © 2014, Ohmoto and At, Inc. Reproduced with permission

11.5.1 A Facilitative Decision-Making Support Agent

We used MMDAgent[1] as the interface for a facilitative decision-making support agent. MMDAgent is a toolkit for building voice interaction systems, and includes Julius, Open J Talk, and a number of other systems. We developed a control system that received inputs from MMDAgent (recognized voice data) and Polimate (LF/HF data and SCR data) and sent outputs of motion and speech commands to MMDAgent. The inputs for the facilitative agent were automatically captured, with the exception of determining whether a user's utterance was positive or negative and whether the user's utterance was a question because we could not robustly determine them in real-time. These responses were determined by a WOZ (Wizard of Oz). The agent automatically generates verbal and nonverbal behavior that had been previously designed, with the exception of the answers to the questions. The answer was selected by a WOZ operator. The architecture of the facilitative agent and the displayed screen are shown in the Fig. 11.11.

11.5.1.1 Method to Realize Divergent and Convergent Processes in an Interaction

The facilitative agent supports the user's decision-making during the interaction. The agent uses social signals for active listening and teaming to realize divergent and convergent processes in the interaction. The used signals are the frequency of providing a new proposal, recommendation from the agent, mimicry of nodding motions, and utterances. The agent's behavior depends on how it recognizes the discussion status:

Agent's behavior in the divergent process: The agent provides a small nod once in reaction to the user's utterance. The frequency of providing a new proposal is low.

[1] http://www.mmdagent.jp/.

The agent provides a new proposal after she explains three emphasizing points. The furthest proposal from the previous one is selected as a new proposal. The degree of emphasis decreases if the emphasizing point is not explained in the previous proposal.

Agent's behavior in the convergent process: The agent provides two large nods in reaction to the user's utterance. The frequency of providing a new proposal is high. The agent provides a new proposal after she explains one emphasizing point, which is a recommendation. The nearest proposal to the previous one is selected as a new proposal. The degree of emphasis decreases only when the emphasizing point is clearly refused in the previous proposal.

The agent starts the interaction with a divergent process. The agent switches from the divergent process to a convergent process when she detects the following situations:

- There are more than three emphasizing points, with a degree of emphasis of more than one, and the degree of emphasis does not change during the explanation.
- The user offers a convergent opinion such as "I want to see like this one" and "I want to determine".

The emphasizing points and the degree of emphasis are the subjective opinions of the agent. The emphasizing points are set to the values of the recent proposal at the time when the agent switches from the divergent process to the convergent process. This causes the agent to search the neighbor of the last proposal of the divergent process during the convergent process. The degree of emphasis decreases when the emphasizing point is clearly refused by the user.

11.5.2 Experiment

The purpose of this experiment was to investigate how the interaction process during the transition from divergent to convergent, which includes providing extrinsic subjective information, affects the final goal of the decision-making and impressions of the process. In the experiment, to strictly control the verbal and nonverbal expressions, we used two types of agents: a facilitative agent who provided subjective opinions to realize divergent and convergent processes in decision-making and an estimation agent who only provided proposals that reflected the emphasizing points of each participant. We explained the facilitative agent in the previous section. Both of the ECAs implemented DEEP. The estimation agent is similar to the agent used in previous studies (Ohmoto et al. 2011, 2012, 2013). Here, we analyze the reaction behavior of participants and questionnaire responses.

Participants were asked to design gift-wrapping for a valentine present. The participants did not know what was appropriate gift-wrapping. The participants interacted with the agent to design the gift-wrapping. We identified 30 factors that the participants considered when they designed the wrapping. We expected that the emphasizing

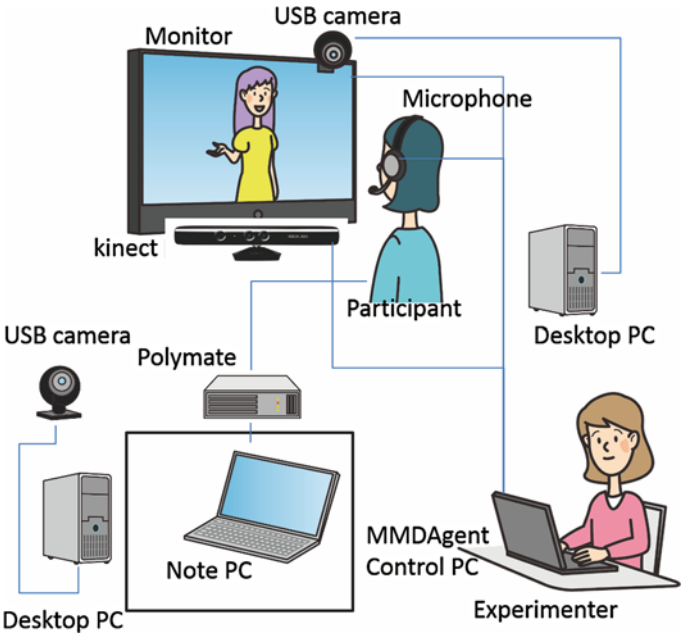


Fig. 11.12 Experimental settings to evaluate the facilitative agent. © 2014, Ohmoto and At, Inc. Reproduced with permission

points would change during the interactions and the participants would take advice from the agent because they tried to predict what the receiver of the gift would like.

The experimental setting is shown in Fig. 11.12. The participant sat in front of a 60-inch monitor displaying the ECA. The experimenter sat out of view of the participant and entered the stimuli via a WOZ interface. Two video cameras recorded the participant's behavior: one was placed on the monitor to record the participant's behavior and another was placed behind the participant to record the ECA's behavior. The kinect was placed under the monitor and captured the nodding motion of the participant. The participant's voice was recorded using microphones, which were placed under the monitor. Polymate was used to measure SCR and the electrocardiogram. The experimenter instructed the participants to keep their left arm on an armrest.

The participants in this experiment were 20 Japanese college students (all female), aged between 20 and 27 years (the average age was 21.4 years). They did not know about gift-wrapping. The participants were divided into two groups: one interacted with the facilitative agent and the other with the estimation agent.

After a brief explanation of the experiment, the experimenter began the experiment, and the recording of the video and physiological indices. The participant repeatedly asked questions about the proposal and considered the proposals provided by the ECA until one of the proposals satisfied the participant. At the conclusion of the experiment, the participant completed a questionnaire regarding evaluations of the interaction process.

11.5.2.1 Results of Reaction Analyses

We focused on reaction latency and changing emphasizing points to study how differently participants reacted to the two agents.

Reaction latency: We extracted a reaction latency for each participant. The reaction latency was defined as the time from the end of the utterance of the agent to the start of the participant's reaction. Reaction latency data were classified into two categories: data in the first half of the interaction and data in the second half. We

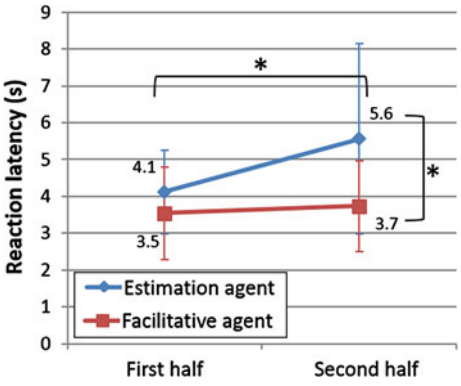


Fig. 11.13 Results of the t-test for reaction latency

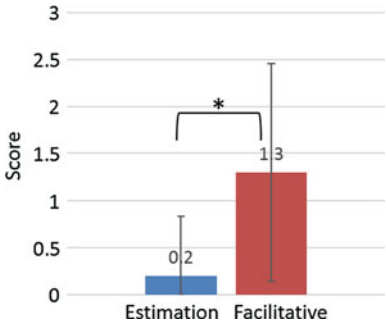


Fig. 11.14 Result of t-test of the number of the changed emphasizing points

conducted a t-test to compare the data from the facilitative agent group with the data from the estimation agent group. The results are shown in Fig. 11.13.
In the first half of the interaction, there is no significant difference between the reaction latency in the facilitative agent group and that in the estimation agent group. In contrast, there is a significant difference in the second half. In addition, there is a significant difference between the reaction latency in the estimation agent group in the first half and that in the second half.
We interpreted these results as follows. When the participant interacted with the estimation agent, she carefully thought about the proposal in the second half of the interaction. As the participant had already obtained a lot of information from the agent in the first half, she did not pay any further attention to the interaction with the agent. In contrast, when the participant interacted with the facilitative agent, she actively interacted with the agent in the second half. As the participant regarded the subjective opinions of the facilitative agent as helpful information, she continuously interacted throughout the whole interaction. Therefore, we can confirm that subjective information was helpful in interactive decision-making.

Changing emphasizing points: At the end of the experiment, the participants chose emphasizing points that they changed during the interaction. We then calculated the number of changed emphasizing points for each participant. We conducted a t-test to compare the number in the facilitative agent group with that in the estimation agent group. The results are shown in Fig. 11.14.
The number in the facilitative agent group was significantly higher than that in the estimation agent group ($t = -2.63369$, $p < 0.05$). It seems that there were less changes in the facilitative agent group because the facilitative agent provided similar proposals in the second half of the interaction. We discuss this further below.

As the results of the reaction latency analysis have shown, the participants in the estimation agent group carefully considered the proposal in the second half of the interaction. It would seem that as they made their decision only based on intrinsic emphasizing points, they could not recognize changes to the emphasizing points. Similarly, the participants in the facilitative agent group did not recognize some changes because the total number of the changes reported by them was small. However, in the facilitative agent group, the agent provided extrinsic subjective opinions. Therefore, they could explicitly recognize some of the changes.

11.5.2.2 Results of Questionnaires

The participants answered three rating questions on the ECA's behavior using a seven-point scale. The scale was presented as seven ticks on a black line without numbers, which we scored from 1 to 7. The results are summarized in Fig. 11.15. The detailed analysis is as follows:

Participant's satisfaction of interaction with the ECA: Participants reported how satisfied they were with the interaction with the ECA. As a result of a Wilcoxon signed-rank test, the facilitative agent provided significantly more satisfactory interactions than the estimation agent ($z = 3.5$, $p < 0.001$).

Naturalness of ECA's interaction: Participants answered how natural the sequence of proposals was. As a result of a Wilcoxon signed-rank test, the facilitative agent provided significantly more natural interactions than the estimation agent ($z = 2.3$, $p < 0.05$).

Appropriateness as a decision-making adviser: Participants answered how appropriate the ECA was as a decision-making adviser. As a result of a Wilcoxon signed-rank test, the facilitative agent provided significantly more appropriate than the estimation agent ($z = 2.0$, $p < 0.05$).

Realizing divergent thinking and convergent thinking: Participants answered how useful the proposals by the agent were for divergent thinking and convergent thinking. Wilcoxon signed-rank tests shows that the facilitative agent was significantly more useful for divergent interactions and convergent interactions than the estimation agent (divergent: $z = 2.5$, $p < 0.05$; convergent: $z = 2.0$, $p < 0.05$).

The questionnaire results show that the interaction process with the facilitative agent was better than that with the estimation agent. This suggests that the convergent interaction process, where subjective opinions are expressed, produces a better impression of the interaction process. Of particular interest is the result stating how useful the agent's proposals were for divergent thinking. This means that the convergent interaction process contributed to divergent thinking. We consider that one of the reasons for this result is that the participants felt they finished searching the whole of the problem space by switching from the divergent process to the convergent process.

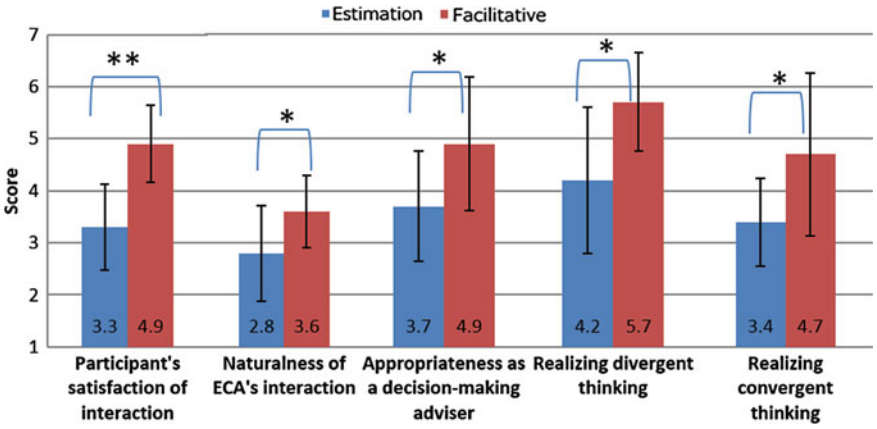


Fig. 11.15 Means and the results of Wilcoxon signed-rank tests on the questionnaires

11.5.3 Discussion

In this study, we evaluated a facilitative agent who provided subjective opinions to realize divergent and convergent processes in decision-making, and found that this led to higher scores for participant satisfaction regarding ECA interactions, the naturalness of ECA's interaction, and impressions of the decision-making process. From the results of the reaction analyses and questionnaires, we confirmed that the participants who interacted with the facilitative agent recognized the divergent thinking process more explicitly than those who interacted with the estimation agent. From these results, we suggest that we can partly achieve the interaction based on "bubbling intention" model discussed in Sect. 11.1 in this chapter.

We can explain the results of this study from the perspective of "bubbling intention" model which means the decision or intention is extemporarily shaped by extrinsic stimulus and intrinsic pressure. For example, in the case of the results of the reaction analyses, the participants did not receive extrinsic stimulus, especially in the second half of the interaction when they interacted with the estimation agent. It took a long time to recognize their own emphasizing points and shape their decision; therefore, the reaction latency grew longer and they could not recognize the changes in the emphasizing points. Regarding the results of the questionnaires, the participants could clearly recognize the divergent thinking, which was intrinsic pressure, because switching from the divergent process to the convergent process was triggered by extrinsic stimuli provided by the facilitative agent.

From a different perspective, the concept means that we do not need to precisely estimate the inner states (e.g., emphasizing points, emotions, and intentions) of communication partners during natural communication. This can be used for the design of interaction artifacts like the facilitative agent.

In this study, we investigated the effect of the divergent and convergent interaction process in interactive decision-making. We used ECAs to evaluate the effect because it is difficult for human agents to achieve tightly controlled interactions with participants. We conducted an experiment that compared the results of interactive decision-making with two types of ECAs: a facilitative agent who provided subjective opinions to realize divergent and convergent processes in decision-making and an estimation agent who only provided proposals that reflected the emphasizing points of each participant. The facilitative agent encouraged decision-making by intrinsic and extrinsic factors. As a result, we can confirm that the facilitative agent increased the participant's satisfaction with the ECA interaction, the naturalness of ECA's interaction, and the impression of decision-making process. In addition, we developed a hypothesis called the "Bubbling intention". We will verify the concept of the bubbling intention from various perspectives in future research.

11.6 Summary

In this chapter, we presented how the cognitive design framework can be applied to decision support. First, we consider the method to support human agent interaction based on the human activities. We then focused on the facilitation which is an extrinsic approach to support decision-making. In addition, we proposed a model of joint intention named "bubbling intention". In the model, the decision or intention is extemporarily shaped by an extrinsic stimulus (e.g., a partner's behavior or new information) and intrinsic pressure (e.g., reflection of one's own activity or a strong sense of purpose), based on the underlying and ambiguous wish (which is one of the sources of the decision and intention) through the interaction. Second, we experimentally investigated the facilitation by human in the decision-making situation. As a result, we could confirm that the facilitation could support human decision-making and we could classify the facilitation behavior into four typical categories. Next, to encourage an intrinsic activities, we proposed a method to estimate human preferential structure (emphasizing points) in human-agent interaction. This method was evaluated by using actual conversational agents in two situations; one-to-one interaction and one-to-group interaction. In both cases, we could confirm the method was helpful to support decision-making. Finally, we constructed a conversational agent which could support human decision-making based on the facilitation behavior which was an extrinsic factor and estimation of emphasizing points which was an intrinsic factor. We conducted an evaluation experiment and we suggested that the facilitative agent could support smooth decision-making and improve subjective impressions of decision-making processes. From these researches, we can suggest the decision-making support applied the bubbling intention model is helpful in human-agent interaction.

Chapter 12
Discussions

Abstract In this chapter, we discuss some high-level issues left beyond the scope of this book but deemed critical for future research. We first place conversational intelligence and conversational informatics in a larger picture of conversational knowledge circulation and social intelligence design to discuss issues from a wider perspective. We then single out ethical issues as centric to social issues and discuss the role of conversational informatics in the context of moral agents in human-agent symbiotic society. Finally, we come back to empathic agents as discussed in Chap. 1 and elaborate a road map for future study.

Keywords Conversational knowledge circulation · Community-maintained artifacts of lasting value · Social intelligence design · Ethical aspects · Empathy

12.1 Conversational Knowledge Circulation

Conversation is a powerful method that mankind has ever invented for communicating the meaning and expression of knowledge. Conversational knowledge circulation is an application of conversational informatics that employs conversational interactions as a primary means of communication to realize evolutionary collective knowledge in the society. In order for conversational knowledge circulation to function properly, social aspects need to be considered so that the technologies can be properly embedded into the society.

Conversational knowledge circulation depends on a method of capturing and presenting information in conversational situations (Nishida 2010b). Figure 12.1 shows a simplified view of how conversational knowledge circulation might be applied to an industrial environment, where communication among customers and engineers is critical. Emphasis is placed on enhancing the lower layers of the community knowledge process. It illustrates how conversations at the design, presentation, and deployment stages might be supported by conversational knowledge circulation.

At the design stage, the product is designed and possible usage scenarios are developed by discussions among engineers and sales managers. The discussions contain valuable pieces of knowledge, such as intended usage or tips, that may also

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_12

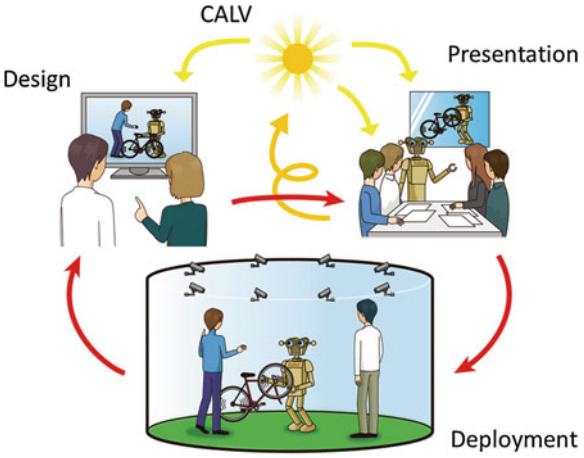


Fig. 12.1 Conversational knowledge circulation applied to industrial environment. Adapted from Nishida (2010b). © 2014, Toyoaki Nishida and At, Inc. 2014. Reproduced with permission

be useful to the users. Conversational content about the product and service can be composed as a result of the design phase. Conversational content may also be used as an additional information source in the fabrication phase to help developers understand the intention of the design.

In the presentation stage, the product and service are displayed to the potential customers in an interactive fashion. To make the interactive presentation widely available on the net, embodied conversational agents may be used as a virtual presenter. Embodied conversational agents will be able to cope with frequently asked questions using a collection of conversational content prepared in advance. When questions cannot be answered based on the prepared conversational content, the engineer may control the presenter agent as an avatar to create a proper reply. Such communication logs can be saved so that the service division may extend the "FAQ" conversational content for future questions. Embodied conversational agents may be used as a surrogate of the customer to ensure that the communication from the user is anonymous. The presentation stage can also be employed to train novices when the product and service is introduced to the user.

At the deployment stage, conversations may contain various pieces of knowledge sources, such as the real usage scenario, an evaluation from the user, complaints about the current service, and demands for new services. The conversation between the user and system engineer may be captured by intelligent sensing devices. Service robots may be deployed to help the user as well as to collect usage data. The collected conversational content may be fed back to the design phase for improvement and further product and service development.

Note that the collection of (potential) customers, salespersons, and engineers forms a community that shares a common product and service. *Community-maintained artifacts of lasting value* (CALVs) (Cosley et al. 2006) are expected to be created

as a result of the conversational knowledge circulation. The more information and knowledge is circulated, the richer the CALVs may be obtained.

The idea of a primordial soup of conversation discussed in Chap. 1 and its implementation presented in Chaps. 5–11 may be used as a basis for building a system of conversational knowledge circulation. Conversational content servers are needed to accumulate conversational content for distribution. Ideally, they may be equipped with a self-organization mechanism so that new conversation content may be automatically associated with an existing collection of conversational content, and the entire collection of conversational content may be organized systematically. A less ambitious goal is to provide a visualizer and an editor that allow the user to browse the collection of conversational content, organize it into topic clusters, and create new conversational content from the existing collection.

In addition to the above-mentioned basic elements, high-level functions may be introduced to allow the user to utilize the collection of conversation content in collaboration, discussion, and decision making. This issue has been addressed in social intelligence design, as discussed in the next section.

12.2 Social Intelligence Design

Social intelligence design is a field of research aiming at understanding and augmenting social intelligence based on a bilateral definition of social intelligence as an individual's ability to live in a social context and a group's ability to collectively solve problems and learn from experiences (Nishida 2007c, 2010b).

Previous studies on social intelligence design research focus on five topics. The first is about theoretical aspects of social intelligence design, involving understanding group dynamics and consensus formation of knowledge creation, theory of common ground in language use, and social learning. The second is about methods of establishing the social context by such means as awareness of connectedness, circulating personal views, or sharing stories. The third is about embodied conversational agents for knowledge exchange, mediating discussions, or learning. The fourth is about collaboration design by integrating the physical space, electronic content, and interaction. Multiagent systems might be used to help people in a complex situation. The fifth is about public discourse. Social intelligence design may be concerned with visualization, social awareness support, democratic participation, web mining, and social network analysis (Nishida 2001).

Further topics such as mediated communication and interaction (Fruchter et al. 2005), natural interaction (Nijholt and Nishida 2006), collaboration technology and multidisciplinary perspectives (Fruchter et al. 2007), evaluation and modeling (Miura and Matsumura 2009), ambient intelligence (Nijholt et al. 2009), designing socially aware interaction (Cavallin et al. 2010), and situated and embodied interactions for symbolic and inclusive societies (Katai et al. 2011) have been added to the scope in subsequent workshops.

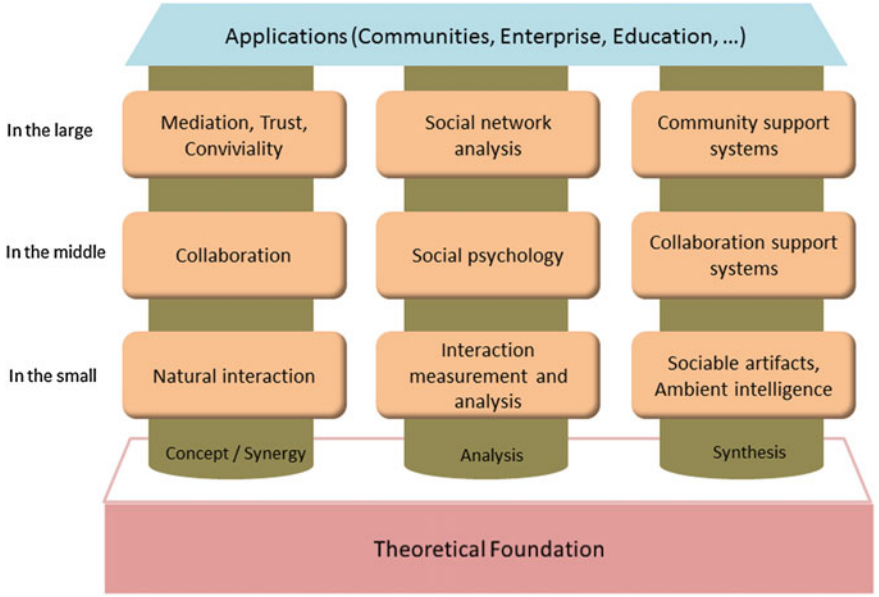


Fig. 12.2 Framework of social intelligence design

Social intelligence design can be discussed at different levels of granularity, as illustrated in Fig. 12.2. Social intelligence design at the macroscopic level is about social networking and knowledge circulation in a community. Social intelligence design at the mesoscopic level is about collaboration in small groups and teams. Social intelligence design at the microscopic level is about fast social interactions in a social discourse.

12.2.1 The Fast Interaction Loop on the Microscopic Level

Social intelligence design at the microscopic level is concerned with fast social interactions in the face-to-face interaction environment. It provides an opportunity to look at conversational interactions from a new angle, i.e., social intelligence and interaction. This is where conversational informatics comes in. Major issues other than those discussed in the context of conversational informatics include interactive social assistants, and social artifacts and multiagent systems.

Interactive social assistants help the user with social activities. S-Conart (Shoji and Hori 2005) supports conception and decision making of the user while online shopping. PLASIU (Shoji et al. 2009) is designed to support job-hunters' decision making based on observations from their actual job-hunting process. StoryTable (Gal et al. 2009) is a co-located cooperation-enforcing interface, designed to facilitate collaboration and positive social interaction for children with autistic spectrum disorder.

Social artifacts aim at embodying social intelligence to interact with people or other social agents. Xu et al. (2007) presented a two-layered approach to enhance the robot's capability of involvement and engagement. Xu et al. (2009) described a WOZ experiment setting that allows the mutual adaptation procedure between humans to be observed and understood. Mohammad and Nishida (2009b) presented NaturalDraw that uses interactive perception to attenuate noise and unintended behavior components of sensor signals by creating a form of mutual alignment between a human and a robot. Mohammad and Nishida (2009f) discussed combining autonomy and interactivity for social robots. Yamashita et al. (2006) evaluates how much a conversational form of presentation aids comprehension, particularly for long sentences and when the user has little knowledge about the topic. Poel et al. (2009) reported design and evaluation of iCat's gaze behavior. Nomura et al. (2006) studied negative attitudes toward robots.

Multiagent systems fully automate a computational theory of social agents. Roest and Szirbik (2009) showed an interaction-oriented agent architecture and language that makes use of an interaction pattern, such as escape/intervention. Rehm and Endrass (2009) integrated social group dynamics into the behavior modeling of multiagent systems. Mao and Gratch (2009) studied social judgment in multiagent systems. Pan et al. (2007) presented a multiagent-based framework for simulating human and social behavior during an emergency evacuation. Cardon (2001) argued that the emerging structure or the morphological agent organization reflects the meaning of the communications between the users.

12.2.2 The Structured Interactions at the Mesoscopic Level

Social intelligence design at the mesoscopic level is concerned with collaboration support in structured interactions of a group or team. Major issues include design and analysis of global teamwork, collaboration support tools, and meeting support and smart meeting rooms.

Design and analysis of global teamwork is a major concern in many industrial applications. Fruchter (2001) characterized collaboration support systems for global teamwork in terms of bricks (physical spaces), bits (electronic content), and interaction (the way people communicate with each other). Fruchter and Cavallin (2006) described a methodology for analyzing discourse and workplace in distributed computer-mediated interaction. Fruchter et al. (2007) formalized the concept of reflection in interaction during communicative events among multiple project stakeholders. The observed reflection in interaction is prototyped as TalkingPaper^{TM}.

Cornillon and Rosenberg (2005) investigated the conceptual design of a feedback advisor suggesting the knowledge co-construction aspect of a debate and notes that various aspects of social intelligence are coded into the dialogue, such as repetitions encoding awareness of connectedness. Cornillon and Rosenberg (2007) analyzed how people work together at a distance using a collaborative argumentation graph. They find that the number of turning actions (those changing the structure of an

argumentative graph) greatly varies between the face-to-face and remote conditions, while that of building actions (those contributing new information on the screen) does not.

In the network era, workplaces are enhanced with information and communication technologies. To enable people to flexibly interact with one another in a hybrid workplace, communication in the real-life workplace needs to be analyzed in terms of physical space, communication space, and organizational space (Rosenberg et al. 2005). People's behavior in coping with multitasking and interruptions in the workplace is studied in depth by Mark and her colleagues (González and Mark 2004; Mark et al. 2008; Su and Mark 2008).

Various collaboration support tools have been proposed to facilitate collaboration from different angles. Martin et al. (2005) identified story telling as a vehicle for tacit-to-tacit knowledge transfer in architectural practice and proposes the Building Stories methodology. Fruchter (2005) proposed RECALL, a multimodal collaboration technology that supports global team work. Heylighen et al. (2007) presented Dynamic Architectural Memory Online (DYNAMO), an interactive platform for sharing ideas, knowledge, and insights in the form of concrete building projects. Stock et al. (2009) presented a co-located interface for narration reconciliation in a conflict by making tangible the contributions and disagreements of participants and constraints imposed by the system to jointly perform some key actions on the story. Merckel and Nishida (2009a) presented a framework for situated knowledge management. A low-cost three-dimensional pointer allows the user to associate information with arbitrary points on the surface of physical equipment. Analysis is as important as synthesis. Pumareja and Sikkel (2006) studied the effects of long-term use of a groupware. The paradigm of social constructivism and the perspective of structuration are proposed as a framework of analysis. The findings from the case study suggest that collaboration technology can serve as a change agent in transforming the culture and structure of social interaction, through the various meanings people construct when interacting with technology and in benefiting from the structural properties of a system. Cavallin et al. (2007) investigated how subjective usability evaluation across applications can be affected by the conditions of evaluation and finds that scenarios not only affect the task-solving level but also prime the subjective evaluation of an application.

Meeting support and smart meeting rooms have large practical potential. Suzuki et al. (2009) discussed the social relation between the moderator and interviewees. Nijholt et al. (2006) described research on meeting rooms and its relevance to augmented reality meeting support and virtual reality generation of meetings. Reidsma et al. (2007) discussed three uses of Virtual Meeting Room: to improve remote meeting participation, to visualize multimedia data, and as an instrument for research into social interaction in meetings. Rienks et al. (2009) presented an ambient intelligent system that uses a conflict management meeting assistant. Wizard of Oz experiments are used to determine the detailed specification of the acceptable behaviors of the meeting assistant and to obtain a preliminary evaluation of the effect of the meeting assistant. Gill and Borchers (2003) studied use of interaction media.

12.2.3 The Networked Interactions at the Macroscopic Level

Social intelligence design at the macroscopic level is concerned with understanding and supporting communities where knowledge evolves as a result of interaction among members. Major issues include community knowledge management, and design and analysis of computer-mediated communication (CMC).

Community knowledge management is concerned with understanding and enabling an organizational approach to identify, foster, and leverage insights and experiences shared in a community. It should recognize best practice in a community (Davenport and Prusak 2000) and enhance the knowledge spiral between formal and tacit knowledge (Nonaka and Takeuchi 1995). CMC tools should be amalgamated with organizational structure and process. Tacit knowledge might be better formalized into formal knowledge with CMC tools with face-to-face communication functions, while formal knowledge might be better internalized into tacit knowledge with anonymous communication means (Azechi 2005). Caire (2009) pointed out that conviviality promotes values such as empathy, reciprocity, social cohesion, inclusiveness, and participation. Katai et al. (2007) introduced a framework of social improvisational acts toward communication aiming at creative and humanistic communities.

CMC tools support various phases of the knowledge process in a community. A corporate-wide meeting may not be possible without powerful CMC tools. FaintPop (Ohguro et al. 2001) is designed to provide social awareness. Nakata (2001) discussed a tool for raising social awareness through position-oriented discussions. Nijholt (2001) discussed the design of a virtual reality theater environment for a virtual community. At "World Jam," IBM's corporate-wide discussions held for three days and participated in by over 53,600 employees, a system called "Babble" was deployed that assisted synchronous and asynchronous text communications. Each participant was represented as a colored dot. The position of a dot within a visualization called a "social proxy" was designed to allow each participant to grasp who else is present and which topics are being discussed (Thomas 2001; Erickson 2009). In the DEMOS project, Survey, Delphi, and Mediation methods are combined to connect political representatives and citizens, experts and laymen. These methods are expected to strengthen the legitimacy and rationality of democratic decision-making processes by using CMC tools to inspire and guide large-scale political debates (Luehrs et al. 2001). Public Opinion Channel was proposed as a CMC tool for circulating small talk in a community (Fukuhara et al. 2001). Kanshin was designed to allow for extracting social concerns (Fukuhara et al. 2007). In order to cope with the digital divide, the culture of the user needs to be investigated with the greatest care and sensitivity (Blake and Tucker 2006).

CMC tools need to be analyzed for understanding and bringing about better community communication. In general, statistical or social network analysis may be applied to understand the structure and features of community communication (Fujihara 2001). Notsu et al. (2009) used visual assessment of clustering tendency (VAT) to analyze the balance of the network modeling of conceptualization. Miura

and Shinohara (2005) found that medium-density congestion with a relevant topic might activate communication by experienced participants in online chat and suggests the cognitive process in the course of communication congestion. Miura et al. (2006) suggested that information retrieval behaviors may vary depending on task-related domain-specific knowledge in information retrieval. If the retriever has sufficient knowledge, he or she will cleverly limit the scope of retrieval and extract more exact information; otherwise, s/he will spend much effort on comprehending the task-related domain for efficient retrieval. Matsumura et al. (2005) revealed that the dynamic mechanism of a popular online community is driven by two distinct causes: discussion and chitchat. Hofte et al. (2006) investigated place-based presence (presence enhanced with concepts from the spatial model of interaction). The lessons learned include the following: place-based presence applications should be designed as an extension of existing PIM applications to allow people to control the exchange of place-based presence information; a place-based presence system should keep the user effort to a minimum, since trust in presence status may be lowered otherwise; and wider presence and awareness scopes may be needed to allow people to see each other since they will easily lose track of each other otherwise. Morio and Buchholz (2009) made a cross-cultural examination of online communities in the USA and Japan, and found that Japanese people would prefer to discuss or display their opinions when there is a lack of identifiability, while US people feel a much lower need for anonymity. Furutani et al. (2009) investigated the effects of internet use on self-efficacy. The results suggest that a belief of finding people with different social background may positively affect self-efficacy (the cognition about one's capabilities to produce designated levels of performance), while staying in low-risk communication situations with homogeneous others might undermine self-efficacy. Moriyama et al. (2009) studied the relationship between self-efficacy and learning experiences in information education. They suggest that self-efficacy and information utilization abilities may enhance each other. In addition, creativity and information utilization skills might promote self-efficacy.

One of the essential issues for social intelligence design at the macroscopic level is a moral theory that may be used by participants to negotiate either explicit or implicit social conflicts. Ethical issues in human society have a long history of discussions. In contrast, discussions have just started on ethical issues in the forthcoming symbiotic society where robots and other forms of autonomous agents (hereafter, simply "autonomous agents") play an integral role in the society.

12.3 Ethical Aspects

As autonomous agents become increasingly popular around us, the boundary between decision making and the simple pursuit of tasks becomes more blurred. Nishida (2009) suggested that we need to consider a serious trade-off between convenience and ethical decision making. Even though the relationship between humans and artifacts are not as intimate as in the case of Brain Computer Interfaces (BCIs)

(Tamburrini 2009), the more the agent becomes akin to humans, the more ethical problems may arise at the mental level, as pointed out by Turkle (2011). Several ethical concerns come in here. First, developers should design autonomous agents so that they will not harm humanity. Second, developers should design autonomous agents so that they do not take advantage of their superiority in perceptual or intellectual capabilities. This principle will make sense when autonomous agents are equipped with high-resolution sensing techniques that can read the private internal mental status of the user, for such a high-performance sensing technique might be abused by a salesperson to deceive a customer. In particular, privacy should be protected in a similar vein, for it may be easily disclosed by the powerful sensing mechanisms of autonomous agents (Carew et al. 2008).

The most fundamental approach to the above problem is to endow autonomous agents with abilities to communicate with humans and other citizens in the symbiotic society. What kind of ethical principles should underlie human-agent communication?

In human society, the discussion of ethical principles has been primarily concerned with protecting the autonomy and dignity of humans. Kant's formula of humanity (Kant 1788) implies the ethical inadmissibility of humans being used solely as a *means* to achieve an *end* (Decker 2008). This ethical principle can be construed as entailing the norm of protecting the weak from being used solely as a means to an end by the strong.

For ethical principles to be effective, they should be interpreted as entailing moral rules of behavior to be executed by humans, provided that humans are to have an exclusive right and responsibility for that execution. However, this approach will lead to several difficulties. Unethical people may simply neglect ethical principles. Even for those trying to comply with ethical principles, it is often unclear whether an action may violate some. In addition, humans are fallible; they are not free from contradictions and dilemmas, which makes the pursuit of ethical behavior even harder. In reality, humans may often behave unethically, intentionally or unintentionally. Furthermore, the strong may not be well aware of the consequences of their dominance over the weak. For example, in hospitals it is often pointed out that doctors might not be aware of the pain of the patients and treat the patients as if they were objects; in schools, teachers might not be aware of students not understanding the subjects; at home, parents might not be aware of the frustrations of their children; and in care-giving situations, caregivers may not be aware of the lack of freedom of elderly people, to name just a few.

As new technology is accompanied by the introduction of new criteria for the weak and the strong, we cannot possess a priori knowledge for distinguishing who are the weak and who are not. Although autonomous agents are often taken as a threat to humanity, it should be noted that there are many reports of older people being mistreated by humans. Some of these events are treated as crimes, while others are regarded as a failure of ethical and respectful treatment. Thus, ethical principles become much harder to apply than before, as the environment surrounding us becomes more complex. Ethical principles will be in crisis unless they are applied in real life.

Conversational intelligence may be designed to augment *ethical intelligence*. It is also interesting to discuss whether the schema also applies to service domains in general by replacing "care" with "service." Whether service-providing people are replaced by autonomous machines in general is not clear, but in reality, there are already many examples in modern society where services are given by machines, for example, music boxes, vending machines, industrial robots, auto-drivers, sushi robots, and sushi-go-round.

A key idea for using autonomous agents to facilitate the following of ethical principles by users is to have autonomous agents mediate communication between the service providers and recipients so that ethical awareness of the service recipient will be fed back to the providers. This is a use of autonomous agents as an aid to encourage ethical behavior. We are not proposing to build a strong ethical intelligence, an autonomous agent that can actually make ethical decisions or enforce people's behavior in compliance with ethical codes. We do not challenge the conclusion of Torrance (2008) that robotic agents will never be capable of being taken as moral beings of the first order, due to their lack of organic embodiment. Instead, we propose to work seriously on building a *partially ethical agent* or an autonomous agent with weak ethical intelligence that will provide ethical information or awareness. A partially ethical agent is more like an intelligent information retrieval system, deemed to be heuristic in nature, not deemed to be complete or perfect.

For example, in schools such a partially ethical agent will invoke questions on behalf of students in addition to conveying personalized pedagogical information. It will inform teachers of exactly what and how students misunderstand and what kinds of frustration they are suffering from in the learning environment.

Androids might be used as a vivid conveyer of the service recipient's ethical experiences. A good example is the medical simulation humanoid robot named Simroid (ABC News 2007, Nov. 28). Simroid can demonstrate pain when it is mistreated by a clinical dental trainee in a training course. Even though the pain is mechanically created by simulating a real person's behavior, human neural mechanisms, presumably the mirror system, allow humans to feel pain even if they only look at painful expressions.

The role of the partially ethical agents in the above examples is twofold. The first is to demonstrate how people may come across difficulties so that their colleagues can understand the situation. This should be effective in protecting the weak, when the strong ignore the difficulties of the weak. Even in normal contexts, they should draw attention whenever they run into unethical activities. The second is to demonstrate what the problems are and what actions might be taken in order to prevent them from happening. Both will be effective in generating ethical awareness. Techniques for weak AI, such as case-based reasoning, may be used to retrieve a relevant case and adapt it for the current situation.

Consider the doctor-patient scenario, where a doctor is the strong and the patient the weak, in terms of knowledge and authority regarding medical care. A doctor might be interested in introducing a partially ethical agent that stays with the patient not only to monitor his or her health condition and recommend him or her what to do and not to do at each time but also to communicate with the patient regarding

health care and a future plan. Both the patient and the doctor might be happy with the situation, for the partially ethical agent may not only realize the every-time medical care with little overhead but also bring about better communication on an equal footing.

In the teacher-student scenario, the teacher is the strong and the student the weak, in terms of knowledge and experience regarding the subject. A teacher might be interested in introducing a partially ethical agent to provide an interactive tutoring service as well as to get feedback about how much the student understands the subject. The student might gain the freedom of learning, while the teacher might gain better awareness of the student's situation.

In the parent-child scenario, the parent is the strong and the child the weak, in terms of experience, physical power, and financial capability. A parent might be interested in introducing a partially ethical agent with weak ethical intelligence to provide better support as well as to get feedback about how much physical or mental frustration the child is suffering due to the parent. The child might gain freedom in activities, while the parent might gain better understanding of his or her family.

In the caregiver-elderly person scenario, the caregiver is the strong and the elderly person the weak, in terms of physical and perceptual ability. A caregiver might be interested in introducing a partially ethical agent to deliver better care as well as to get feedback about how much physical or mental frustration the elderly person is suffering due to the caregiver. The elderly person might gain freedom of life, while the caregiver might gain better understanding of the elderly person.

Although building partially ethical agents with weak ethical intelligence should be easier than building autonomous agents with strong ethical intelligence as they need not be able to distinguish good from bad, or ethical from non-ethical, it should be beneficial for us to have a detailed and clear image of a partially ethical agent. In order to be partially ethical, autonomous agents need to meet several requirements. First, they need to be fluent in communication; they need to interact with people who solicit, monitor, and affect their intention. At the very least, the partially ethical agents need to convince people either verbally or nonverbally that they have goodwill. Second, they should not scare people or other social or biological agents. Third, they should be able to be aware of the emotional status of people and create emotional responses for better communication. Fourth, partially ethical agents should be able to provide useful information for ethical decision making.

Difficulties in meeting the above mentioned requirements may depend on what is shared between humans and robots. The first and most basic level is the physical and informational level. From the viewpoint of environmental ethics, it is important to know that the shared environment has only limited resources and that unconstrained pursuit of personal benefit may result in the tragedy of the commons (Hardin 1968). Partially ethical agents may play an active role in this context so long as regulations and ethical codes can be reflected in their programming so that they do not allow the user to violate those regulations and codes or so that the user's behavior can at least be recorded for later review. In other words, autonomous agents can be regarded as a medium that connects people.

Most problems at this level might be handled pragmatically and practically by combining socio-ethical and technical considerations. Incorporating faithful communication into partially ethical agents might be the key issue to compensate for the intellectual superiority of partially ethical agents. Even though partially ethical agents cannot be perfect, issues of liability and responsibility might be resolved by damage insurance, after the best efforts are made. The better the performance partially ethical agents exhibit, the lower the insurance premium required.

The second level is concerned with perception and cognition. It forms the ground of our epistemology that in turn serves as a ground of ethical decision making. In principle, agents need to share the structure and properties of embodiment in order to share this space, for perception and cognition are formed through learning with an embodiment. Without the same embodiment, the different agents may not be able to infer the emotions caused by the given situations and may have difficulties communicating their emotions with each other. Although they can simulate emotions by simulating the sensory-interpretation-motor mechanism, it may not be evident whether they share perception and cognition.

The third level is concerned with value. Although an extensive discussion of this issue is beyond the scope of this book, I suspect that the sense of value originates from the origin of the species, and consequently, it is very challenging to share this space with artificial agents. Even though we try to program our space of value into autonomous agents so as to make ethical decisions based on the space of value, it may not effectively work, for there are infinitely many situations and we will not be able to foresee all of them in advance.

The moral being at each level needs to be able to ground decisions on its own experiences and explain its decisions at that level, based on ethical principles. Building a moral being on the first level might be quite feasible using existing methods in artificial intelligence or cognitive technologies. Ontology will help define concepts at this level. Symbolic knowledge representation can be used to code ethical logic. Machine learning methods can be used to help in programming or enable the resulting autonomous agent to adapt to the given environment, with shared understanding of the risk.

In contrast, it is less feasible to build moral beings on the second or third levels. A moral being at the third level may not be feasible at all. A moral being on the second level can only be built with the perception and cognition of simulated embodiment. Even though the embodiment may be similar on the surface, it significantly differs beneath the surface. Although the resulting ethical decision making may be limited, it would be both beneficial and challenging to implement autonomous agents compliant with ethical principles (Rosenberg 2008) and ethical codes (Nagenborg et al. 2008). Another challenge is to implement a mechanism for grounding linguistic expressions based on simulated perception and cognition.

An interesting question remains concerning whether we will eventually advance beyond weak ethical intelligence to implement an autonomous ethical agent. Calverley (2008) argued that a non-biological machine having legal independence is theoretically possible. Wallach (2008), Wallach et al. (2008), Whitby (2008) discussed issues to be addressed in order to implement artificial moral agents.

12.4 Empathy

As discussed in Chap. 1, empathy is key to harmony in the symbiotic society. Let us consider a situation illustrated in Fig. 12.3, where the person on the right (A) talks about a difficulty and her friend (B) expresses concern, a typical scene of empathy. There may be a wide variety of possible causes, ranging from the physiological to the ethical. The response of the friend might have come out of mere politeness without accompanying much thought. On the contrary, it might have come from concern at a much deeper level resulting from compassionate thoughts. For example, B may have imagined what has happened from what she has heard from A and felt the way A might have felt (Fig. 12.4). In order to reach this level, one needs to sense and interpret social signals or subtle cues caused by a partner (Pentland 2008). According to theory of mind (ToM) (Baron-Cohen et al. 1985), people normally attribute mental states, such as desire or intention, to interpret and predict the behavior of other people. In general, reasoning about a mental process needs to be supported by background knowledge. How can one acquire this knowledge?

According to the *theory theory* hypothesis in cognitive development (Gopnik 2003), children develop their everyday knowledge of the world in the same way as adults develop scientific theories. A computational model for this hypothesis might be realized with a classic framework of symbolic knowledge-based abductive reasoning that will maintain one or more line of defeasible explanation for the observed behaviors of a partner, supported by the given knowledge base (Fig. 12.4). Unfortunately, however, this approach may run into difficulties when little evidence is available for choosing among many possible interpretations.

The *simulation theory* suggests that a human uses her/his own embodiment and mental model to simulate the behavior and mental processes of other people to understand them. For example, B may want to examine each step of the story she guessed by replacing the image of A by herself in the story, using a model about herself. B's nervous system centered on the limbic system will allow her to experience emotional reactions similar to what A has experienced (Damasio 1994).

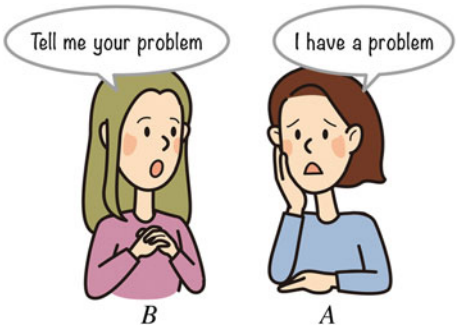


Fig. 12.3 Empathy is in our daily conversations. Drawing inspired by Nishida (2013). © 2014, At, Inc. Reproduced with permission

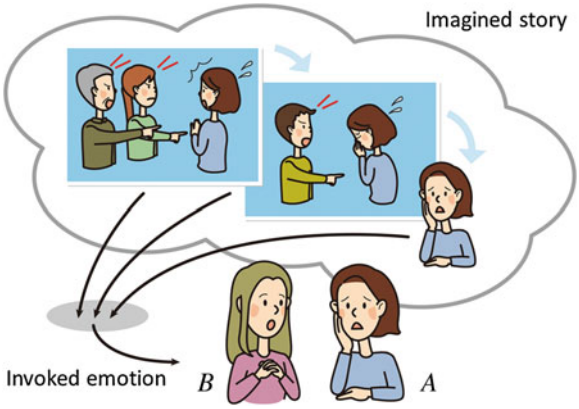


Fig. 12.4 Imagine what has happened from what is heard. *A* might have been discouraged and wanted to talk with *B*, for she was criticized not only by colleagues but also by her boss. Drawing inspired by Nishida (2013). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

Recent findings of neuroscientists suggest that *mirror neurons* help us understand the emotions of other people by some form of inner imitation. Mirror neurons fire both when one performs an action and when one sees that action. They were first found in the ventral premotor cortex (area F5) of the macaque brain and then in the human brain (Rizzolatti and Sinigaglia 2008). Gallese et al. (2007) suggested that intentional attunement or embodied simulation enabled by the dynamics of our embodiment and neural system, including mirror neurons, might cause empathy. Iacoboni (2008) suggested that mirror neurons help us understand the mental state of other people by making some form of inner imitation to pretend to be "in other people's shoes" (the mirror neuron hypothesis of empathy).

Nishida (2013) suggested that the computational model which reflects neuroscientists' arguments on empathy might be the one shown in Fig. 12.5. The architecture of an individual person's cognition may include an internal theater. In addition to the routine work of human information processing that takes information from receptors and produces motor commands, the input is sent to the internal theater where what is happening is reproduced for, e.g., reflection and planning. The communication partner's behavior is reproduced with the assistance of mirror units in the internal theater and interpreted by using the actor's own mechanism for generating emotional appraisals.

Unfortunately, the above framework does not retain all the advantages of symbolic knowledge-based abductive reasoning. In particular, it does not allow one interpretation to be contrasted against another, which is often considered useful in figuring out the most plausible interpretation.

After all, the theory theory and simulation theory components need to be integrated into a coherent mechanism as shown in Fig. 12.6, where the two components play roles that are complementary to each other. The theory theory component will handle hypothesis making and reasoning at the cognitive level, while the simulation theory

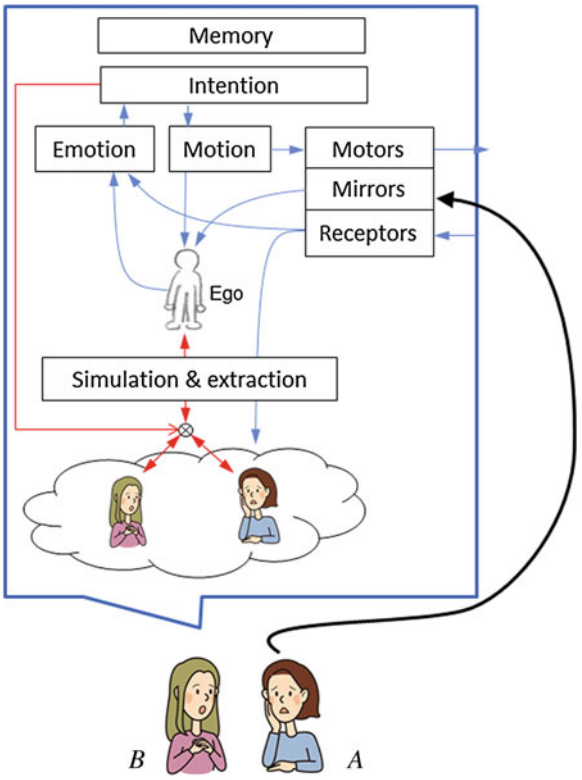


Fig. 12.5 The computational model that reflects neuroscientists' arguments on empathy. Drawing inspired by Nishida (2013). © 2014, Toyoaki Nishida and At, Inc. 2014. Reproduced with permission

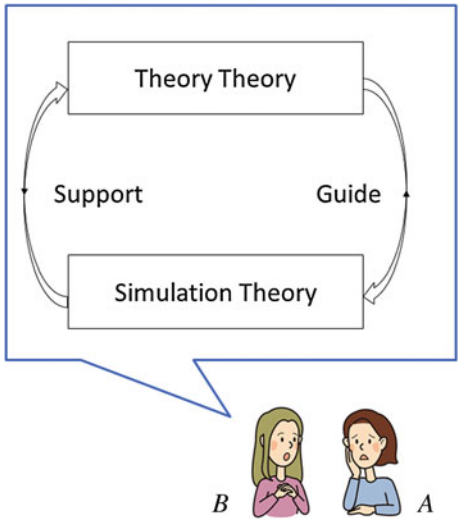


Fig. 12.6 An integrated model for empathy. Drawing inspired by Nishida (2013). © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

component carries out simulation with imaginary embodiment and emotions to evaluate hypotheses. The theory theory component guides and controls the simulation theory component, while the latter provides information about the appraisal of the given hypothesis based on the agent's own embodiment and physiology.

Even though people are gifted with a native mechanism for empathy, empathy may still not result because of blocking factors. First, the universe of discourse may not be established between the participants, preventing one from figuring out what has happened to the partner. Establishing the shared universe of discourse is not always trivial. When the participants are talking over the phone, for example, it is often difficult for them to reconstruct the shared discourse, probably because the subject is not amenable to verbal description or because they cannot share the environment to which they may refer to communicate their thoughts. Even in face-to-face conversation where the environment of the conversation is shared, the participants may have difficulty in interpreting what is said about objects or events due to their lack of shared background.

Second, empathy might be blocked by the failure to share a first-person view with other people. It may prevent one from perceiving the universe of discourse from different perspectives and resolving conflicts caused by the discrepancy of the perceived universe. For example, you might not be able to feel how children experience the world unless you look at the world at the same altitude as their eyes; you might not understand the difficulties of people who have lost their sight until you try to move around a hazard with your eyes closed.

Third, inferior knowledge or skill level might not allow one to interpret meaning in the same way as an expert. Even though one may enjoy the play by an expert, he or she might not able to perceive or be aware of events that make sense to experts nor affect the world in the way experts do.

Fourth, empathy may be prevented by differences in communication style, or the way intentions are encoded into social signals ranging from the way verbal and nonverbal signals are coordinated to constitute an utterance, to the way the discourse is structured to satisfy the speaker's intention. Unlike behaviors whose meaning can be inferred on anthropological or ethological grounds, the meaning of rituals (Goffman 1967) cannot be understood without being taught by a member of the community because they result from arbitrary choices the community has made from multiple alternatives. Failure to share communication style and rituals will not allow one to exchange clearly defined messages with communication partners, or to predict how other agents may behave in a given communication environment, resulting in blocking empathy. Mishaps might be observed in intercultural communication where social signals generated by the speaker are interpreted by the hearer in a different way, which might cause double bind and distress the hearer.

Finally, empathy may not be induced if there is a discrepancy in the way the value is determined for events and objects. According to cognitive appraisal theory (Ortony et al. 1988), emotion is determined by evaluation of incoming events according to the value system. At higher levels of the mental process, the value system is used to make decisions. Empathy might not be achieved if the observed behaviors do not follow the value system believed or at least approved by the observer.

In order to build an empathic agent, it is evident that we need to go beyond building a mindless, intelligent agent, closer to building an autonomous agent that can be perceived as possessing its own mind to sense other minds. At least, our target should possess the sense of itself, or the ego sense, as a mind may be grounded on an ego sense that uniquely resides in each individual, produces the sense of self, and serves as a reference to sense other individuals with an ego sense. In addition, such an autonomous agent with ego sense should be able to reason about how people and other autonomous agents with ego sense might behave under given circumstances.

It appears that the more technology removes these blocking factors and the more is shared among the participants, the more empathy is gained (the sharing hypothesis (Nishida 2013)). Information and communication technologies have a large potential for bringing about various kinds of sharing that have not existed before the information age. The Internet and web technologies have brought about significant impacts on helping people share the universe of discourse, first-person view, knowledge and skills, the communication style and rituals, and the value system. The sharing hypothesis emphasizes the potentials that technology may realize in the future, rather than enumerating a list of blocking factors. Indeed, we have a long list of challenges.

After all, the empathic agent, if completed, will have a fairly complex structure as shown in Fig. 12.7. The components for empathic agents should comprise not just problem-solving intelligence connected to the external world through sensors and motors, but also a fully embodied artificial mind containing components such as perception, cognition, language, judgment, memory, imagination, and consciousness

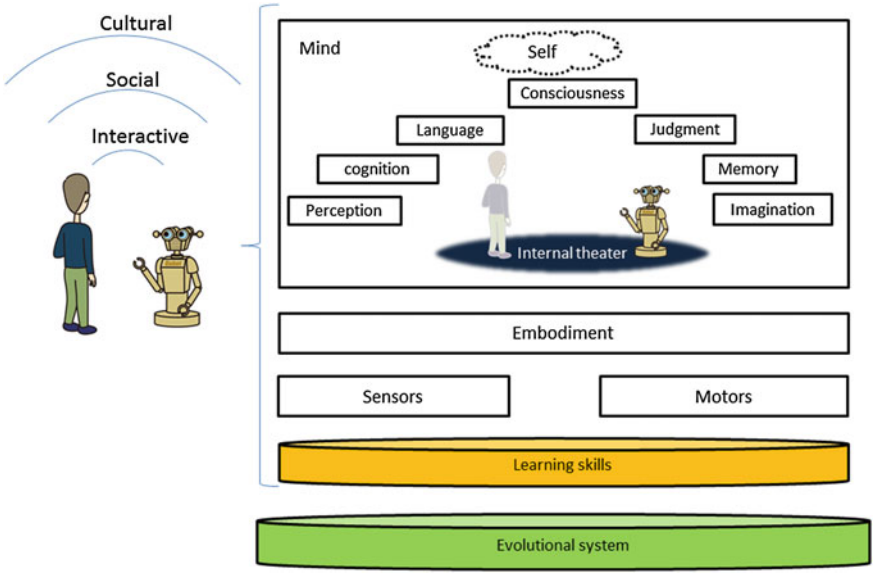


Fig. 12.7 The prospective architecture of empathic agent. © 2014, Toyoaki Nishida and At, Inc. Reproduced with permission

reflecting the sense of self. In order to be able to evolve itself as it learns, the empathic agent needs to be supported by learning skills and an evolutionary process. We believe that some kind of mental device that may be called the internal theater may be mandatory for the empathic agent not just to behave reactively but also to behave proactively based on reasoning about the state of the world and the mental states of other actors cohabiting with it, to share common ground for interaction in the social and cultural context.

12.5 Summary

In this chapter, we have discussed high-level issues left beyond the scope of this book. We have started with arguing for the use of conversational intelligence as a means for building socially intelligent systems. We believe that conversational informatics plays a key role in constructing conversational knowledge circulation whose goal is to build, maintain, and harness community-maintained artifacts of lasting value. Towards this end, we need to fit the whole enterprise in a larger picture. Social intelligence design is a field of research aiming at understanding and augmenting social intelligence based on a bilateral definition of social intelligence as an individual's ability to live in a social context and a group's ability to collectively solve problems and learn from experiences. We have surveyed previous work in social intelligence design at the three levels ranging from the micro to the macro. We need to be sensitive about the ethical aspects of the technology, not only to protect but also promote humanity. We believe that one of the important feasible goal of conversational informatics to build a partially ethical agents. Finally, we have come back to empathic agents. We elaborated our thoughts on our perspectives for empathic agents.

Chapter 13
Conclusion

This book is the first systematic presentation of conversational informatics. It not only compiled the major outcomes resulting from research and development activities of our group, it also identified the foundations on which we have been relying so far as well as potential directions of future research on this subject. Topics are laid from the fundamentals to the advanced technical issues followed by discussions about future work.

In the introductory chapter, we characterized conversational informatics as a field of research focusing on communicative aspects of intelligence, aiming at understanding and augmenting conversations—a fundamental human activity. In contrast to conventional research on artificial intelligence oriented towards autonomous intelligence, conversational informatics attempts to bridge human society and computational intelligence. We identified our engineering goal as building empathic agents that can dynamically establish empathic relationships with humans and other agents by accumulating conversations. We argued that building an ever-evolving primordial soup of conversations, an ensemble of mechanized humans and human-like machines, by exploiting abundance of data rapidly growing to cover a wide spectrum of our conversational behaviors is a promising approach, if we can rely on the sharing hypothesis—the more common ground is shared, the more empathy is gained. Evolution of a primordial soup of conversation will be supported by synergy of the common ground and conversational intelligence. A strong engagement by participants is deemed a key to success.

In order to build a good engineered system, we need to know the nature of the phenomenon into which our artifact is to be incorporated. To overview existing work on conversation, we employed five viewpoints to overview a vast collection of previous work, namely, verbal communication, nonverbal communication, social discourse, narratives and content flow, and cognitive processes. We believe that such topics as memes, narratology, discursive psychology, and social constructionism are relevant to conversational informatics, as storytelling aspects of conversation is quite important to characterize the role of conversation in our social life. Goffman's seminal work on unfocused and focused interactions in gathering helps us delineate the outer

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2_13

appearance of conversation. Types of participation introduced by Goffman allows us to observe conversation in a structured fashion. Focused gathering comprises a core of conversation. Studies on verbal interactions belongs to the domain of language use. Our approach draws on Clark's joint activity theory brings about a rich repertoire of concepts, such as levels, tracks and layers, for interpreting activities in conversation. Studies on nonverbal communication centers on the identification of social signals and their interpretation. We need to know cognitive aspects, such as theory of mind and emotions, to understand what is happening beneath the surface of conversations.

Our literature study on the engineering aspects consists of two parts: a bird's eye view on the historical development of conversational systems and technology-oriented survey. Our historical survey identifies three lines of research following the initial success of natural language question answering systems in the 1960s. The first is directed towards development of interactional systems in search for better human-computer interface that aims at providing natural user interfaces. Speech dialogue systems, multimodal interfaces and embodied conversational agents or intelligent virtual agents have been developed in this vein. The second line of research is on transactional systems, story understanding and generation shed light on the content. Knowledge based methods and the dynamic memory model were proposed. Third line of research is on cognitive systems oriented toward realization of lifelikeness, emotion, intentionality and agency.

On the other hand, our technology-oriented survey highlighted five aspects: the architecture, scripts and markup languages, corpus-based approaches, behavior learning using machine learning techniques, and the evaluation methodology. We made a rather comprehensive list of components of a full-fledged conversational system. We pointed out that a generic control structure such as the blackboard system is mandatory to support a complex structure of invocation and coordination among components. We described how scripts and markup languages are used to specify behavior of conversational systems. Standardization allows researchers to share data and codes. We described a corpus-based approach that is mandatory to base target behaviors of conversational agents on actual conversations among people. We presented a mathematical framework of machine learning techniques that is applied to corpora to identify basic communication acts from a large amount of noisy data. We discussed that the importance of evaluation of conversational systems for solid understanding of technology. Techniques such as the implicit association test allows for estimating participants' implicit attitudes in evaluation.

The journey into the technology we developed starts with a conceptual introduction of conversation quantization. There are four main ideas underlying conversation quantization: conversation quantum as a package of information for describing the meaning and expressions of a significant segment of conversation, the use of conversation quantum as a component of complex conversational scenes, generic procedures for obtaining, manipulating, and utilizing conversation quanta, and amenability of implementation at different levels of automation. The actual implementation may vary depending on granularity, depth and breadth of annotation, representational fidelity and generality. We discussed four aspects of conversation quantization: representation, production and consumption, manipulation, and circulation. The

application of conversation quantization centers on the implementation of a shared conversation space using the mixed reality technology. We discussed shared virtual meeting space, virtual interaction game, and tele-presence as typical examples. Finally, we included a historical note, as the idea of conversation quantization has slowly evolved over a decade.

We then presented the smart conversation space. An open conversation space enhanced by augmented reality and an immersive conversation space are used to project situations and activities in the shared virtual space. The situated knowledge media is an early implementation of the smart conversation space to enhance online customer service not only to benefit information consumers but also empower information producers. 3DCCbyMK is a three-dimensional conversation capture that uses the Kinect technology to measure and record the entire conversation space including multiple persons and the surroundings. ICIE implements the idea of an immersive conversation space that can augment, record, and measure conversation in a fully controlled immersive interaction environment. DEAL is a software platform that allows for extension with function plugins and control jacks. We also presented three implemented scenarios using ICIE: the filming robotic agent scenario, the cooperative multiagent interaction scenario, and the tele-presence scenario.

Conversational informatics relies on visual techniques for recognizing human behavior and producing character animations. The first topic was facial expressions. On the recognition side, we showed several techniques of face detections by combining the Haar-like features and the weak-classifier. On the synthesis side, we introduced the two types of facial expression and synthesis techniques: the semantic/symbolic descriptor-based based on the descriptors of facial parts movements such as FACS and FAP, and the data-driven approaches using the 2D/3D database of facial movements. The semantic/symbolic descriptor-based approach is simple and suitable for real-time/mobile systems, while the resulting picture is not realistic. In contrast, the data-driven approach can generate photo-realistic pictures but requires more computational time and large size of data. We also surveyed gesture recognition and synthesis techniques. The major techniques for gesture recognition are characterized in terms of feature vectors and gesture recognition methods. Similar to facial expression synthesis methods, we classified approaches to gesture synthesis into the symbol-based and the data-driven, and surveyed major techniques.

Multi-modal interaction analysis is mandatory to better understanding of conversation and the case studies. Following a generic introduction to designing experiments for analysis of multi-modal interaction, we explained two advanced techniques: collaborative allocation for reliable and efficient corpus building and physiological signal analysis for obtaining objective estimation of mental states. Then, we presented three case studies. The first case study is about the use of physiological-signals based metric for measuring naturalness in conversational contexts. The second case study is about measuring social atmosphere by combining visual information and physiological indices. The third case study involved the analysis for extracting evaluation criteria for ballroom dance.

In order to benefit from abundance of data, we need a strategy. Based on the analysis of existing research in cognitive science, neuroscience, psychology and

developmental studies, we argued for an intimate relation between imitation as a learning strategy, simulation as a behavior generation mechanism and interaction or conversation as basic constructs of human sociality. Based on this argument, we proposed an architecture for conversational agents that utilizes simulation as the main building block of its behavior generation allowing it to combine seamlessly information from lower and higher layers in the cognitive hierarchy. The proposed architecture also utilizes imitation to learn the computational processes. Our SILI framework centers on the connection between imitation, simulation and conversation. Simulation is used as the basis for agent's behavior generation while imitation is utilized to develop the internal processes of the agent that are used by the simulation engine. We have presented the SILI architecture consisting of interaction perception processes, perspective taking processes, forward basic interaction acts, reverse basic interaction acts, interactive control processes, shared variables, mirror trainer, interaction structure learner and interactive adaptation manager. We have shown key algorithms.

We then reported a couple of case studies. The first case study was concerned with learning to control the eye-gaze using the SILI architecture. It was shown that this approach can outperform a carefully designed gaze controllers and user expectations. The second case study was conducted in search for a more human-oriented form of learning from demonstration. We discussed imitation in social context that allows a robot to learn throughout their lives not only from explicit teaching but also be just watching what humans do in their environment more like children. Fluid imitation, as we proposed, uses the same building blocks of SILI (namely, motif discovery, change point discovery and causality analysis) but it does not utilize the simulation theoretic behavior generation mechanism, but it tries to leverage the ability to interact with people.

Human-agent interaction will be better supported if the agent can estimate the cognitive process beneath the surface of interaction and reflect it in interactional behaviors. Towards this end, we first conducted a general discussions on facilitating agents. We introduced a bubbling-intention model, which suggests that the decision or intention is extemporarily shaped by an extrinsic stimulus (e.g., a partner's behavior or new information) and intrinsic pressure (e.g., reflection of one's own activity or a strong sense of purpose), based on the underlying and ambiguous wish (which is one of the sources of the decision and intention) through the interaction. Then, we experimentally investigated the facilitation by human in the decision-making situation. As a result, we could confirm that the facilitation could support human decision-making and we could classify the facilitation behavior into four typical categories. We proposed a method to estimate human preferential structure (emphasizing points) in human-agent interaction. We constructed a conversational agent that could support human decision-making based on the facilitation behavior which was an extrinsic factor and estimation of emphasizing points which was an intrinsic factor. The experimental evaluation suggested that the facilitative agent could support smooth decision-making and improve subjective impressions of decision-making processes.

Finally, we discussed high-level issues left beyond the scope of this book. We argued for the use of conversational intelligence as a means for building socially intelligent systems. To fit the enterprise of conversational informatics in a larger context, we surveyed previous work in social intelligence design to obtain potential demands for further development. We discussed the ethical aspects of the technology and identified building partially ethical agents as one of the important feasible goals of conversational informatics. Then, we came back to empathic agents and elaborated our thoughts on our perspectives for empathic agents.

References

Abbeel P, Coates A, Ng AY (2010) Autonomous helicopter aerobatics through apprenticeship learning. Int J Robot Res 29(13):1608–1639

Aleotti J, Caselli S (2008) Grasp programming by demonstration: a task-based quality measure. In: Proceedings of the 17th IEEE international symposium on robot and human interactive communication (RO-MAN 2008), pp 383–388

André E, Rist T (1995) Generating coherent presentations employing textual and visual material. Artif Intell Rev 9(2–3):147–165

Andre-Obrecht R (1988) A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Trans Acoust Speech Sig Process 36(1):29–40

Argall BD, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robot Autonom Syst 57(5):469–483

Argyle M (2001) Bodily communication, 2nd edn. Routledge, London

Atienza R, Zelinsky A (2005) Intuitive human-robot interaction through active 3D gaze tracking. In: Dario P, Chatila R (eds) Robotics Research, Springer Tracts in Advanced Robotics 15, pp 172–181

Austin J (1962) How to do things with words. Harvard University Press, Cambridge

Aydogan R, Yolum P (2007) Learning consumer preferences using semanticsimilarity. In: Proceedings of the 6th international joint conference on autonomous agents and multiagent systems, AAMAS '07. ACM, New York, pp 1–8

Aylett R, Louchart S, Dias J, Paiva A, Vala M (2005) Fearnot!: an experiment in emergent narrative. In: Panayiotopoulos T, Gratch J, Aylett R, Ballin D, Olivier P, Rist T (eds) Intelligent virtual agents. Springer, Berlin, pp 305–316

Aylett R, Paiva A (2012) Computational modelling of culture and affect. Emot Rev 4(3):253–263

Azechi S (2005) Informational humidity model: explanation of dual modes of community for social intelligence design. AI SOC 19(1):110–122

Ball G, Ling D, Kurlander D, Miller J, Pugh D, Skelly T, Stankosky A, Thiel D, Van Dantzich M, Wax T (1997) Lifelike computer characters: the persona project at microsoft research. In: Bradshaw JM (ed) Software agents. AAAI/MIT Press, Menlo Park, CA

Baron-Cohen S (1995) Mindblindness: an essay on autism and theory of mind. MIT Press, Cambridge

Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a 'theory of mind'? Cognition 21:37–46

Basseville M, Kikiforov I (1993) Detection of abrupt changes. Printice Hall, Englewood Cliffs

Bates J (1994) The role of emotion in believable agents. Commun ACM 37(7):122–125

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2

Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: Proceedings of the computer vision–ECCV 2006. Springer, pp 404–417
Becker-Asano C (2008) WASABI: affect simulation for agents with believable interactivity. IOS Press, Amsterdam
Billing EA (2010) A formalism for learning from demonstration. Paladyn 1(1):1–13
Biscaldi M, Rauh R, Irion L, Jung N, Mall V, Fleischhaker C, Klein C (2013) Deficits in motor abilities and developmental fractionation of imitation performance in high-functioning autism spectrum disorders. Eur Child Adolesc Psychiatry 1–12
Blake E, Tucker W (2006) User interfaces for communication bridges across the digital divide. AI SOC 20(2):232–242
Blumberg BM (1997) Old tricks, new dogs: ethology and interactive creatures. Unpublished doctoral dissertation, Institute of Technology, Massachusetts
Bolt RA (1980) "Put-that-there": voice and gesture at the graphics interface. SIGGRAPH Comput Graph 14(3):262–270
Bono M, Sumi Y, Nishida T (2007) Towards achieving complex medical engineering to understand conversational dynamics. In: Proceedings of the IEEE/ICME international conference on complex medical engineering (CME 2007), pp 474–478
Bosma W, André E (2004) Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the 9th international conference on intelligent user interfaces, pp 85–92
Bradley M, Lang P (2000) Cognitive neurosceince of emotion. Oxford University Press, New York
Breazeal C, Buchsbaum D, Gray J, Gatenby D, Blumberg B (2005) Learning from and about others: towards using imitation to bootstrap the social understanding of others by robots. Artif Life 11(1–2):31–62
Breazeal C, Scassellati B (2002) Robots that imitate humans. Trends Cog Sci 6(11):481–487
Brooks R (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom 2(1):14–23
Brown P, Levinson SC (1978) Politeness: some universals in language usage. Cambridge University Press, Cambridge
Buhler J, Tompa M (2002) Finding motifs using random projections. J Comput Biol 9(2):225–242
Bull PE (1987) Posture and gesture. Pergamon Press, Oxford
Caire P (2009) Designing convivial digital cities: a social intelligence design approach. AI SOC 24(1):97–114
Calverley D (2008) Imagining a non-biological machine as a legal person. AI SOC 22(4):523–537
Canelas A, Neves R, Horta N (2013) Multi-dimensional pattern discovery in financial time series using SAX-GA with extended robustness. In: Proceeding of the 15th annual conference companion on genetic and evolutionary computation conference companion, pp 179–180
Cardon A (2001) A distributed multi-agent system for the self-evaluation of dialogs. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 43–50
Carew P, Stapleton L, Byrne G (2008) Implications of an ethic of privacy for human-centred systems engineering. AI SOC 22(3):385–403
Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V, Kraaij W, Kronenthal M, Lathoud G, Lincoln M, Lisowska A, McCowan I, Post W, Reidsma D, Wellner P (2006) The AMI meeting corpus: a pre-announcement. Proceedings of the second international conference on machine learning for multimodal interaction. Springer, Berlin, pp 28–39
Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H, Yan H (1999) Embodiment in conversational interfaces: Rea. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 520–527
Cassell J, Sullivan J, Prevost S, Churchill E (eds) (2000) Embodied conversational agents. The MIT Press, Cambridge
Catalano J, Armstrong T, Oates T (2006) Discovering patterns in real-valued time series. In: Proceedings of the knowledge discovery in databases (PKDD 2006), pp 462–469

Cavallin H, Fruchter R, Nishida T (2010) The multiple faces of social intelligence design. AI SOC 25(2):141–143
Cavallin H, Martin W, Heylighen A (2007) How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability. AI SOC 22(2):227–235
Charman T, Baron-Cohen S, Swettenham J, Baird G, Cox A, Drew A (2000) Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. Cogn Dev 15(4):481–498
Chen L, Rose R, Qiao Y, Kimbara I, Parrill F, Welji H, Han T, Tu J, Huang Z, Harper M, Quek F, Xiong Y, McNeill D, Tuttle R, Huang T (2006) VACE multimodal meeting corpus. In: Renals S, Bengio S (eds) Machine learning for multimodal interaction. Springer, Berlin, pp 40–51
Chiu B, Keogh E, Lonardi S (2003) Probabilistic discovery of time series motifs. Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining (KDD '03). ACM, New York, pp 493–498
Clark H (1996) Using language. Cambridge University Press, Cambridge
Clawson VK, Bostrom RP (1993) The facilitation role in group support systems environments. In: Proceedings of the 1993 conference on computer personnel research, pp 323–335
Cohen PR (1995) Empirical methods for artificial intelligence. MIT Press, Cambridge
Cooley M (2007) From judgment to calculation. AI SOC 21(4):395–409
Cornillon J, Rosenberg D (2005) Dialogue organisation in argumentative debates. AI SOC 19(1):48–64
Cornillon J, Rosenberg D (2007) Experiment in social intelligence design. AI SOC 22(2):197–210
Cosley D, Frankowski D, Terveen L, Riedl J (2006) Using intelligent task routing and contribution review to help communities build artifacts of lasting value. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1037–1046
Costa PT Jr, McCrae RR (1992) Four ways five factors are basic. Pers Individ Differ 13(6):653–665
Cruz-Neira C, Sandin DJ, DeFanti TA (1993) Surround-screen projection-based virtual reality: the design and implementation of the cave. In: Proceedings of the 20th annual conference on computer graphics and interactive techniques, pp 135–142
Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceesdings of the IEEE computer society conference on computer vision and pattern recognition, 2005 (CVPR 2005), vol. 1, pp 886–893
Damasio AR (1994) Descartes' error: emotion, reason, and the human brain. Grosset, New York
Davenport T, Prusak L (2000) Working knowledge. Harvard Business School Press, Boston
Davies M, Stone T (1995) Mental simulation: evaluations and applications–reading in mind and language. Blackwell Publishers, Oxford
Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
de Jong G (1977) Frump..frump..frump. SIGART Bull 61:54–55
Decker M (2008) Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. AI SOC 22(3):315–330
Dennett DC (1989) The international stance. The MIT press, Cambridge
Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E (2008) Querying and mining of time series data: experimental comparison of representations and distance measures. Proc VLDB Endow 1(2):1542–1552
Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Classifying facial actions. IEEE Trans Pattern Anal Mach Intell 21(10):974–989
Dorrance BR, Zentall TR (2002) Imitation of conditional discriminations in pigeons (columba livia). J Comp Psychol 116(3):277–285
Duncan S Jr (1972) Some signals and rules for taking speaking turns in conversations. J Pers Soc Psychol 23(2):283–292
Duncan S Jr (1974) On the structure of speaker-auditor interaction during speaking turns. Lang Soc 3:161–180
Duncan S Jr, Niederehe G (1974) On signaling that its your turn to speak. J Exp Soc Psychol 10:234–247

Edwards D (1997) Discourse and cognition. Sage, London
Edwards G, Taylor C, Cootes T (1998) Interpreting face images using active appearance models. In: Proceedings of 3d IEEE international conference on automatic face and gesture recognition, 1998, pp 300–305
Eisenberg N, Eggum N, Di Giunta L (2010) Empathy-related responding: associations with prosocial behavior, aggression, and intergroup relations. Soc Issues Policy Rev 4(1):143–180
Ekman P (1992) An argument for basic emotions. Cogn Emotion 6(3/4):169–200
Ekman P, Friesen W (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto
Elliott CD (1992) The affective reasoner: a process model of emotions in a multi-agent system. Unpublished doctoral dissertation, Evanston
Erickson T (2009) social systems: designing digital systems that support social intelligence. AI SOC 23(2):147–166
Erman LD, Hayes-Roth F, Lesser VR, Reddy DR (1980) The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty. ACM Comput Surv 12(2):213–253
Essa I, Pentland A (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7):757–763
Fazio RH, Sanbonmatsu DM, Powell FR (1986) On the automatic activation of attitudes. J Pers Soc Psychol 50(2):229–238
Fodor JA (1981) The modularity of mind. MIT, MA
Freeman WT, Freeman WT, Roth M, Roth M (1994) Orientation histograms for hand gesture recognition. In: International workshop on automatic face and gesture recognition, pp 296–301
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Fruchter R (2001) Bricks & bits & interaction. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 35–42
Fruchter R (2005) Degrees of engagement in interactive workspaces. AI SOC 19(1):8–21
Fruchter R, Cavallin HE (2006) Developing methods to understand discourse and workspace in distributed computer-mediated interaction. AI SOC 20(2):169–188
Fruchter R, Nishida T, Rosenberg D (2005) Understanding mediated communication: the social intelligence design (sid) approach. AI SOC 19(1):1–7
Fruchter R, Nishida T, Rosenberg D (2007) Mediated communication in action: a social intelligence design approach. AI SOC 22(2):93–100
Fruchter R, Swaminathan S, Boraiah M, Upadhyay C (2007) Reflection in interaction. AI SOC 22(2):211–226
Fujihara N (2001) How to evaluate social intelligence design. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 75–82
Fukuhara T, Murayama T, Nishida T (2007) Analyzing concerns of people from weblog articles. AI SOC 22(2):253–263
Fukuhara T, Nishida T, Uemura S (2001) Public opinion channel: a system for augmenting social intelligence of a community. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 51–58
Furutani K, Kobayashi T, Ura M (2009) Effects of internet use on self-efficacy: perceived network-changing possibility as a mediator. AI SOC 23(2):251–263
Gal E, Bauminger N, Goren-Bar D, Pianesi F, Stock O, Zancanaro M (2009) Enhancing social communication of children with high-functioning autism through a co-located interface. AI SOC 24(1):75–84
Gallese V, Eagle MN, Migone P (2007) Intentional attunement: mirror neurons and the neural underpinnings of interpersonal relations. J Am Psychoanal Assoc 55(1):131–175
Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119(2):593–609

Gallese V, Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci 2(12):493–501
Georgeff MP, Ingrand FF (1990) Real-time reasoning: the monitoring and control of spacecraft systems. Proceedings of the sixth conference on artificial intelligence applications. IEEE Press, Piscataway, pp 198–204
Gill S, Borchers J (2003) Knowledge in co-action: social intelligence in collaborative design activity. AI SOC 17(3–4):322–339
Goffman E (1955) On face work: an analysis of ritual elements in social interaction. Psychiatry 18(3):213–231
Goffman E (1963) Behavior in public places. The Free Press, New York
Goffman E (1967) Interaction ritual: essays face-to-face behavior. Aldine, Chicago
Goffman E (1981) Forms of talk. University of Pennsylvania Press, Pennsylvania
Goldman AI (2006) Simulating minds: the philosophy, psychology, and neuroscience of mindreading. Oxford University Press, Oxford
Gombay E (2008) Change detection in autoregressive time series. J Multivar Anal 99(3):451–464
González VM, Mark G (2004) "Constant, constant, multi-tasking craziness": managing multiple working spheres. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 113–120
Goodwin C (1981) Conversational organization: interaction between speakers and hearers. Acadamic Press, New York
Gopnik A (2003) The theory as an alternative to the innateness hypothesis. In: Antony LM (ed) Chomsky and his critics. Blackwell, Oxford
Gordon RM (1986) Folk psychology as simulation. Mind Lang 1(2):158–171
Green BF Jr, Wolf AK, Chomsky C, Laughery K (1961) Baseball: an automatic question-answerer. In: Presented papers at the western joint IRE-AIEE-ACM computer conference. ACM, New York, pp 219–224, May 9–11
Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the implicit association test: I. an improved scoring algorithm. J Pers Soc Psychol 85(2):197–216
Hachimura K, Nakamura M (2001) Method of generating coded description of human body motion from motion-captured data. In: Proceedings of the 10th IEEE international workshop on robot and human interactive communication, pp 122–127
Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248
Hayes-Roth B (1985) A blackboard architecture for control. Artif Intell 26(3):251–321
Hayes-Roth B (1998) Jennifer james, celebrity auto spokesperson. In: ACM SIGGRAPH '98 conference abstracts and applications. ACM, New York, p 136
Hayes-Roth B, van Gent R (1997) Story-making with improvisational puppets. In: Agents, pp 1–7
Heyes C (2010) Where do mirror neurons come from? Neurosci Biobehav Rev 34(4):575–583
Heylen D, Kopp S, Marsella SC, Pelachaud C, Vilhjálmsson H (2008) The next step towards a function markup language. Proceedings of the 8th international conference on intelligent virtual agents. Springer, Berlin, pp 270–280
Heylighen A, Heylighen F, Bollen J, Casaer M (2007) Distributed (design) knowledge exchange. AI SOC 22(2):145–154
Hirano S, Tsumoto S (2002) Mining similar temporal patterns in long time-series data and its application to medicine. Proceedings of the 2002 IEEE international conference on data mining (ICDM'02). IEEE Computer Society, Washington DC, pp 219–226
Hjelmås E, Low BK (2001) Face detection: a survey. Comput Vis Image Underst 83(3):236–274
Hofstede G (2001) Culure's consequences, 2nd edn. Sage Publications, Thousand Oaks
Hofstede G, Hofstede GJ (2005) Cultures and orgnizations: software of the mind. McGraw-Hill, New York
Hofstede GJ, Pedersen PB, Hofstede G (2002) Exploring culture: excercises, stories and synthetic cultures. Intercultural Press, Boston
Hofte GH, Mulder I, Verwijs C (2006) Close encounters of the virtual kind: a study on place-based presence. AI SOC 20(2):151–168

Huang H-H, Nishida T, Cerekovic A, Pandzic IS, Nakano Y (2008) Proceedings of the 7th international joint conference on autonomous agents and multiagent systems. International Foundation for Autonomous Agents and Multiagent Systems, Richland, pp 128–135

Huang Z, Eliens A, Visser C (2004) STEP: a scripting lan-guage for embodied agents. In: Prendinger H, Ishizuka M (eds) Life-like characters: tools, affective functions, and applications. Springer, pp 87–109

Iacoboni M (2008) Mirroring people: the new science of how we connect with others. Farrar, Straus and Giroux, New York

Iacoboni M (2009) Imitation, empathy, and mirror neurons. Ann Rev Psychol 60:653–670

Ide T, Inoue K (2005) Knowledge discovery from heterogeneous dynamic systems using change-point correlations. In: Proceedings of the SIAM international conference on data mining, pp 571–575

Imai M, Kanda T, Ono T, Ishiguro H, Mase K (2002) Robot mediated round table: analysis of the effect of robot's gaze. In: Proceedings of 11th IEEE international workshop on robot and human interactive communication, pp 411–416

Ishizuka M, Prendinger H (2006) Describing and generating multimodal contents featuring affective lifelike agents with mpml. New Gen Comput 24(2):97–128

Iwaki M, Arakawa S, Kiryu T (2008) Influence on biosignal and working efficiency of sound environment in typewriting. Technical report. IEICE, vol. 108, No. 52, MBE2008-5, pp 19–24 (in Japanese)

Jensen KL, Styczynxki MP, Rigoutsos I, Stephanopoulos GN (2006) A generic motif discovery algorithm for sequenctial data. BioInformatics 22(1):21–28

Jones SS (2009) The development of imitation in infancy. Philos Trans R Soc B Biol Sci 364(1528):2325–2335

Kadambe S, Boudreaux-Bartels G (1992) Application of the wavelet transform for pitch detection of speech signals. IEEE Trans Inf Theor 38(2):917–924

Kanade T (1973) Picture processing system by computer complex and recognition of human faces. Unpublished doctoral dissertation, Kyoto University, Kyoto

Kanakogi Y, Itakura S (2011) Developmental correspondence between action prediction and motor ability in early infancy. Nat Commun 2:341

Kanda T, Ishiguro H, Ono T, Imai M, Nakatsu R (2002) Development and evaluation of an interactive humanoid robot "robovie". In: Proceedings of IEEE international conference on robotics and automation (ICRA '02), vol. 2, pp 1848–1855

Kaner S (2007) Facilitator's guide to participatory decision-making. Wiley, New York

Kant I (1788) Critique of practical reason. Cambridge University Press, Cambridge (Translated by Mary Gregor, 1997)

Katai O, Minamizono K, Shiose T, Kawakami H (2007) System design of ba-like stages for improvisational acts via leibnizian space-time and peirces existential graph concepts. AI SOC 22(2):101–112

Katai O, Nishida T, Fruchter R (2011) Situated and embodied interactions for symbiotic and inclusive societies. AI SOC 26(3):193–196

Kendon A (1967) Some functions of gaze-direction in social interaction. Acta Psychol 26:22–63

Kendon A (1972) Some relationships between body motion and speech. Stud Dyadic Commun 7:177

Kendon A (1990) Conducting interaction: patterns of behavior in focused encounters. Cambridge University Press, Cambridge

Kendon A (2004) Gesture. Cambridge University Press, New York

Kenward B (2012) Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. J Exp Child Psychol 112(2):195–207

Keogh E, Lin J, Fu A (2005) HOT SAX: efficiently finding the most unusual time series subsequence. Proceedings of the fifth IEEE international conference on data mining. IEEE Computer Society, Washington DC, pp 226–233

Khalifa M, Kwok R-W, Davison R (2002) The effects of process and content facilitation restrictiveness on gss-mediated collaborative learning. Group Decis Negot 11(5):345–361
Kidd CD, Breazeal C (2005) Human-robot interaction experiments: lessons learned. In: Robot compaions: hard problems and open challenges in robot human interaction symbosioum of social intelligence and interaciton in animals robots and agents, pp 141–142
Kipp M (2001) ANVIL: a generic annotation tool for multimodal dialogue, pp 1367–1370
Kipp M (2004) Gesture generation by imitation: from human behavior to computer character animation. Dissertation.com, Boca Raton
Kipp M, Neff M, Kipp K, Albrecht I (2007) Towards natural gesture synthesis: evaluating gesture units in a data-driven approach to gesture synthesis. In: Pelachaud C, Martin J-C, André E, Chollet G, Karpouzis K, Pel D (eds) Intelligent virtual agents. Springer, Berlin, pp 15–28
Kitamura M, Shimohata S, Sukehiro T, Ikeno A, Sakamoto M, Orihara I, Murata T (2008) Design and development of dialogue system for laddering search service, vol. 108. IEICE technical report. Natural language understanding and models of communication
Kitsuregawa M, Nishida T (2010) Special issue on information explosion. New Gener Comput 28(3):207–215
Knapp ML, Hall JA (2010) Nonverbal communication in human interaction. Wadsworth/Cengage, Boston
Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on artificial intelligence, vol 2. Morgan Kaufmann Publishers Inc, San Francisco, pp 1137–1143
Kopp S, Krenn B, Marsella S, Marshall AN, Pelachaud C, Pirker H, Thórisson KR, Vilhjálmsson H (2006) Towards a common framework for multimodal generation: the behavior markup language. Proceedings of the 6th international conference on intelligent virtual agents. Springer, Berlin, pp 205–217
Kopp S, Tepper P, Striegnitz K, Ferriman K, Cassell J (2007) Trading spaces: how humans and humanoids use speech and gesture to give directions. In: Nishida T (ed) Conversational informatics: an engineering approach. Wiley, Chichester, pp 133–160
Kovar L, Gleicher M, Pighin F (2002) Motion graphs. ACM Trans Graph 21(3):473–482
Kubota H, Nishida T (2002) EgoChat Agent: a talking virtualized member for supporting community knowledge creation. In: Dautenhahn K, Bond A, Canamero D, Edmonds B (eds) Socially intelligent agents: creating relationships with computers and robots. IOS Press, Amsterdam
Kubota H, Nishida T, Koda T (2000) Exchanging tacit community knowledge by talking-virtualized-egos. Proceedings of the fourth international conference on autonomous agents. ACM, New York, pp 285–292
Kubota H, Nomura S, Sumi Y, Nishida T (2007) Sustainable memory system using global and conical spaces. j-jucs 13(2):135–148
Kurata Y (2010) Interactive assistance for tour planning. In: Hölscher C, Shipley TF, Olivetti Belardinelli M, Bateman JA, Newcombe NS (eds) Spatial cognition VII. Springer, Berlin, pp 289–302
Laban R, Ullmann L (1960) Mastery of movement. Princeton Book Company Publishers, NJ
Lakoff G (1987) Women, fire, and dangerous things: what categoris reveal about the mind. The University of Cicago Press, Cicago
Lala D, Nishida T (2013) Modeling communicative virtual agents based on joint activity theory. In: Proceedings of IEEE international conference on cognitive informatics cognitive computing (ICCI*CC 2013), pp 8–16
Lang PJ (1995) The emotion probe: studies of motivation and attention. Am Psychol 50(5):285–372
Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2008 (CVPR 2008), pp 1–8
Leslie AM (1987) Pretense and representation: the origins of “theory of mind”. Psychol Rev 94(4):412–426

Li Y, Lin J (2010) Approximate variable-length time series motif discovery using grammar inference. In: Proceedings of the 10th international workshop on multimedia data mining. ACM, New York, pp 10:1–10:9

Liao W, Zhang W, Zhu Z, Ji O (2005) A decision theoretic model for stress recognition and user assisstance. In: AAAI 2005, pp 529–534

Lim M, Dias J, Aylett R, Paiva A (2012) Creating adaptive affective autonomous npcs. Auton Agents Multi-Agent Syst 24(2):287–311

Limayem M (2006) Human versus automated facilitation in the GSS context. ACM SIGMIS Database 37(2–3):156–166

Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, pp 2–11

Lin J, Keogh E, Lonardi S, Patel P (2002) Finding motifs in time series. In: Proceedings of the the 2nd workshop on temporal data mining, pp 53–68

Lin J, Keogh E, Wei L, Lonardi S (2007) Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Disc 15(2):107–144

Lin T, Omata M, Hu W, Imamiya A (2005) Do physiological data relate to traditional usability indexes? In: Proceedings of the 17th Australia conference on computer-human interaction: citizens online: considerations for today and the future, Narrabundah, Australia, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia, pp 1–10

Loke L, Larssen AT, Robertson T (2005) Labanotation for design of movement-based interaction. Proceedings of the second Australasian conference on interactive entertainment. Creativity and Cognition Studios Press, Sydney, pp 113–120

Lowe D (1999) Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision, vol. 2, pp 1150–1157

Luehrs R, Malsch T, Voss K (2001) Internet, discourses, and democracy. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 67–74

Maes P, Blumberg B, Darrell T, Pentland A, Wexelblat A (1995) Modeling interactive agents in ALIVE. Proceedings of the 14th international joint conference on artificial intelligence, vol 2. Morgan Kaufmann Publishers Inc, San Francisco, pp 2073–2074

Maimone A, Fuchs H (2011) Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras. In: Proceedings of the 10th IEEE international symposium on mixed and augmented reality (ismar), 2011, pp 137–146

Mandryk RL, Inkpen KM (2004) Physiological indicators for the evaluation of co-located collaborative play. In: Proceedings of the 2004 ACM conference on computer supported cooperative work (CSCW '04), pp 102–111

Mao W, Gratch J (2009) Modeling social inference in virtual agents. AI SOC 24(1):5–11

Margolis E, Samuels R, Stich SP (2012) The oxford handbook of philosophy of cognitive science. Oxford University Press, Oxford

Mark G, Gudith D, Klocke U (2008) The cost of interrupted work: more speed and stress. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 107–110

Martin W, Heylighen A, Cavallin H (2005) The right story at the right time. AI SOC 19(1):34–47

Matsumura K, Nakano YI, Nishida T (2005) The explanatory experiment for evaluation of SPOC system from contents creators' perspective. Proceedings of the second international conference on intelligent media technology for communicative intelligence. Springer, Berlin, pp 79–90

Matsumura N, Miura A, Shibanai Y, Ohsawa Y, Nishida T (2005) The dynamism of 2 channel. AI SOC 19(1):84–92

Maurer H (2007) Some ideas on ICT as it influences the future. Talk given at NEC Technology Forum, Tokyo

McCann J, Pollard NS (2007) Responsive characters from motion fragments. ACM Trans Graphics 26(3):417:426

McCrae RR, Costa PT (1987) Validation of the five-factor model of personality across instruments and observers. J Pers Soc Psychol 52(1):81–90
McCrae RR, Coasta PT (1997) Personality trait structure as a human universal. Am Psychol 52(5):509–516
McGuigan N, Makinson J, Whiten A (2011) From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. Br J Psychol 102(1):1–18
McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago
McNeill D (2005) Gesture and thought. The University of Chicago Press, Chicago
Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14(4):261–292
Meltzoff AN (2005) Imitation and other minds: the like me hypothesis. Perspect Imitation Neurosc Soc Sci 2:55–77
Meltzoff AN, Moore MK (1997) Explaining facial imitation: a theoretical model. Early Dev Parenting 6:179–192
Merckel L, Nishida T (2009a) Enabling situated knowledge management for complex instruments by real-time reconstruction of surface coordinate system on a mobile device. AI SOC 24(1):85–95
Merckel L, Nishida T (2009b) Low-overhead 3D items drawing engine for communicating situated knowledge. In: Liu J, Wu J, Yao Y, Nishida T (eds) Active media technology. Springer, Berlin, pp 31–41
Merckel L, Nishida T (2010) Multi-interfaces approach to situated knowledge management for complex instruments: first step toward industrial deployment. AI SOC 25(2):211–223
Microsoft Corporation (1999) Microsoft agent software development kit
Minnen D, Starner T, Essa I, Isbell C (2007) Improving activity discovery with automatic neighborhood estimation. In: Proceedings of international joint conference on artificial intelligence, pp 6–12
Mita T, Kaneko T, Hori O (2005) Joint haar-like features for face detection. In: Proceedings of the 10th IEEE international conference on computer vision (ICCV 2005), vol. 2, pp 1619–1626
Mitra S, Acharya T (2007) Gesture recognition: a survey. IEEE Trans Sys Man Cybern Part C Appl Rev 37(3):311–324
Miura A, Fujihara N, Yamashita K (2006) Retrieving information on the World Wide Web: effects of domain specific knowledge. AI SOC 20(2):221–231
Miura A, Matsumura N (2009) Social intelligence design: a junction between engineering and social sciences. AI SOC 23(2):139–145
Miura A, Shinohara K (2005) Social intelligence design in online chat communication: a psychological study on the effects of "congestion". AI SOC 19(1):93–109
Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81(3):231–268
Mohammad Y, Nishida T (2008a) Constrained motif discovery. In: International workshop on data mining and statistical science (DMSS2008), pp 16–19
Mohammad Y, Nishida T (2008b) Toward agents that can learn nonverbal interactive behavior. In: IAPR workshop on cognitive information processing, pp 164–169
Mohammad Y, Nishida T (2009a) Constrained motif discovery in time series. New Gener Comput 27(4):319–346
Mohammad Y, Nishida T (2009b) Interactive perception for amplification of intended behavior in complex noisy environments. AI SOC 23(2):167–186
Mohammad Y, Nishida T (2009c) Learning interaction structure using a hierarchy of dynamical systems. In: IEA/AIE, pp 253–258
Mohammad Y, Nishida T (2009d) Measuring naturalness during close encounters using physiological signal processing. In: IEA/AIE, pp 281–290

Mohammad Y, Nishida T (2009e) Robust singular spectrum transform. In: The twenty second international conference on industrial, engineering and other applications of applied intelligent systems (IEA-AIE 2009), pp 123–132
Mohammad Y, Nishida T (2009f) Toward combining autonomy and interactivity for social robots. AI SOC 24(1):35–49
Mohammad Y, Nishida T (2010a) Controlling gaze with an embodied interactive control architecture. Appl Intell 32(2):148–163
Mohammad Y, Nishida T (2010b) Mining causal relationships in multidimensional time series. In: Szczerbicki E, Nguyen NT (eds) Smart information and knowledge management: advances, challenges, and critical issues. Springer, pp. 309–338
Mohammad Y, Nishida T (2010c) Using physiological signals to detect natural interactive behavior. Appl Intell 33:79–92. doi:10.1007/s10489-010-0241-4
Mohammad Y, Nishida T (2011a) Discovering causal change relationships between processes in complex systems. In: Proceedings of the IEEE/SICE international symposium on system integration, pp 12–17
Mohammad Y, Nishida T (2011b) On comparing SSA-based change point discovery algorithms. In: Proceedings of the IEEE/SICE international symposium on system integration, pp 938–945
Mohammad Y, Nishida T (2012a) Fluid imitation: discovering what to imitate. Int J Soc Robot 4(4):369–382
Mohammad Y, Nishida T (2012b) Unsupervised discovery of basic human actions from activity recording datasets. In: Proceedings of the IEEE/SICE international symposium on system integration, pp 402–409
Mohammad Y, Nishida T (2013a) Learning sensorimotor concepts without reinforcement. In: AAAI summer symposium on life long machine learning
Mohammad Y, Nishida T (2013b) Tackling the correspondence problem: closed-form solution for gesture imitation by a humanoid's upper body. In: Yoshida T, Kou G, Skowron A, Cao J, Hacid H, Zhong N (eds) Active media technology, vol. 8210. Springer International Publishing, pp 84–95
Mohammad Y, Nishida T (2014a) Exact discovery of length-range motifs. In: The 6th Asian conference on intelligent information and database systems
Mohammad Y, Nishida T (2014b) Learning where to look: autonomous development of gaze behavior for natural human-robot interaction. Interact Stud 14(3):419–450
Mohammad Y, Nishida T, Okada S (2009) Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction. Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems (IROS'09). IEEE Press, Piscataway, pp 2537–2544
Mohammad Y, Nishida T, Okada S (2010) Autonomous development of gaze control for natural human-robot interaction. Proceedings of the 2010 workshop on Eye Gaze in intelligent human machine interaction. ACM, New York, pp 63–70
Mohammad Y, Xu Y, Matsumura K, Nishida T (2008) The H^3R explanation corpus: human-human and base human-robot interaction dataset. In: Proceedings of the 4th international conference on intelligent sensors, sensor networks and information processing (ISSNIP2008), pp 201–206
Molenberghs P, Cunnington R, Mattingley JB (2009) Is the mirror neuron system involved in imitation? a short review and meta-analysis. Neurosci Biobehav Rev 33(7):975–980
Mori S, Ohmoto Y, Nishida T (2011) Constructing immersive virtual space for hai with photos. In: Proceedings of the IEEE international conference on granular computing (grc), pp 479–484
Morio H, Buchholz C (2009) How anonymous are you online? examining online social behaviors from a cross-cultural perspective. AI SOC 23(2):297–307
Moriyama J, Kato Y, Aoki Y, Kito A, Behnoodi M, Miyagawa Y, Matsuura M (2009) Self-efficacy and learning experience of information education: in case of junior high school. AI SOC 23(2):309–325
Mower E, Feil-Seifer DJ, Mataric MJ, Narayanan S (2007) Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements. In: Proceedings of the 16th international conference on robot and human interactive communication, pp 1125–1130

Mueen A (2013) Enumeration of time series motifs of all lengths. In: Proceedings of the IEEE international conference on data mining, pp 547–556
Mueen A, Keogh E, Bigdely-Shamlo N (2009) Finding time series motifs in disk-resident data. In: Proceedings of the 9h IEEE international conference on data mining (ICDM'09), pp 367–376
Mueen A, Keogh E, Zhu Q, Cash S, Westover B (2009) Exact discovery of time series motifs. In: Proceedings of 2009 SIAM international conference on data mining, pp 1–12
Nagai Y (2005) Joint attention development in infant-like robot based on head movement imitation. In: Proceedings of the 3rd international symposium on imitation in animals and artifacts, pp 87–96
Nagai Y, Rohlfing KJ (2007) Can motionese tell infants and robots what to imitate? In: Proceedings of the 4th international symposium on imitation in animals and artifacts, pp 299–306
Nagaoka C, Yoshikawa S, Komori M (2006) Embodied synchrony of nonverbal behaviour in counseling: a case study of role playing school counseling. In: Proceedings of the 28th annual conference of the cognitive science society (CogSci 2006), pp 1862–186
Nagenborg M, Capurro R, Weber J, Pingel C (2008) Ethical regulations on robotics in europe. AI SOC 22(3):349–366
Nagy E, Molnar P (2004) Homo imitans or homo provocans? human imprinting model of neonatal imitation. Infant Behav Dev 27(1):54–63
Nakano YI, Murayama T, Okamoto M, Kawahara D, Li Q, Kurohashi S, Nishida T (2006) Cards-to-presentation on the web: generating multimedia contents featuring agent animations. J Netw Comput Appl 29(2):83–104
Nakano YI, Reinstein G, Stocky T, Cassell J (2003) Towards a model of face-to-face grounding. Proceedings of the 41st annual meeting on association for computational linguistics, vol 1. Association for Computational Linguistics, Stroudsburg, pp 553–561
Nakata K (2001) Enabling public discourse. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 59–66
Nass C, Steuer J, Tauber ER (1994) Computers are social actors. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 72–78
Nehaniv C, Dautenhahn K (1998) Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In: Demiris J, Birk A (eds) Proceedings European workshop on learning robots 1998 (EWLR-7), pp 64–72
Nevill-Manning CG, Witten IH (1997) Identifying hierarchical structure in sequences: a linear-time algorithm. J Artif Int Res 7(1):67–82
Niederman F, Beise CM, Beranek PM (1993) Facilitation issues in distributed group support systems. In: Proceedings of the 1993 conference on computer personnel research, pp 299–312
Nijholt A (2001) From virtual environment to virtual community. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 19–26
Nijholt A, Akker R, Heylen D (2006) Meetings and meeting modeling in smart environments. AI SOC 20(2):202–220
Nijholt A, Nishida T (2006) Social intelligence design for mediated communication. AI SOC 20(2):119–124
Nijholt A, Stock O, Nishida T (2009) Social intelligence design in ambient intelligence. AI SOC 24(1):1–3
Nishida T (2001) Social intelligence design: an overview. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 3–10
Nishida T (2002) Social intelligence design for the web. Computer 35(11):37–41
Nishida T (2007a) Conversational informatics and human-centered web intelligence. IEEE Intell Inf Bull 8(1):19–28
Nishida T (ed) (2007b) Conversational informatics: an engineering approach. Wiley, London
Nishida T (2007c) Social intelligence design and human computing. In: Huang TS, Nijholt A, Pantic M, Pentland A (eds) Artifical intelligence for human computing. Springer, pp 190–214

Nishida T (2009) Towards robots with good will. In: Nagenborg M, Capurro R (eds) Ethics and robotics. IOS Press, Amsterdam
Nishida T (2010a) Modeling machine emotions for realizing intelligence: an introduction. In: Nishida T, Jain L, Faucher C (eds) Modeling machine emotions for realizing intelligence. Springer, Berlin, pp 1–15
Nishida T (2010b) Social intelligence design for knowledge circulation. In: Kikuchi S, Sachdeva S, Bhalla S (eds) Databases in networked information systems. Springer, Berlin, pp 122–142
Nishida T (2012) Artificial intelligence research in the second half century. J Inf Process Manage 55(7):461–471 (in Japanese)
Nishida T (2013) Toward mutual dependency between empathy and technology. AI SOC 28(3):277–287
Nishida T, Fujihara N, Azechi S, Sumi K, Yano H, Hirata T (1999) Public opinion channel for communities in the information age. New Gener Comput 17(4):417–427
Nishida T, Hirata T, Maeda H (1998) CoMeMo-community: a system for supporting community knowledge evolution. In: Ishida T (ed) Community computing and support systems. Springer, Berlin, pp 183–200
Nishida T, Nishida R (2007) Socializing artifacts as a half mirror of the mind. AI SOC 21(4):549–566
Nishida T, Terada K, Tajima T, Hatakeyama M, Ogasawara Y, Sumi Y, Xu Y, Mohammad YFO, Tarasenko K, Ohya T, Hiramatsu T (2006) Toward robots as embodied knowledge media. IEICE Trans 89-D(6):1768–1780
Nomura T, Kanda T, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. AI SOC 20(2):138–150
Nonaka I, Takeuchi H (1995) The knowledge-creating company: how Japanese companies create the dynamics of innovation. Oxford University Press, Oxford
Notsu A, Ichihashi H, Honda K, Katai O (2009) Visualization of balancing systems based on nave psychological approaches. AI SOC 23(2):281–296
Numaguchi N, Nakazawa A, Shiratori T, Hodgins JK (2011) A puppet interface for retrieval of motion capture data. Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation. ACM, New York, pp 157–166
Nunthanid P, Niennattrakul V, Ratanamahatana C (2011) Discovery of variable length time series motif. In: Proceedings of the 8th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON), pp 472–475
Oates T (2002) PERUSE: an unsupervised algorithm for finding recurring patterns in time series. In: Proceedings of the IEEE international conference on data mining (ICDM 2002), pp 330–337
Ohguro T, Kuwabara K, Owada T, Shirai Y (2001) FaintPop: in touch with the social relationships. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 11–18
Ohmoto Y, Kataoka M, Miyake T, Nishida T (2011) A method to dynamically estimate emphasizing points and degree by using verbal and nonverbal information and physiological indices. In: Proceedings of the 2011 IEEE international conference on granular computing, pp 508–514
Ohmoto Y, Kataoka M, Nishida T (2013) Extended methods to dynamically estimate emphasizing points for group decision-making and their evaluation. Procedia-Soc Behav Sci 97:147–155
Ohmoto Y, Lala D, Saiga H, Ohashi H, Mori S, Sakamoto K, Kinoshita K, Nishida T (2013) Design of immersive environment for social interaction based on socio-spatial information and the applications. J Inf Sci Eng 29(4):663–679
Ohmoto Y, Miyake T, Nishida T (2010) Judgement as to whether or not people are involved, enjoying and excited, based on the visual and physiological information. In: Nishida T, Jain L, Faucher C (eds) Modeling Machine Emotions for Realizing Intelligence. Springer, Berlin Heidelberg, pp 35–52
Ohmoto Y, Miyake T, Nishida T (2012) Dynamic estimation of emphasizing points for user satisfaction evaluations. In: Proceedings of the 34th annual conference of the cognitive science society, pp 2115–2120

Ohmoto Y, Toda Y, Ueda K, Nishida T (2011) Analyses of the facilitating behavior by using participant's agreement and nonverbal behavior. J Inf Process Soc Japan 52(12):3659–3670
Ong SC, Ranganath S (2005) Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans Pattern Anal Mach Intell 27(6):873–891
Oostenbroek J, Slaughter V, Nielsen M, Suddendorf T (2013) Why the confusion around neonatal imitation? a review. J Reprod Infant Psychol 31(4):328–341
Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
Page ES (1954) Continuous inspection schemes. Biometrika 44:100–115
Pan X, Han C, Dauber K, Law K (2007) A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. AI SOC 22(2):113–132
Pandzic IS, Forchheimer R (eds) (2003) MPEG-4 facial animation: the standard, implementation and applications. Wiley, New York
Papillo JF, Shapiro D (1990) The cardiovascular system. In: Cacioppo JT, Tassinary LG (eds) Principles of psychophysiology: physical, social and inferential elements. Cambridge University Press, Cambridge
Pavesi G, Mauri G, Pesole G (2001) Methods for pattern discovery in unaligned biological sequences. Briefings Bioinf 2(4):417
Pentland AS (2008) Honest signals: how they shape our world. The MIT Press, Cambridge
Perrow C (1984) Normal accidents: living with high-risk technologies. Princeton University Press, NJ
Pevzner PA, Sze S-H et al (2000) Combinatorial approaches to finding subtle signals in DNA sequences. In: ISMB, pp 269–278
Picard RW (1997) Affective computing. MIT Press, Cambridge
Plutchik R (1980) Emotion: a psychoevolutionary synthesis. Harper and Row, New York
Poel M, Heylen D, Nijholt A, Meulemans M, Breemen A (2009) Gaze behaviour, believability, likability and the icat. AI SOC 24(1):61–73
Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–526
Prendinger H, Ishizuka M (eds) (2004) Life-like characters: tools, affective functions and applications. Springer, Heidelberg
Prendinger H, Ishizuka M (2005) The empathic companion: a character-based interface that addresses users' affective states. Appl Artif Intell 19(3–4):267–285
Pumareja D, Sikkel K (2006) Getting used with groupware: a first class experience. AI SOC 20(2):189–201
Qazi Z, Wang Z, Haq I (2006) Human likeness of humanoid robots exploring the uncanny valley. In: Proceedings of the international conference on emerging technologies (ICET '06), pp 650–656
Ramachandran VS, Oberman LM (2006) Broken mirrors: a theory of autism. Sci Am 295(5):62–69
Rao RPN, Shon AP, Meltzoff AN (2004) A bayesian model of imitation in infants and robots. Cambridge University Press, Cambridge, pp 217–247
Rawls J (1999) A theory of justice. Oxford University Press, Oxford
Ray E, Heyes C (2011) Imitation in infancy: the wealth of the stimulus. Dev Sci 14(1):92–105
Reagan-Cirincione P (1994) Improving the accuracy of group judgment: a process intervention combining group facilitation, social judgment analysis, and information technology. Organ Behav Hum Decis Process 58(2):246–270
Rehm M, Endrass B (2009) Rapid prototyping of social group dynamics in multiagent systems. AI SOC 24(1):13–23
Rehm M, Nakano Y, André E, Nishida T, Bee N, Endrass B, Wissner M, Lipi A, Huang H-H (2009) From observation to simulation: generating culture-specific behavior for interactive systems. AI SOC 24(3):267–280
Reidsma D, Akker R, Rienks R, Poppe R, Nijholt A, Heylen D, Zwiers J (2007) Virtual meeting rooms: from observation to simulation. AI SOC 22(2):133–144

Richmond VP, McCroskey JC, Mickson III ML (2004) Nonverbal behavior in interpersonal relations, 7th edn. Pearson Education Inc, New York
Rienks R, Nijholt A, Barthelmess P (2009) Pro-active meeting assistants: attention please!. AI SOC 23(2):213–231
Rizzolatti G, Sinigaglia C (2008) Mirrors in the brain: how our minds share actions, emotions, and experience. Oxford University Press, Oxford
Roest G, Szirbik N (2009) Escape and intervention in multi-agent systems. AI SOC 24(1):25–34
Rosenberg D, Foley S, Lievonen M, Kammas S, Crisp M (2005) Interaction spaces in computer-mediated communication. AI SOC 19(1):22–33
Rosenberg R (2008) The social impact of intelligent artefacts. AI SOC 22(3):367–383
Rosner B (1975) On the detection of many outliers. Technometrics 17(2):221–227
Rowe DW, Sibert J, Irwin D (1998) Heart rate variability: indicator of user state as an aid to human-computer interaction. Proceedings of the SIGCHI conference on human factors in computing systems. ACM Press/Addison-Wesley Publishing Co, New York, pp 480–487
Rowley HA, Baluja S, Kanade T (1998) Neural network-based face detection. IEEE Trans Pattern Anal Mach Intell 20(1):23–38
Ruttkay Z, Pelachaud C (2004) From brows to trust: evaluating embodied conversational agents. Springer, Berlin
Sabbagh MA (2004) Understanding orbitofrontal contributions to theory-of-mind reasoning: implications for autism. Brain Cogn 55:209–219
Sacks H, Schegloff EA, Jefferson GA (1974) A simplest systematics for the organization of turn-taking in conversation. Language 50:996–735
Sagot M-F (1998) Spelling approximate repeated or common motifs using a suffix tree. In: Lucchesi C, Moura A (eds) LATIN'98: theoretical informatics. Springer, Berlin, pp 374–390
Saito K, Kubota H, Sumi Y, Nishida T (2007) Analysis of conversation quanta for conversational knowledge circulation. J UCS 13(2):177–185
Schank R (1990) Tell me a story: a new look at real and artificial memory. John Brockman Associates Inc, New York
Schank RC (1982) Dynamic memory: a theory ofreminding and learning in computers and people. Cambridge University Press, Cambridge
Schank RC, Abelson R (1977) Scripts, plans, goals, and understanding. Lawrence Erlbaum Associates, Hillsdale
Schank RC, Riesbeck CK (1981) Inside computer understanding: five programs plus miniatures. Erlbaum, Hillsdale
Schegloff EA (1968) Sequencing in conversational openings. Am Anthropol 70:1075–1095
Schegloff EA (1999) Discourse, pragmatics, conversation, analysis. Discourse Stud 1(4):405–435
Schegloff EA, Jefferson G, Sacks H (1977) The preference for self-correction in the organization of repair in conversation. Language 53(2):361–382
Schegloff EA, Sacks H (1973) Opening up closings. Semiotica VII 1(4):289–327
Schuman SP (1996) What to look for in a group facilitator. Qual Prog 29(6):69–76
Sculley J, Byrne JA (1987) Odyssey: pepsi to apple: a journey of adventure, ideas, and the future. Harpercollins, New York
Searle J (1969) Speech acts. Cambridge University Press, Cambridge
Sellars W et al (1956) Empiricism and the philosophy of mind. Minn Stud philos Sci 1:253–329
Shen X, Li Q, Yu T, Geng W, Lau N (2005) Mocap data editing via movement notations. Proceedings of the ninth international conference on computer aided design and computer graphics. IEEE Computer Society, Washington DC, pp 463–470
Shi Y, Choi EHC, Ruiz N, Chen F, Taib R (2007) Galvanic skin respons (GSR) as an index of cognitive load. In: CHI 2007, pp 2651–2656
Shiratori T, Nakazawa A, Ikeuchi K (2006) Dancing-to-music character animation. Comput Graph Forum 25(3):449–458
Shoji H, Fujimoto K, Hori K (2009) PLASIU: a system that facilitates creative decision-making in job-hunting. AI SOC 23(2):265–279

Shoji H, Hori K (2005) S-conart: an interaction method that facilitates concept articulation in shopping online. AI SOC 19(1):65–83
Sidner CL, Kidd CD, Lee C, Lesh N (2004) Where to look: a study of human-robot engagement. Proceedings of the 9th international conference on intelligent user interfaces (IUI'04). ACM, New York, pp 78–84
Siegel LS (1981) Infant tests as predictors of cognitive and language development at two years. Child Dev 52(2):545–557
Simpson A, Riggs KJ (2011) Three-and 4-year-olds encode modeled actions in two ways leading to immediate imitation and delayed emulation. Dev psychol 47(3):834–840
Soussignan R, Courtial A, Canet P, Danon-Apter G, Nadel J (2011) Human newborns match tongue protrusion of disembodied human and robotic mouths. Dev Sci 14(2):385–394
Staden R (1989) Methods for discovering novel motifs in nucleic acid sequences. Comput Appl Biosci 5(4):293–298
Staudte M, Crocker MW (2010) When robot gaze helps human listeners: attentional versus intentional account. In: Proceedings of the 32nd annual meeting of the cognitive science society, pp 1637–1642
Stefik M, Foster G, Bobrow DG, Kahn K, Lanning S, Suchman L (1987) Beyond the chalkboard: computer support for collaboration and problem solving in meetings. Commun ACM 30(1):32–47
Stein A, Thiel U (1993) A conversational model of multimodal interaction in information systems. In: Proceedings of the 11th national conference on artificial intelligence. AAAI Press, pp 283–288
Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, Hoymann G, Rossano F, de Ruiter JP, Yoon K-E, Levinson SC (2009) Universals and cultural variation in turn-taking in conversation. Proc Nat Acad Sci 106(26):10587–10592
Stock O, Zancanaro M, Rocchi C, Tomasini D, Koren C, Eisikovits Z, Goren-Bar D (2009) The design of a collaborative interface for narration to support reconciliation in a conflict. AI SOC 24(1):51–59
Stone M, DeCarlo D, Oh I, Rodriguez C, Stere A, Lees A, Bregler C (2004) Speaking with hands, creating animated conversational characters from recordings of human performance. ACM Trans Graph 23(3):506
Su NM, Mark G (2008) Communication chains and multitasking. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 83–92
Sumi Y, Yano M, Nishida T (2010) Analysis environment of conversational structure with nonverbal multimodal data. In: Proceedings of the international conference on multimodal interfaces and the workshop on machine learning for multimodal interaction. ACM, New York, pp 44:1–44:4
Suzuki K, Morimoto I, Mizukami E, Otsuka H, Isahara H (2009) An exploratory study for analyzing interactional processes of group discussion: the case of a focus group interview. AI SOC 23(2):233–249
Tamburrini G (2009) Brain to computer communication: ethical perspectives on interaction models. Neuroethics 2(3):137–149
Tang H, Liao SS (2008) Discovering original motifs with different lengths from time series. Know Based Syst 21(7):666–671
Thomas J (2001) Collaborative innovation tools. In: Terano T, Ohsawa Y, Nishida T, Namatame A, Tsumoto S, Washio T (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 27–34
Tompa M (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In: ISMB, vol. 99, pp 262–271
Tompa M, Buhler J (2001) Finding motifs using random projections. In: Proceedings of the 5th international conference on computational molecular biology, pp 67–74
Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York
Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schroeder M (2012) Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans Affect Comput 3(1):69–87

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (CVPR 2001), vol. 1, pp I-511-I-518
Wachsmuth I, de Ruiter J, Jaecks P, Kopp S (eds) (2013) Alignment in communication. John Benjamins Publishing Company, Amsterdam
Waibel AH, Stiefelhagen R (eds) (2009) Computers in the human interaction loop. Springer, Berlin
Wallace R (2003) The elements of AIML style: ALICE (AI Foundation. http://www.alicebot.org)
Wallach W (2008) Implementing moral decision making faculties in computers and robots. AI SOC 22(4):463–475
Wallach W, Allen C, Smit I (2008) Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI SOC 22(4):565–582
Waterman MS, Arratia R, Galas DJ (1984) Pattern recognition in several sequences: consensus and alignment. Bull Math Biol 46(4):515–527
Weber W, Rabaey J, Aarts EH (2010) Ambient intelligence. Springer, Boston
Weise T, Bouaziz S, Li H, Pauly M (2011) Realtime performance-based facial animation. In: ACM SIGGRAPH 2011 Papers. ACM, New York, pp 77:1–77:10
Weizenbaum J (1966) ELIZA–a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
Whitby B (2008) Computing machinery and morality. AI SOC 22(4):551–563
Wiemann J, Knapp ML (2007) Turn taking in conversations: communication Theory. Transaction Publishers, NJ, pp 226–244
Wiese E, Wykowska A, Zwickel J, Müller HJ (2012) I see what you mean: how attentional selection is shaped by ascribing intentions to others. PloS one 7(9):e45391
Willsky AS (1976) A survey of design methods for failure detection in dynamic systems. Automatica 12(6):601–611
Wimmer H, Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13(1):103–128
Winograd T (1972) Understanding natural language. Academic Press Inc, Orlando
Woods WA (1970) Transition network grammars for natural language analysis. Commun ACM 13(10):591–606
Woods WA (1973) Progress in natural language understanding: an application to lunar geology. Proceedings of the national computer conference and exposition. ACM, New York, pp 441–450 June 4–8, 1973
Xu Y, Hiramatsu T, Tarasenko K, Nishida T, Ogasawara Y, Tajima T, Hatakeyama M, Okamoto M, Nakano Y (2007) A two-layered approach to communicative artifacts. AI SOC 22(2):185–196
Xu Y, Ueda K, Komatsu T, Okadome T, Hattori T, Sumi Y, Nishida T (2009) Woz experiments for understanding mutual adaptation. AI SOC 23(2):201–212
Yamaguchi O, Fukui K, Maeda K (1998) Face recognition using temporal image sequence. In: Proceedings of the 3rd International Conference on Face and Gesture Recognition. IEEE Computer Society, Washington DC, p 318
Yamashita K, Kubota H, Nishida T (2006) Designing conversational agents: effect of conversational form on our comprehension. AI SOC 20(2):125–137
Yamazaki A, Yamazaki K, Kuno Y, Burdelski M, Kawashima M, Kuzuoka H (2008) Precision timing in human-robot interaction: coordination of head movement and utterance. Proceeding of the twenty-sixth annual sigchi conference on human factors in computing systems (CHI '08). ACM, New York, pp 131–140
Yang M-H, Kriegman D, Ahuja N (2002) Detecting faces in images: a survey. IEEE Trans Pattern Anal Mach Intell 24(1):34–58
Yankov D, Keogh E, Medina J, Chiu B, Zordan V (2007) Detecting time series motifs under uniform scaling. In: Proceedings of the 13 ACM SIGKDD international conference on knowledge discovery and data mining, pp 844–853
Yngve V (1970) On getting a word in edgewise. In: 6th Chicago Linguistic Society, pp 567–578

Yonezawa T, Yamazoe H, Utsumi A, Abe S (2007) Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. Proceedings of the 9th international conference on multimodal interfaces (ICMI '07). ACM, New York, pp 140–145

Zentall TR (2003) Imitation by animals how do they do it? Curr Dir Psychol Sci 12(3):91–95

Zhang Q, Liu Z, Guo B, Shum H (2003) Geometry-driven photorealistic facial expression synthesis. Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on computer animation. Eurographics Association, Aire-la-Ville, pp 177–186

Zhang S, Wu Z, Meng H, Cai L (2010) Facial expression synthesis based on emotion dimensions for affective talking avatar. In: Nishida T, Jain L, Faucher C (eds) Modeling machine emotions for realizing intelligence, vol. 1. Springer, Berlin, pp 109–132

Zhang Z (1997) Parameter estimation techniques: a tutorial with application to conic fitting. Image Vis Comput 15(1):59–76

Zoric G, Smid K, Pandžić IS (2007) Facial gestures: taxonomy and application of nonverbal, non-emotional facial displays for embodied conversational agents. In: Conversational informatics. Wiley, pp 161–182

Index

T. Nishida et al., *Conversational Informatics*,
DOI: 10.1007/978-4-431-55040-2