Book Reviews

Robots that Talk and Listen

Judith A. Markowitz (editor)
(President of J. Markowitz, Consultants, Chicago, IL)

Berlin/Boston/Munich: Walter de Gruyter, 2015, hardbound, ISBN 978-1-61451-603-3; PDF, e-ISBN 978-1-61451-440-4 EPUB; e-ISBN 978-1-61451-915-7

Reviewed by
Martha Evens
Illinois Institute of Technology

I volunteered to review this book because I found Dr. Markowitz's 1995 book, *Using Speech Recognition*, extremely helpful when I first became involved in medical applications of speech. In fact, I made all my students in that area read it. This book is very different, but in its own way, just as useful. In fact, it is a must-read for anyone contemplating a new speech application or enhancing a current one, as well as for anyone searching for new directions in dialogue research.

This book is divided into four parts, followed by a separate conclusion. It is always hard to do justice to all the papers and authors in a collection like this—with twelve varied papers addressing many different parts of the complex problem of enabling a robot to carry on a spoken dialogue—but I will try to give at least a brief taste of each.

Part I is about "Images." In the first chapter of Part I, Steve Mushkin, the founder and president of Latitude Research describes an experiment in which 348 children aged 8–12 from six digitally-advanced countries were asked to draw and describe in words how they would like to interact with a personal robot in the future. In the second chapter Markowitz herself reviews the language capabilities and behavior of powerful cultural icons, such as Frankenstein, the Hebrew Golem, the Japanese karakuri, and Pygmalion's beloved statue, Galatea. This is highly relevant, because science fiction, whether on paper or in the movies, part of traditional mythology, or local social history, has largely shaped our expectations about what new technology promises or threatens. In the third chapter, David Dufty, author of the well-known book *How to Build an Android: The True Story of Philip K. Dick's Robotic Resurrection*, believes that robots are intrinsically part of art and entertainment, and that is the major reason why we should build them.

Part II is called "Frameworks and Guidelines." The first of the three chapters in this part, written by Bilge Mutlu and two of his students at the University of Wisconsin, Madison, discusses the requirements for designing a robot capable of sustaining productive dialogue. They test their framework in robot-human interactions; first using a robot as a teacher, then to assess linguistic factors that contribute to expertise. Director of Columbia University's Japanese Language Program, focuses on the knowledge that any teacher must have to be effective in teaching a foreign language. Based on her series of research studies on the teaching of Japanese, her concern is that a robot (or human) teacher must understand the bond between language and culture, which she sees as inseparable. She emphasizes the importance of using human-like facial expressions in the teaching machine, at first for teaching pronunciation, for communicating emotional

doi:10.1162/COLI_r_00224

reactions to the student's contributions, and making the student feel comfortable with the machine. In this excellent paper, chock full of valuable references, Nazikian unintentionally reminds us of the huge and inexplicable divide between the world of Computer-Aided Language Learning (CALL) and the world of AI in education. Art Graesser and his group at the University of Memphis (Link et al., 2001) have carried out a long series of experiments studying the effect of using expressive faces on the screen coordinated with the dialogue, but the work at Memphis is never mentioned in this paper and there are no references to papers in *Discourse Processes* or the *Journal of the Learning Sciences*. The third paper in Part II, written by Manfred Tscheligi and his student, Nicole Mirnig, at the University of Salzburg, focuses on the problems of Comprehension, Coherence, and Consistency. These qualities depend, they argue, on the use and transfer of mental models, which they find essential to effective dialogue with any robot capable of learning what human beings want and giving it to them.

Part III is about constructing a speech-enabled robot capable of useful, real-world tasks. In Chapter 7, Jonathan Connell of IBM's T. J. Watson Research Center explains how he defined these capabilities for a robot named ELI (the Extensible Language Interface). First of all, such a robot needs to be able to accept and synthesize multimodal information. It needs to recognize speech, gestures, and object manipulations and put this information together into an algorithm for getting you a cup of coffee? It has to find your kitchen, pick up the right coffee pot, heat the water in a microwave safe measuring cup, and put in just the right amount of coffee, cream, and sugar. In the process, it will need to learn new nouns (the names that you use for objects and rooms in your house) and new verbs (for activities like opening bottles and boiling water). Their robot parser is a finite state machine, but it can handle new names and new activities. The emphasis is on the grounding of new language in a new environment. In Chapter 8, Alan Wagner, a Senior Research Scientist at Georgia Tech, outlines what a robot has to know before you can teach it to tell lies or recognize lies produced by others. Unlike ELI, this robot does not yet exist and I find it hard to imagine wanting to build it or even use it myself, but the description of deceptive language in this chapter has charm. Joerg Wolf and Guido Bugmann, at the University of Plymouth in the UK, recount their experience getting university students to teach their off-the-shelf robot how to play a simple card game. They began by collecting a corpus of sessions in which one university student taught another to play the game. They used this corpus to develop the rules for their robot, grammar rules, rules for anaphora, dialogue rules, rules for dealing cards, and then they carried out a series of experiments in which university students tried to teach the robot the game. One of many problems that they had not anticipated was that when the robot did not understand what the student wanted him to do, the student raised his or her voice and tried to simplify the explanation. The result was many out-of-grammar and out-of-vocabulary errors.

Part IV contains two very different papers, one about improving audition (so that the robot can understand what you are telling it, even when several people are talking at once or machines are clanking) and the other about how design factors, such as robot voices and gesture speed affect memory and engagement in both children and adults. The experts in audition, François Grondin and François Michaud from the Robotics Laboratory at the Canadian Université de Sherbrooke, have developed algorithms for localization (figuring out where a sound is coming from), tracking (following sounds as people or robots move), and separation (extracting sounds from a given person or machine when several are making noises at once). They have achieved remarkable success in a number of experiments run on a system with a bank of separate microphones and

parallel hardware capable of processing all this information at once. Sandra Okita, then at Stanford, now at Columbia University, and Victor Ng-Thow-Hing from the Honda Research Institute USA in Mountain View teamed up to carry out a series of experiments with robot voices. The objective of their studies was to determine how design choices like robot voices and gesture speed effect how humans of all ages respond to a robot. They tried out two voices with young children, one robot-like (monotone) and one more human-like. Both robots followed exactly the same script, but the children remembered much more of what the humanoid voice said and they were much more likely to be willing to talk to that robot again. Teenagers preferred the humanoid voice, but there was much less difference between the two groups interacting with the different voices. Adults seemed to be affected even less. Similar experiments with gesture speed showed that robots who gestured rapidly were judged to be happier and more pleasant to spend time with. Again, small children were affected the most, but teenagers also had strong preferences here. Again, adults were affected less.

We badly need more work of the kind described by Okita and Ng-Thow-Hing. Too many people believed Reeves and Nass's (1996) Media Equation, which is the claim that people interact with computers in the same way that they interact with people; and decided that they could substitute research on human interaction for research on interactions between people and machines. Our own experiments (Bhatt, Argamon, and Evens, 2004) have convinced us that the Media Equation is certainly not true for the students interacting with our intelligent tutoring system, who are constantly polite to human tutors, whether their professors or their peers, but often very rude to our computer tutor.

The concluding chapter in the book, written by Roger K. Moore, Professor of Computer Science at the University of Sheffield and the Editor-in-Chief of Computer Speech and Language, takes us back to the future where the first chapter began. He argues that we have a long way to go to meet the goal of intelligent communication with machines. We need to build robots that understand human behavior and are capable of generating the language necessary to change it.

Moore essentially focuses this chapter on his own view of the major conflict of opinion and approach separating the experts featured in this book. It will never be enough to simply tack on a multi-purpose language module to an existing robot, he argues. Spoken communication is a fundamental and integral part of ordinary human beings; and if we are to succeed in our goal of constructing "intelligent communicative machines," we need to make the communication function a central part of the machine design and not simply an add-on function.

There is a major gulf between Moore's argument and the experts quoted by Dufty back in the third chapter (page 55). Dufty quotes Sylvia Solon, deputy editor of *Wired* in the UK, as saying, "There is no point making robots look and act like humans." Dufty adds that Martin Robbins, who blogs for the *Guardian*, under the name "The Lay Scientist," wrote: "Humanoid robots are seen as the future, but for almost any specific task you can think of, a focused simple design will work better." Of course, Dufty goes on to tell us about the use of robots in art and entertainment and about the workings of the Philip K. Dick android, so he may not agree entirely with Solon and Robbins.

There is a cogent case on the opposite shore of this gulf in at least three of the other chapters (4, 7, and 9). In Chapter 4, Mutlu's group at the University of Wisconsin describe the simple semantic grammar used by their robot (a Wakamaru), along with more sophisticated dialogue and behavior models and algorithms to control gaze and gestures. Jonathan Connell begins Chapter 7 with "Suppose you buy a general fetch-and-carry robot from Sears and take it home" and continues in this vein. In

Chapter 9, Wolf and Bugmann designed their system to operate on components made of production rules, but they still have a lot to teach us.

Personally, I am on a raft drifting in the middle of this gulf. I think that we need thoughtful experiments with existing robots now, to help us discover the problems, but I believe that really satisfactory solutions will require much more research. I cannot agree with Moore, however, when he argues that it is unethical to attempt to imitate human communication by stitching together the limited technologies we have today. It seems to me that much of what we now know about human communication comes from attempting to do just that.

In the process of reading this book I discovered that I had not heard about two other collections edited by Markowitz and published by Springer in 2013. Both of these were edited by Neustein & Markowitz jointly. Dr. Amy Neustein is founder and CEO of Linguistic Technology Systems and Editor-in-Chief of the International Journal of Speech Technology. The first is entitled *Mobile Speech and Advanced Natural Language Solutions* (hardbound ISBN 978-1-4614-6017-6; e-Book ISBN 978-1-4614-6018-3); the second, *Where Humans Meet Machines: Innovative Solutions for Knotty Natural Language Problems* (ISBN 978-1-4614-6933-9, eBook ISBN 978-1-4614-6934-6).

References

Bhatt, K.; Argamon, S.; and Evens, M. W. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In *Proceedings of COGSCI 2004*, Chicago, IL. 114–119.

Link, K. E.; Kreuz, R. J.; Graesser, A. C.; and the Tutoring Research Group. (2001). Factors that influence the perception of feedback delivered by a pedagogical agent. *International Journal of Speech Technology*, 4, 145–153.

Reeves, B. and Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge, UK: Cambridge University Press.