The challenge of conversational machines: from MUSA to Jibo

Roberto Pieraccini
Director of Advanced Conversational Technologies, Jibo

ABSTRACT

About 35 years ago I built my first conversational system at CSELT in Italy. It was based on a PDP 11-60 computer, a FP-100 Array Processor, a custom built AD/DA converter, and MUSA, one of the first standalone Text-to-Speech systems that had made the national news only a few years earlier. Since then my career brought me back and forth between research and industry on a quest for building systems that would converse the way we do among humans, within realistic tasks of arbitrary complexity, and outside a controlled laboratory environment.

My driving philosophy through all these years has been that of trying to automate as much as possible the development of such conversational systems. My work on statistical language understanding in the early 1990s, and learning dialog strategies a few years later, were among the first attempts to effectively use machine learning right there where handcrafted rules had been "the rule." But that was still in the research realm.

In 1999 I joined SpeechWorks—one of the two main *ancestors* of what is now known as Nuance. There I had my first direct experiences in building real systems for real customers, and there I realized for the first time the importance of Voice User Interface (VUI) design in copying with the idiosyncrasies of a then—and still today to a certain extent— quite imperfect speech recognition technology. At SpeechWorks I understood the power of a well-defined development process, and the overall difficulty inherent in bringing a system outside of the labs and in front of thousands of users.

After a short interlude back in research at IBM, I joined an early stage startup called SpeechCycle as their CTO. At SpeechCycle I was at the front line in trying to bring to life complex systems for demanding and paying clients to serve hard to please, and often recalcitrant, final users. The payback in terms of experience towards the productization of real systems and the opportunity to have available much more data than what we, speech, language and dialog scientists, could ask for at that time was priceless. With hundreds of thousands of weekly calls per system, and very clear and measurable optimality criteria, my team and I had the luxury of being able to apply statistical machine learning techniques to high volume production systems, and actually see the caller experience and task completion improve before our very eyes.

After another interlude in advanced research at ICSI, I went back to what I consider my biggest conversational challenge until now: building Jibo, the first consumer social robot for the home.

Jibo is an extremely complex device that includes both embedded and cloud speech recognition, natural language understanding, text-to-speech, and speaker identification, but not only. Jibo has a moving body that helps him communicate more effectively, express emotions, and create social bonds with his users. Jibo has cameras and microphones to make sense of the world around him, including detecting where sounds come from and recognizing and tracking faces and people. He has a display to show images, words, pictures, and videos, an eye that can morph into shapes at will, and a touch interface that provides a complementary input modality. Jibo encompasses the ultimate human-machine interface, with the potential of becoming a new interaction paradigm. And, even more important, Jibo is a platform with a development system that can be creatively used by third parties to build and distribute different applications.

In this talk I will take the audience through my journey in conversational technology, from the first years of bulky and slow computers, for which the recognition of the ten digit in real time was a task of gargantuan proportion, to Jibo, a social robot that has the promise of being a first significant step towards a more pleasant, accessible, and human interaction with machines. I will conclude with the lessons learned, and how they shaped my vision for the evolution of conversational technology now and for the next few years.

ABOUT THE SPEAKER

Roberto Pieraccini, a scientist, technologist, and the author of “The Voice in the Machine,” (MIT Press, 2012) has been at the forefront of speech, language, and machine learning innovation for more than 30 years. He is widely known as a pioneer in the fields of statistical natural language understanding and machine learning for automatic dialog systems, and their practical application to industrial solutions. As a researcher he worked at CSELT (Italy), Bell laboratories, AT&T Labs, and IBM T.J. Watson. He led the dialog technology team at SpeechWorks Int.l, he was the CTO of SpeechCycle, and the CEO of the International Computer Science Institute (ICSI) in Berkeley. He now leads the Advanced Conversational Technologies team at Jibo. http://robertopieraccini.com