IEEE CIS

AMD Newsletter

Robotics
Psychology
Neuroscience
Machine Intelligence

The Newsletter of the Autonomous Mental Development Technical Committee

VOLUME 2, NUMBER 2 ISSN 1550-1914 OCTOBER 2005

Message From The Editor

Welcome to a new issue of the AMD TC Newsletter. In the past few years, we have been seeing steady and encouraging development of the AMD research community, including the regular annual conference, the growth of the AMD TC membership, and the forthcoming special issue on AMD in IEEE Transactions on Evolutionary Computation, among many, many other things. The AMD TC Newsletter has been and will be continuously communicating new developments and hot research topics to her audience. Featured in this Newsletter issue is a dialog column guest-moderated by Stephen Levinson on speech and language, in addition to our other regular columns. Enjoy!

-Yilu Zhang, Editor of the AMD Newsletter

Committee News

- July 20, 2005: The AMD Technical Committee (TC) Annual Meeting was held during the 4th ICDL in Osaka, Japan. The participants were TC members and the major organizers of the 4th ICDL. The TC chair Juyang Weng chaired the meeting. The participants applauded the successful organization of the 4th ICDL by Minoru Asada, Koh Hosada, and many others. Olaf Sporns, as the representative of the 5th ICDL (2006) organizers, presented the preparation of the forthcoming conference. Weng summarized the written TC progress report to the IEEE Computational Intelligence Society.

- August 4, 2005: The progress report of the AMD Technical Committee was presented at the Annual Adcom Meeting of the IEEE Computational Intelligence Society by the TC Chair Juyang Weng.

Dialog Column

Dialog: **Can Robots Learn Language the Way Children Do**?

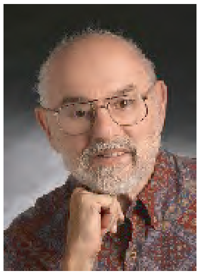
Steve Levinson, Department of Electrical and Computer Engineering and Beckman Institute,
University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA

Speech recognition machines are in use in more and more devices and services. Airlines, banks, and telephone companies provide information to customers via spoken queries. You can buy hand-held devices, appliances, and PCs that are operated by spoken commands. And, for around $100, you can buy a program for your laptop that will transcribe speech into text. Unfortunately, automatic speech recognition systems are quite error prone, nor do they understand the meanings of spoken messages in any significant way. I argue that to do so, speech recognition machines would have to possess the same kinds of cognitive abilities that humans display. Engineers have been trying to build machines with human-like abilities to think and use language for nearly 60 years without much success. Are all such efforts doomed to failure? Maybe not. I suggest that if we take a radically different approach, we might succeed. If, instead of trying to program machines to behave intelligently, we design them to learn by experiencing the real world in the same way a child does, we might solve the speech recognition problem in the process.

Do you consider this a reasonable approach to the problem of speech recognition? If not, why not? If so, what experimental steps could be taken to demonstrate the principle?

Dialog Column

Reply to Dialog: "Can Robots Learn Language the Way Children Do?"



Stan Franklin, Department of Computer Science and the Institute for Intelligent Systems,
University of Memphis, Memphis, TN 38152, USA

My software agent, IDA (Intelligent Distribution Agent), negotiates in unstructured natural language (English) with sailors in the process of finding new jobs toward the end of their current tour of duty [2]. Due to the relatively poor quality of speech to text transcription systems, IDA communicates with sailors by email. She doesn't learn at all. Her language generation is done by means of filling in blanks in prepared scripts and ordering them. She "understands" incoming emails in that she knows how to pick out significant content, and how to act on it. Glenberg [4] and others would argue that we humans also understand in that sense.

More recent work by my research team is aimed at transforming the IDA technology into a domain independent Learning IDA (LIDA) technology capable of learning in human-like ways [1]. This includes perceptual learning [3], episodic learning [5], and procedural learning. Our intent is to use the LIDA technology to control real-world, cognitive robots that would learn continuously, and pass through a developmental period as a human child does.

Could such a LIDA controlled robot "solve the speech recognition problem" as Levinson suggests? I believe that it, or similar learning robots, could in principle do so, provided certain conditions are met:

1. The control architecture must, as LIDA does, model a broad swath of human cognition.
2. The robot must learn to speak as well as to understand.
3. The robot's senses, actuators, and motivations must be specifically designed to be suitable for this task.
4. The robot's developmental environment must include human children with whom it could "play."

References:
[1] S. K. D'Mello and S. Franklin, "A cognitive architecture capable of human like learning." (submitted)
[2] S. Franklin, "A 'Consciousness' Based Architecture for a Functioning Mind." In *Visions of Mind*, ed. D. Davis. Hershey, PA: Information Science Publishing, 2005.
[3] S. Franklin, "Perceptual Memory and Learning: Recognizing, Categorizing, and Relating," presented at American Association for Artificial Intelligence (AAAI) Symposium on Developmental Robotics, Stanford University, Palo Alto CA, USA; March 21-23, 2005.
[4] A. M. Glenberg, "What memory is for." *Behavioral and Brain Sciences* vol. 20, pp. 1-19, 1997.
[5] U.Ramamurthy, S. K. D'Mello, and S. Franklin, "Modified Sparse Distributed Memory as Transient Episodic Memory for Cognitive Software Agents." In *Proc. of the International Conference on Systems, Man and Cybernetics*. Piscataway, NJ: IEEE, 2004.

Reply to Dialog: "Can Robots Learn Language the Way Children Do?"

Dave Touretzky, Department of Computer Science and Center for the Neural Basis of Cognition,
Carnegie Mellon University, Pittsburgh, PA 1521, USA

> *"If, instead of trying to program machines to behave intelligently, we design them to learn by experiencing the real world in the same way a child does, we might solve the speech recognition problem in the process."*

There are so many things wrong with this statement, it's hard to know where to begin. It opens with a false dichotomy: contrasting "programming" with "design." Programming is difficult and error-prone because one must be very explicit about every step of the computation being performed. "Design," on the other hand, is comfortably vague: one can paint with broad strokes, focusing on fundamental principles. Unfortunately, design must be followed by implementation - and that requires programming. Second, the proposal begs the question of what it means to “experience” the world. One can easily plug a microphone and webcam into a computer and place it in someone’s living room to “observe” the life of the household. But where is the homunculus that will be looking at these images and listening to the accompanying dialog? We don’t even have a design for such a thing. We don’t know what goes on in children’s heads.

There has been recent work in the machine learning community by Tom Mitchell and others on exploiting large quantities of unlabeled training data in conjunction with smaller quantities of labeled data. Child language learning might be seen as an example of this. Normal children learning to speak receive very little direct instruction; most of their linguistic “training” comes from experiencing family interactions. Holding up objects in front of a TV camera while speaking their names would be analogous to direct instruction - and assumes we’ve already solved the vision problem so that the computer can recognize what it’s looking at. Letting the computer passively observe human interactions would be a more naturalistic approach. But children are not passive learners. They speak, and observe the effects of their utterances. So perhaps we will need to equip the computer with a sound card as well. This sounds suspiciously like what Richard Feynman called “cargo cult science:” if we understand nothing about the design and operation of an object, but we imitate its outward appearance, somehow that should magically be enough to replicate its function. It didn’t work for the South Sea Islanders, and it’s unlikely to work here.

The real problem with fluent speech recognition is that the noisy and ambiguous audio signal lacks sufficient information to support flawless transcription. Human beings rely extensively on context, and vast amounts of world knowledge, to fill in the missing bits. Children acquire this knowledge, and the ability to exploit it, through a cognitive development process lasting more than a decade. Understanding this process is a far more difficult problem than building statistical speech recognizers. To reiterate: we don’t know what goes on in children’s heads.

Audition and Language are Tightly Intertwined in Autonomous Development

Juyang Weng, Embodied Intelligence Laboratory, Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA

Existing applications of speech recognition devices, like applications of visual inspection devices, have been limited to constrained environments (e.g., a single speaker on a specific subject) and have been error prone as Levinson stated. Yes, simulating autonomous development by embodied robots is not only a reasonable approach to speech recognition, but also an opportunity to go beyond the status quo. Autonomous development is the only way through which humans acquire speech recognition capabilities.

However, autonomous development is not a superficial simulation of child learning using traditional machine learning approaches (e.g., traditional learning in existing speech recognition systems). It is true that the field does not yet have the complete knowledge about how the human brain develops, but some major characteristics are known. Some major principles set autonomous development apart from traditional machine learning:

First, fully autonomous internal self-organization is necessary. The brain fully self-organizes internal representation throughout its development. However, a typical traditional speech recognition method requires the human trainer to manually assign the function (e.g., for recognizing a specific word) to every internal model (e.g., HMM). This assignment also requires him to control the system internal signal flow so that every internal model listens only to signals from its assigned class during training, but it listens to signals from all classes during performance. This practice of partial manual development limits the capability of adapting to open, unconstrained auditory environments (e.g., multiple speakers) because of the limitation of human static manual organization. In contrast, our developmental audition program [1] dynamically self-organizes the auditory input space so that there is no need for the human trainer to manually assign the function to each representation component, or to manually manage the internal signal flow.

Dialog Column

Second, a typical traditional method uses symbolic representation (e.g., a symbolic word) between the speech recognition stage (early processing) and the language stage (later processing), but an autonomous developmental program should not. Autonomous development must self-generate representation for new and unknown tasks; but any design of symbolic representation requires human manual work. Therefore, the development cannot be autonomous if symbolic representation is used for internal representation. Symbolic abstraction of human auditory capability is demonstrated by human behaviors (e.g., producing a required class of behaviors), not necessarily supported by a symbolic internal representation. The richness of sensory context needs to be conveyed from the early cortical processing stage to later cortical processing stages for robust cognition. The nervous pathways between speech recognition and language processing are of a significantly higher dimension than what a symbolic representation can handle. In our developmental program [1], high dimensional pathways are used between the early and later processing stages.

Third, audition and language acquisition are tightly intertwined throughout development. The programmer of a traditional language processing system needs to model language-specific syntax and semantics. However, this is not the case with the programmer of a developmental program. He does not need to understand the syntax and semantics of a language that the robot ends up learning. The capability of understanding and using a language (auditory, visual, or written) is, in essence, the capability of handling the association of longer sequences of multimodal sensory and motor contexts. Language acquisition is a natural outcome of the development of grounded sensory and motor processing. It is crippled without grounded sensory and motor experiences. Human language-specific knowledge, including semantics and syntax, consists of real-time multimodal associations between the last context and the future possible contexts, learned through context-dependent experiences. With this view, Weng [2] proposed a theory and a developmental architecture for dealing with reasoning and planning in any mode (auditory, visual, or written), with language acquisition as a special case of many other possible applications. Since sophisticated language processing capabilities are not fully developed until later stages of development, the realization of language acquisition should be first demonstrated in early development, for small vocabulary, syntax-free but semantics-rich simple-language settings, e.g., action chaining (i.e., skill transfer) demonstrated by Zhang & Weng [3].

References:

[1] Y. Zhang, J. Weng, and W. Hwang, "Auditory Learning: A Developmental Method," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 601-616, 2005.

[2] J. Weng, "A Theory of Developmental Architecture," in *Proc. 3rd International Conference on Development and Learning (ICDL 2004)*, La Jolla, CA, Oct. 20-23, 2004.

[3] Y. Zhang and J. Weng, "Action Chaining by a Developmental Robot with a Value System," in *Proc. IEEE 2nd International Conference on Development and Learning (ICDL'02)*, Cambridge, MA, pp. 53-60, June 12-15, 2002.

Response to "Dialog: Can Robots Learn Languages the Way Children Do?"

Xie

Kandhasamy

Leong

Ming Xie, Jayakumar Sadhasivam Kandhasamy, and Kok Heng Leong,
School of Mechanical & Aerospace Engineering,
Nanyang Technological University, Singapore 639798

Human beings can virtually learn all the human languages. But, no animal can do so. For instance, studies show that a monkey, or ape, can only master a vocabulary of about 400 to 600 words, in maximum, of a natural language regardless of the effort and duration of training and learning. And, it is a common fact that no domestic animal could master a human language at a reasonable level regardless of how long it lives with its human master. These observations strongly suggest that a biological creature must be designed in an appropriate manner in order to be innately apt to learn human languages.

The importance of design could also be appreciated from the examination of the question of what we mean by learning. In [1], we define "learning" as a process which consists of three steps: a) modeling (i.e. design), b) optimization (i.e. to learn, or to be trained), and c) representation (i.e. organized properties and constraints as the substance of knowledge and skill). Clearly, the modeling aspect of learning depends on design in one way or another. Then, the fundamental question will become: Will human designers be able to design a robot and to make it be innately capable of learning human languages the way children do? Before we make an attempt to answer this question, let's look at these two scenarios:

Let us imagine that we confine a new-born baby to grow up in a room which is only equipped with a screen, a desk, and other necessary facilities for living purposes. When the baby reaches a certain age, we start to let him or her learn a natural language. After a long enough period of memorizing the symbols displayed on the screen, we decide to let him or her learn the sentence "An airplane is a vehicle that is able to fly high, and a fish is an animal that is able to swim fast" by displaying the text on the screen. At a certain point in time, we ask him or her this question: "Who is able to swim?" or, "who is able to fly?" Most likely, the child will give the correct answer. But, surely, the child will not understand the physical meanings of the words in his or her answer, because he or she has never seen the movement of flying or swimming.

In the second scenario, assume that a person has been blind and deaf since birth. When he or she attains a certain age, we teach the child a natural language. For example, we can let him or her touch a pen first, and then touch the Braille dot which represents the word "pen." We repeat the exercise until the child knows how the pen is represented by the Braille dot. Now, we ask him or her the question "what is this pen?" It is certain that he or she will not be able to give a full description of the pen (e.g. the pen's color).

These scenarios show that language learning should not be simply playing with symbols or dotted patterns. The most important aspect of language learning is to associate symbols (or a dotted pattern) with physical meanings. As physical meanings refer to the properties and constraints of entities in the physical world, it is possible to design an organized memory for a robot to represent meanings in the form of a hierarchical architecture of networks such as object network, agent network, behavior network, event network, episode network, lexical network, concept network, and topic network, etc. In other words, we believe that it is possible to design a universal meaning-centric representation for natural languages.

As autonomous learning consists of modeling, optimization, and representation, future success in the design of appropriate model and representation for natural languages will be a great advance toward achieving the goal of making a robot capable of mastering a human language.

References:

[1] M. Xie, J. S. Kandhasamy, and H.F. Chia. "Meaning-centric Framework for Natural Text/Scene Understanding by Robots," *International Journal of Humanoid Robotics,* Vol. 1(2), June 2004.

[2] J. S. Kandhasamy, "Organized Memory for Natural Text Understanding and Its Meaning Visualization by Machine." *PhD thesis (Under Review),* School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, 2005.

Dialog Column

Can machines learn and use contextual information?

Yilu Zhang, Electrical and Controls Integration Lab, R&D Center,
General Motors Corporation, Warren, MI 48090 USA

To successfully exchange information, the parties involved in communication use a great deal of contextual information that is either embedded in the speech environment or has been acquired from previous sensory experiences. Humans use information such as a mutually understood discussion topic to understand each other in situations such as a noisy cocktail party. Many practical automatic speech recognition (ASR) systems also have built-in task-specific heuristics or constraints. For example, a telephone number recognition system has the knowledge of non-existing area codes. The difference is that humans seem to acquire the background knowledge and the skill of using this information naturally and easily from a very young age, while it takes a lot of effort for engineers to build the knowledge and the skill into a machine. In many cases, the engineering efforts are application-specific and largely ad-hoc.

Can the knowledge-building process by machines be generic and systematic? It may. But it relies on at least two important mechanisms that are observed in child development. The first one is multimodal learning. In the process of cognitive development, children take in and integrate the information from all the senses - sight, hearing, smell, touch, and taste. There is evidence showing that if visual, auditory, and tactile inputs never have the chance to occur together, there is no opportunity to develop an integrated knowledge between what is seen, heard, and felt [1]. The second important mechanism is grounding. Grounding means that representations inside an agent should be connected to their references in the external world [2]. For example, the representation of "dog" should be related to the presence of actual dogs in the environment. Grounding is accomplished through real-time sensory experiences.

We have been making some initial pitches to equip a machine with the mechanisms discussed above. In [3], we built a robot that learned to follow verbal instructions based on raw sensory experiences and direct interactions with the environment, which is usually termed as grounded learning. Starting from no knowledge of any auditory stimuli it was going to perceive or the behavior it was going to develop, the robot managed to learn more than a dozen auditory commands, such as raising arm and moving around, within half an hour. In [4], we extended our work from auditory learning to multimodal learning. Mimicking the way a human child learns the concept of the object, the robot learned to correctly answer verbal questions about the properties (such as the size and the names) of the objects it saw. In other words, the robot learned to use both the auditory and visual cues embedded in the environment to improve communication.

What we have achieved so far seems to be limited compared to the commercially available ASR systems in terms of vocabulary size. However, the "radically different approach" (Levinson) we have been taking has the potential to save the large amount of engineering efforts devoted to data collection, transcription, and parameter tuning in the traditional approach. Further, it potentially enables the autonomous learning and usage of the contextual information which may eventually facilitate robust human-like speech communication capability by machines.

References:

[1] B. Bertenthal, J. Campos, and K. Barrett, "Self produced locomotions: an organizer of emotional, cognitive, and social development in infancy." In R. Emde and R. Harmon, (Eds.), *Continuities and Discontinities in Development*. Plenum Press, New York, NY, 1984.

[2] S. Harnard, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335-346, 1990.

[3] Y. Zhang, J. Weng, and W. Hwang, "Auditory learning: a developmental method," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 601-616, 2005.

[4] Y. Zhang, and J, Weng, "Conjunctive visual and auditory development via real-time dialogue," in *Proc. of The Third International Workshop on Epigenetic Robotics*, Boston, MA, August 4-5, 2003.

Dialog Column

Reply and Summary: "Can Robots Learn Language the Way Children Do?"

Steve Levinson, Department of Electrical and Computer Engineering and Beckman Institute,
University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA

The question and premise to which our respondents have given answers was first posed by none other than A. M. Turing in his famous 1950 paper in which he outlined the "Turing Test" for intelligence. The question of mental development doesn't appear until the penultimate paragraph in which he specifically addresses the question of language acquisition with the aid of vision. Interestingly, he does not include motor function explicitly. He concludes the paragraph by saying that he does not know whether fixed designs or adaptive systems are better but that both should be tried [1].

It appears that all of our respondents agree with Turing that both design and learning are necessary. The open question is which functions should be explicitly designed and which should be acquired, leaving open the possibility that all have some fixed and some learned aspects. Even the fixed designs may be adaptive over evolutionary time scales while the learned abilities are acquired in somatic time. I think that there is also a strong case made by the respondents that functions should not be considered in isolation. The brain/mind integrates all sensorimotor functions.

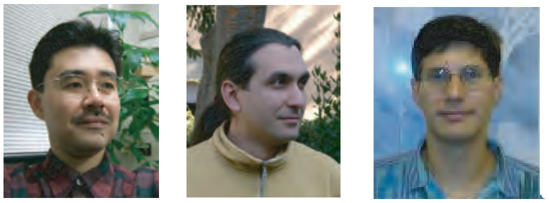
My own research follows these assumptions. My experiments with language acquiring robots are detailed in the final two chapters of my recent book [2]. I find it quite exciting to work in our AMD community and be part of this new and promising scientific enterprise.

References:
[1] A. M. Turing, "Computing Machinery and Intelligence," *Mind* pp. 433-460, 1950.
[2] S. E. Levinson, *Mathematical Models for Speech Technology*. West Sussex, UK: John Wiley and Sons, Ltd., 2005.

Conference Reports

A Really Hot Summer in Osaka

ICDL 2005 Organizers



Hosada Metta Deak

Koh Hosoda, Department of Adaptive Machine Systems,
Osaka University, Suita, Osaka 565-0871, Japan

Giorgio Metta, LIRA-Lab, University of Genova,
16145 Genova, Italy

Gedeon O. Deak, Department of Cognitive Science,
University of California, San Diego, CA 92093-0515, USA

ICDL '05, the fourth International Conference on Development and Learning, was held during the scorching heat of summer in Osaka, Japan, July 19-21, 2005. The site, INTEX Osaka, was located in the Northern Osaka bay area, easily accessible from the city center of Osaka. The event was sponsored by the IEEE CIS society, and conference papers are available on the IEEE Xplore website (http://www.ieeexplore.ieee.org). We enjoyed invited talks from a prominent and diverse group of speakers from Europe, the USA, and Asia: (in order of appearance) Claes von Hofsten (psychology, Uppsala U., Sweden), Yasuo Kuniyoshi (robotics, U. of Tokyo, Japan), Toshio Inui (psychology, Kyoto U., Japan), and Joseph J. Campos (psychology, U. of California, USA). The conference was held right after the Ninth RoboCup International Competitions and Conferences (RoboCup-05, at the same site), and right before the Fifth International Workshop on Epigenetic Robotics (EpiRob-05, in Nara, Japan), which attracted more than 100 participants from 14 countries.

Conference Reports

The meeting was kicked off by a workshop, "Social Cognition: From Humans to Robots," organized by Gordon Cheng (ATR, Japan) and a tutorial, "Autism: social communication disorders," by Hideki Kozima (NICT, Japan). Following the workshops there was a joint session with the RoboCup Symposium, featuring talks by Pat Langley (Stanford U., USA) and Giorgio Metta (U. of Genova, Italy). We designed the technical sessions to be consecutive, not concurrent, so as to gather all the participants and generate fruitful interdisciplinary discussion, following the model of ICDL 2004 in San Diego, California. There were 20 oral talks and 33 posters. Thanks to the effort of PC members and reviewers, each paper was carefully reviewed by referees from three regions: Europe, the USA, and Asia. This helped ensure the high quality of papers and presentations. All posters remained in the conference room for the duration of the meeting, which allowed participants to access them at their convenience. Thursday afternoon we enjoyed a panel discussion with the invited speakers and the General chair, Minoru Asada, and the conference was concluded.

Call For Papers

ICDL 2006
The Fifth International Conference on Development and Learning
May 31 – June 3, 2006, Bloomington, Indiana, USA
Submission Deadline: February 6, 2006
http://www.icdl06.org

WCCI 2006 IEEE World Conference on Computational Intelligence
A joint conference of the **IEEE Congress on Evolutionary Computation CEC**
IEEE International Conference on Fuzzy Systems FUZZ-IEEE
International Joint Conference on Neural Networks IJCNN
July 16 – 21 Sheraton Wall Center, Vancouver, BC, Canada
http://www.wcci2006.org

Deadlines for all three WCCI conferences:
Special Session proposal: December 31 2005
Tutorial proposal: January 31 2006
Paper Submission: January 31 2006

WCCI 2006 Special Session on Autonomous Mental Development
Session Chairs: Brian Scassellati (scaz@cs.yale.edu) and John Weng (weng@cse.msu.edu).
Visit the WCCI 2006 web page for more information.

Glossary

Speaker Independence: This term refers to the properties of automatic speech recognition systems which allow them to transcribe the utterances of any speaker without specific training for that individual. No automatic system is speaker independent to the extent that human listeners are. Machines do poorly with children's voices and regional dialects. Many speech recognition systems are generically trained for all voices but use an enrollment process to adjust to a specific speaker for maximum performance.

Dynamic Time Warping: This term refers to a method of making automatic speech recognition systems insensitive to small local variations in the timing of spoken utterances. The process implicitly aligns the phonetic boundaries of different utterances of the same words so that only like sounds are compared. The method is usually based on the well-known method of dynamic programming.

- Supplied by Steve Levinson

Editor: Yilu Zhang **Editorial assistant: Kim Thompson** **Logo: Joe Zimmerman**