The Robot Baby meets the Intelligent Room

David M. W. Powers

School of Informatics and Engineering
Flinders University of South Australia
powers@ieee.org

Introduction

The world of science fiction has long known computers, robots and spaceships that can converse like humans, pass the Turing Test with ease, and interact with their environments with superhuman intelligence. There is an entire spectrum from a Data to a HAL to the ubiquitous Star Trek computer that will dim your lights, serve you tea - or commence the autodestruct sequence. Many of these, like Data, HAL and Astroboy have learned most of their knowledge, including their language capabilities, rather than being programmed [Clark, 1972; Ishiguru 1962], paralleling both early and recent research into the acquisition of language and ontology [McCarthy, Earnest, Reddy and Vicens, 1968; Block, Moulton & Robinson, 1975; Feldman, Lakoff, Stolcke and Hollback Weber, 1990; Steels and Brooks, 1995].

This paper describes a program of research into language acquisition using robot babies and intelligent rooms, and summarizes some preliminary results and applications.

Grounding a Baby

The initial motivation for using real or simulated robots, or embedded computers, for natural language research has been to ground semantics through sensory-motor interaction with a real or simulated environment, sometimes without any ambitions towards learning [Winograd, 1973], but most commonly with the aim of acquiring a realistic semantics and specifically contrasting with approaches that develop a pseudo-semantics or pseudo-ontology that is essentially a thesaurus that groups related words without any claim to understanding them.

Even in current research the simulated robot still dominates the real robot for language learning and semantic modeling, as using real robots and real sensors involves dealing with a great many hardware and sensory-motor issues tangential to the objectives of the project. The experiments with real robots tend therefore to operate with cockroach level robots without auditory or visual capabilities. Even today with cheap videocams, video-processing software and speech recognition software commercially available, it is difficult to make effective use of these resources when the outputs are not what is desired and the modules do not provide the hooks and interfaces to allow the associative learning experiments to be made at different levels of processing.

Many researchers have found simulated worlds and cockroach robotics adequate for exploring the meaning of individual words, and have succeeded in mapping or characterizing the meanings of selected prepositions and verbs in a number of languages. Other researchers [Steels, 1996/97] have explored what kind of language will automatically emerge amongst cooperating robots. Indeed we turn the language learning paradigm on its head, and do not assume that there is a particular standard target language to learn. Rather each learner of "English" actually learns a slightly different idiolect - the learning process is rather seen as "creative invention of language" even in this case: there is no language "English" in any objective sense [Powers, 1985/89; Yngve, 1996]. Our learning model is a growing model in which both the ontology and idiolect develop as an interactive negotiation, socialization and conventionalization process.

One of the problems with simulated environments is that they are simplistic and tend to lead to highly supervised learning paradigms. For example, a major issue is the association of a specific word with a specific scene (or worse some specific part of a representation of a scene). In real environments, various mechanisms are employed to direct attention to a particular part of the scene or a particular word in a sentence, but this is far more subtle. Attention is thus a particular focus of current research into robot learning of semantics [Steels, 1997; Homes, 1998; Kozima and Ito, 1998; Hogan, Diederich and Finn, 1998].

Finally we come to syntax. A major question is whether grammar is learnable in the absence of grounding. Cognitive Linguistics explores the premise that linguistic processing reuses structures and mechanisms, so that linguistic processing is analogous to non-linguistic sensory-motor processing in a deep sense [Deane, 1992].

Previous experiments have demonstrated unsupervised learning of syntactic structure up to the level of noun and verb phrases using solely word, character, phoneme, or speech code vector input [Powers, 1983/91/92]. Classes like noun, verb and preposition can be self-organized

without multi-modal input, and word order and cohesive constraints (e.g. agreement) can be learned by a simple constraint parser [Entwisle and Groves, 1994] but this has not yet been demonstrated in a completely unsupervised paradigm. Thus a major objective of this program is to explore whether multi-modal input, and implicit supervision, can produce more effective syntactic learning. In this work, the same learning mechanisms are being investigated for all perceptual structures, both for intra-modal and cross-modal associations.

Investigating a Baby

Psycholinguistic research is largely built on a program of brief sessions with infants - except for those few who have published a comprehensive study of their own children. A researcher doing a longitudinal study of a child may only have an hour a week of data. Even then, the researcher is primarily focussed on what the child is saying, and glossing it with interpretations that are relatively subjective.

Active experiments to see what the child understands can be conducted, but require careful design - the production side of child speech has received immensely more attention than comprehension. Experiments on how the child learns new words can also be designed, typically using nonsense words - but this kind of experiment actually influences the child's learning and changes the language the child ends up learning (memory for introduced/artificial terms is quite persistent, and a word trained in a single session been detected over a decade later). Moreover, this data concentrates only on the child's capabilities or the parent's interaction during the interview or experiment. To develop an effective model of a child's learning we really need to know everything that the child has experienced, linguistic and non-linguistic.

Another major aim for our robot baby design has therefore been to allow the capture of full sensory-motor/audio-visual records from the perspective of an interacting child. This can take two forms: mothering of an intelligent doll by a child, and parental interactions with a child monitored by the doll and the room.

Designing a Baby

Both for purposes of data collection and interactive learning, the robot baby needs to be supplied with a variety of sensor-motor capabilities. We want to go beyond what can be achieved by simply videotaping a formal session with a child, or even spontaneous interactions between a child and his family/environment. Ideally we want an audio-visual data stream from the child's perspective. Additionally it is useful to have an audio-visual data stream from an external perspective. Our simulated world originally provided for each object to have arbitrarily located eyes, and eyes were provided as standard in front and above the stage.

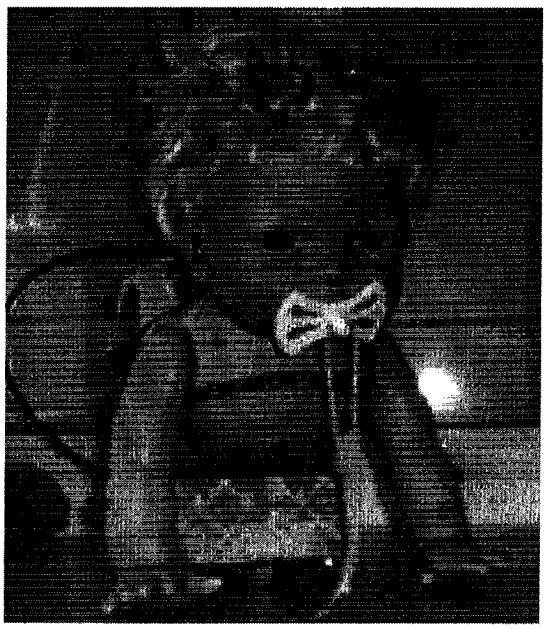

Figure 1. *The first physical implementation of the robot baby has microphones in its ears, crude switches for touch, independent control of the head and each limb, and internal sensors for orientation and acceleration/shock. The electronics is controlled by a 6809HC11 microcontroller.*

Our initial robot baby is designed with multiple electret microphones, touch sensors and motors (one per limb plus one for the head), as well as acceleration, shock and orientation sensors. Currently this information is subject to extreme bandwidth limitations and we can collect only sensory and trivially compressed 8-bit 8kbs stereo audio through the internal 6809HC11 microcontroller (Fig.1).

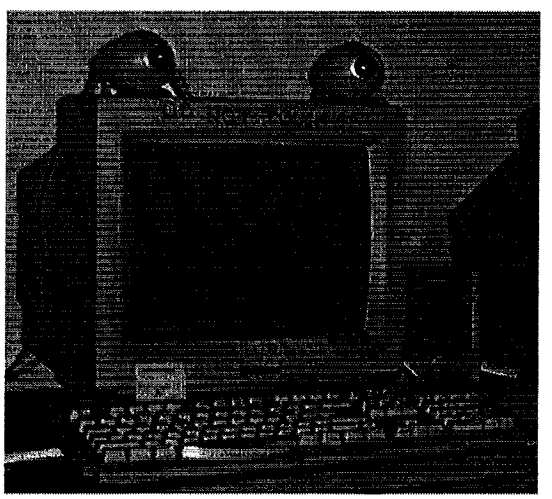

Figure 2. *Philips USB videocams are currently used externally for vision.*

Comprehensive experiments with direct data collection have been performed by an array of Windows PCs using a tetrahedral microphone array (16-bit 22kbs, designed to fit ears, nose and crown of our new doll) in a room wired with two ceiling mikes (16-bit 22kbs), two directional USB mikes (16-bit 44kbs) and two USB videocams located on a monitor (Fig. 2) 1 to 2m. in front of the subject (16-bit 44kbs 20fps 352x288).

This testbed serves two purposes. A new power-hungry robot is being built that will incorporates these stereo videocams along with a 500Mhz Pentium III running Linux. In addition, we are experimenting with data collection and control in an intelligent room.

Evaluating a Baby

The design of our robot baby is still fairly flexible, and we have been running a series of experiments with the initial prototype and test beds over the last two years to ensure that the hardware is sufficient to provide the basic sensory-motor capabilities required for language and ontology learning. This part of the project overlaps with a commercial home automation product prototyped by our group in which devices (lights, television, computer, alarm, watering system) can be controlled using natural language input, either spoken or typed [I2Net Computer Systems, Clipsal].

The doll should be able to identify where a speaker is, using audio and/or visual cues. This is relatively straightforward: given we have four non-coplanar microphones we are able to identify location in 3 dimensions. Even with just the stereo microphones in the original prototype, we have demonstrated the doll turning its head left or right (but not up or down) to face a sound.

A further development of this idea is that we should be able to use noise-canceling, blind signal separation and beam-forming techniques to enhance the signal from a target speaker. Note that we are not limited to blind signal separation techniques, because we assume that we know the relative position of the microphones in the tetrahedral (head) array (typically 20 to 80 cm distant), as well as the location of the microphones around the room (which are typically 1 to 2 metres distant) [Li, Powers and Peach, 2001]. These experiments differ from typical experiments in speech recognition because we are assuming a noisy environment and relatively distant mikes.

Generally speaking, separation of artificially mixed and convolved audio signals is relatively easy, but separation and deconvolution of real signals is still problematic for all algorithms. Another issue is synchronization of recordings made by multiple computers (eventually this will reduce to synchronizing doll and room but currently up to five computers are involved). The synchronization using a clapper or buzzer to signal the start of a recording visually and aurally did not provide sufficient accuracy - the error is of the same order as the echoes we are trying to deal with but not as consistent across recordings. A more effective means of synchronization involved playing close-mike recordings rather than using live speakers, and recording a sine wave sync burst on one track of each computer. The playing of recorded samples rather than the use of live speakers also allows us to perform a more direct evaluation, in terms of signal-to-noise ratio, of the degree of separation achieved by our algorithms.

Our conclusion is that at this point we are unable to significantly improve the SNR using multiple microphones in a noisy environment. Although we are now standardizing on the more expensive directional USB devices designed to record at distances of up to 60cm, we are employing them at distances well outside their specifications, at up to 2m. Nonetheless, speech sounds are comprehensible enough through all microphones, and the USB microphones are clear even at 2m, notwithstanding that commercial speech recognition performance degrades quickly as we move out of the specified range.

An alternative to the BSS approach is to investigate direct Speech Recognition using sensor fusion of multiple microphones and cameras, and taking into account positional information since human speech understanding involves visual cues as well as multidimensional auditory perception (not just stereo). Initially we are examining whether speech recognition under these unfavorable conditions can be enhanced by making use of visual cues.

This task in known as Speech Reading and is known to face difficulties due to catastrophic fusion - usually the results using multimodal input are worse than those achievable using one or other modes alone. However in our preliminary experiments we have demonstrated an increase in recall from 21% to 29% in a phoneme discrimination task comparing audio recognition to audiovisual recognition using simple auditory and visual features. This results from both extremely clear discrimination of the lips using a new 'red-exclusion' contrast technique [Lewis and Powers, 2000] and use of a late fusion technique designed to ensure that training gives more weight to the more significant features of the current input vector.

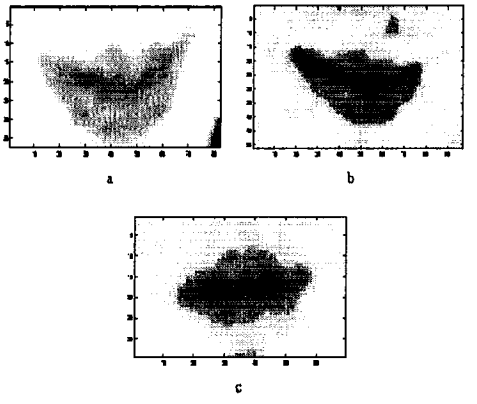

Figure 3. *Mouths as captured from 3 distinct subjects using red exclusion. Note: subject* a *is female, subjects* b *and* c *are male, and subject* b *has a moustache and beard.*

In summary, whilst the audio-visual information available from our robot baby in a normal office environment (containing several computers, several people and the usual collection of furniture and furnishings) is sufficient for people to understand the speech and make use of visual cues, it is well beyond what can be handled by off-the-shelf speech software, standard and custom signal separation and deconvolution algorithms do not handle it well, but the visual cues are promising and have increased phoneme recognition by close to 40% in a study on the recognition of stops and nasals [Lewis, 2000].

Further investigations will concentrate on the sensor fusion of multiple microphones and visual cues for direct phone recognition, rather than speech recognition. The performance of modern speech recognition systems is also overestimated (or ours underestimated) to the extent that they perform word rather than phoneme recognition and depend on higher order models (statistical, grammatical and semantic models) to obtain the published performance - they tend to make up their own stories and produce complete, albeit grammatical, rubbish under even mildly adverse conditions.

Commercial Applications

An offshoot of this project is the development of a speech control product for I2Net Computer Solutions operating with Clipsal home automation products. Whilst excellent performance can be achieved using text input or a headset microphone, the investigations into improving performance using signal separation and speech reading techniques are being undertaken with the support of I2Net with a view to enabling rooms to be wired for hands/headset-free operation

References

Block, H. D., J. Moulton, and G. M. Robinson (1975). Natural Language Acquisition by a Robot. *International Journal of Man-Machine Studies* 7: 571-608.

Clark, A. C. (1972). *The Lost Worlds of 2001*, Sidgwick and Jackson.

Deane, P. (1992). *Grammar in mind and brain: explorations in cognitive syntax.* Mouton

Entwisle, J. and Groves, M. (1994). A method of parsing English based on sentence form. New Methods in Language Processing (NeMLaP-1): 116-122.

Feldman, J. A., Lakoff, G., Stolcke, A., Hollback Weber, S. (1990). Miniature Language Acquisition: A Touchstone for Cognitive Science. TR-90-009. International Computer Science Institute.

Hogan, J. M., J. Diederich and G. D. Finn (1998). Selective Attention and the Acquisition of Spatial Semantics. In D.M.W.Powers (ed), New Methods in Language Processing and Computational Natural Language Learning (NeMLaP-3/CoNLL-98) 235-244, ACL

Homes, D. (1998). Perceptually grounded language learning. Computer Science Honours Thesis, Flinders University, AUS

Hume, D. (1984). Creating interactive worlds with multiple actors. Computer Science Honours Thesis, University of NSW, AUS.

Ishiguru, N. (1962). *Astro Boy.* Nippon Television Cartoon Serial.

Kozima, H and A. Ito, (1998). Towards language acquisition by an attention-sharing robot. In D.M.W.Powers (ed), New Methods in Language Processing and Computational Natural Language Learning (NeMLaP-3/CoNLL-98) 235-244, ACL

Lewis, T. W. and D. M. W. Powers (2000). Lip Feature Extraction using Red Exclusion, Visual Information Processing (VIP2000), University of NSW, Sydney

Lewis, T. W. (2000). Audio-Visual Speech Recognition: Extraction, Recognition and Integration. Computer Science Honours Thesis, Flinders University, SA, AUS.

Li, Y., D. M. W. Powers and J. Peach (2001). Comparison of Blind Source Separation Algorithms. WSES 2001 Neural Networks and Applications (NNA-01), World Scientific Engineering Society, Tenerife.

McCarthy, J., L. D. Earnest, D. R. Reddy and P. J. Vicens (1968). A computer with hands, eyes and ears. AFIPS Conf. Proc. Fall JCC 33#1:329-338.

Powers, D. M. W. and C. C. R. Turk (1989). *Machine Learning of Natural Language.* Springer-Verlag.

Powers, D. M. W. (1991). How far can self-organization go? Results in unsupervised language learning. In D.M.W Powers and L. Reeker (eds), AAAI Spring Symposium on Machine Learning of Natural Language and Ontology: 131-137. Kaiserslautern: DFKI D-91-09

Powers, D. M. W. (1992). On the significance of closed classes and boundary conditions: experiments in Machine Learning of Natural Language. SHOE Workshop on Extraction of Hierarchical Structure: 245-266. Tilburg NL: ITK Proceedings 92/1.

Steels, L. and R. Brooks (eds) (1995). *Building Situated Embodied Agents: the Alife route to AI.*

Steels, L. (1996). A self-organizing spatial vocabulary. *Artificial Life Journal* 3(2).

Steels, L. (1997). *Constructing and Sharing Perceptual Distinctions.* European Conference on Machine Learning.

Winograd, T. (1973). *Understanding Natural Language.* Academic Press.

Yngve, V. H. (1996). *From grammar to science: new foundations for general linguistics.* John Benjamins