Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

Who's Talking? – Efference Copy and a Robot's Sense of Agency

Justin Brody[1], Don Perlis[2], Jared Shamwell[2]
[1]Goucher College [2]University of Maryland
Justin.Brody@goucher.edu
perlis@cs.umd.edu
jared.shamwell@gmail.com

Abstract

How can a robot tell when *it* – rather than another agent – is making an utterance or performing an action? This is rather tricky and also very important for human-robot (or even robot-robot) interaction. Here we outline our beginning attempt to deal with this issue.

Introduction and Approach

In robot-human interaction, it is essential that the robot distinguish its current actions from those of other agents, including any humans it interacts with. Easy solutions involve things like name-tags on effectors, but these are far too readily thwarted: a tag can be misplaced, or misread, or copied. And humans of course do not tell their activity from that of other humans by looking at physical features; each of us "simply" knows that "this" is my action and "that" is someone else's. Imagine that a robot hears the utterance "Can you help me?" It will be crucial to its proper understanding and subsequent behavior, whether it takes this to be an assertion made *to* it, or *by* it.

Here we are referring not to photographs or soundtracks; we certainly can mistake images or other "markings" or behaviors, as to whether they are ours or another's. This is in effect again a sort of name-tag issue. Instead, we refer to knowing that when we take deliberate action A, it is we ourselves who are doing it then and there. This is an aspect of what is called *de se* knowledge (Perry 1979, Lewis 1979).

Others have approached robotic self-recognition as well – for instance Bringsjord (2015) works in the audio domain as do we, yet seems to rely on an equivalent of the name-tag method; and Hart & Scassellati (2011) use an approach similar to ours here, with some notable differences. Specifically, our current work emphasizes the use of an "efference copy" of initiated action rather than motor-proprioceptive information. Indeed, we ultimately intend to apply this model to all cognition, including "thought perception" (Bhargava et al 2012; Shamwell et al 2012).

So, if not by external indications, how can one tell that an action is being done by oneself? We humans just seem to feel ourselves doing it; but are we at a stage in AI where we can capture actual feelings? No, probably not. But we can at least partially unpack what goes on in ourselves here, and apply that to robots

Here is our tack: when we want to do something, we initiate an action A. Not only that – we *know* we have initiated A; it becomes part of our working memory. This is closely related to the aforementioned efference copy: when a motor command is sent from the cortex (to muscles) a "copy" of the command is retained in the cortex where it can be used for comparison with perceptual data about the ensuing action (or lack thereof). That in turn allows the organism to determine whether things are proceeding normally or not, and at times even what corrections might be needed.

And that is the approach we take here: when our robot initiates A, that fact will enter its KB, and its perceptual apparatus will monitor what happens, for comparison with expected results from the success of A. Not only that: the monitoring and the ensuing performance of A will be carried out in parallel over tiny time-steps so that – in the ideal case – there will be a strong cause-like covariance between the two. Thus as we lift our arm, we start the lift and see the start, we continue the lift and see that too, we finish the lift and see our arm high and coming to a stop. This close linkage between a sequence of action-initiation-bits and perception-bits tells me that yes, it is our action that we are doing. (We also *feel* the lift start – this is proprioceptive feedback; see Hart & Scassellati (2011). But for reasons related to our larger research aims, we are fo-

cused primarily on the efference-copy approach, which also seems to bear more immediately on auditory decision processes.)

Our approach can also be viewed through the lens of Gallagher's notions of a "sense of ownership" and "sense of agency" (Gallagher 2000). Our agent is training to recognize its own voice and the control it has over it. These can be considered crucial elements in the development of self-awareness (which is our broader program; see Brody et al (2012) and reasoning about other agents.

Of course one can tell a story where in fact we are not doing the lifting at all – our arm is numb and a hidden string is lifting the arm just as we are sending the (ineffective) lifting signals. So we can be fooled (at least about physical actions); and our robots need not be better than we at this.

Example and Some Technical Details

Here is an example, that arose in our own work: One of our Baxter robots (named Alice) had been programmed to look for – and then point to – another Baxter (Julia) while saying "I see Julia and I am pointing at her", whenever hearing an utterance containing the word "Julia". But we found, most oddly, that in some cases Alice performed as expected, and then a few seconds later spoke and pointed again, and then again, on and on. Finally we realized that Alice was hearing her own utterance which contained the key word, "Julia" which triggered her repeat actions.

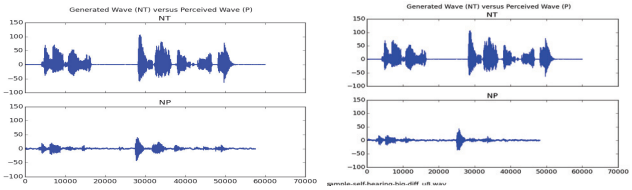
So our solution is this: Alice should be made aware of her own action-initiation efforts via efference copy, including speech acts, so that she can tell when she is hearing an ongoing uttereance that matches her own current effort; from this she can infer that it is not a command from us. But if she hears "Julia" and is *not* uttering it herself then she fairly infers that another agent is doing so, and she then should respond accordingly. Note that is not enough for Alice merely to remember making some particular utterance; for we might make the same utterance later on, and Alice should not regard it as hers but rather as another instance of it uttered by us, which she then should consider for possible response. That is, she should know – in the language of her KB – whether it is I (herself) or Another (e.g., us) who is performing an action right now.

Alice's speech is generated by the Festival Speech Synthesis System (http://www.cstr.ed.ac.uk/projects/festival/). The issue we explored was this: If an utterance is produced on the speaker from an "efference" file WAV-out (such as might in humans go to speech-producing muscles), and a microphone simultaneously picks up that sound signal which then is used to create a file WAV-in, then WAV-out and WAV-in should be close enough that a robot can track the closeness of match in real time and thus infer that it is hearing its own output. But because of noise, the match will be far from perfect. Thus the question is whether a suitable comparison method can separate "good-enough" matches from poor ones.

For this purpose, we had Festival generate WAV-out files for four sentences, and then created four WAV-in files from the four resulting microphone signals. A simple sound-frame-based comparison was done, employing the following utterances:

1. Hello, human. I'm very pleased to meet you.
2. Hello, human. I'm very cheesed to meet you.
3. Hello, Hugh. I meet you.
4. This is entirely different.

The idea is that Alice utters 1 (WAV-out_1), and may *hear* noisy versions of any of 1-4 (WAV_in_1, i=1…4). We then compared WAV-out_1 (our efference copy) with each of WAV-in_i. As it turned out, neither WAV-in_1 nor WAV-in_2 was a truly good match for WAV-out_1. But this is not surprising given the noise inherent in any recording method. We did find that despite this poor match, it was still significantly better than the matches of WAV-in_3 and WAV-in_4 to WAV-out_1. Thus if the robot utters 1 above, while hearing 3 or 4, it can tell that this is not what it is producing, whereas hearing 1 or 2 can pass muster, using the sound-comparison formula we worked with; in fact, 2 scored higher than 1; and 3 and 4 were almost identical in (low) score. Research shows we often do not utter what we think we are saying, nor hear what others are uttering: we infer expected sounds a great deal (McGurk 1976. Niziolec 2013, Pinker 1994). Yet as long as the pattern of stresses has a close temporal similarity in output and input, it is not reasonable to suppose that one is hearing one's own immediate (albeit distorted) speech.



The figure above shows WAV-out (top) compared to WAV-in_1 (left) and to WAV-in_3 (right). Even though the mic has in theory picked up the same signal as was sent to the speaker (on the left) distortions made significant changes; but mathematical massaging was still able to classify WAV-in_1 (and also case 2) as close to WAV-out, compared to cases 3 (right) and 4, which were significantly further away.

Our long-range plans include massive use of efference copy, not only in regard to speech but also other physical

actions and even to the robotic agent's own internal "thinking" actions (Bhargava et al 2012).

Acknowledgments

We thank Carol Espy-Wilson for much-needed advice on processing sound and speech, and ONR for grant support.

References

Bhargava, P., Cox, M., Oates, T., Uran, O., Paisner, M., Perlis, D., and Shamwell. J. 2012. The robot baby and massive metacognition: Future vision. Proceedings of IEEE International Conference on Development, Learning, and Epigetic Robotics.

Bringsjord, S., et al. 2015. Real robots that pass human tests of self-consciousness. In Proceedings of IEEE the 24th International Symposium on Robots and Human Interactive Communications.

Brody, J,. Cox, M., and Perlis, D. 2013. The processual self as cognitive unifier. Proceedings of IACAP 2013.

Gallagher, S. 2000. Philosophical conceptions of the self: implications for cognitive science. Trends in Cognitive Science (4): 14-21

Hart, J., and Scassellati, B. 2011. Robotic models of self. In: Cox and Raja, eds, *Metareasoning: Thinking About Thinking*. Cambridge, MA, MIT Press.

Lewis, D. 1979. Attitudes *De Dicto* and *De Se*. The Philosophical Review: 88(4): 513-543

McGurk, H, and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*: 264(5588): 746-8.

Niziolec, C., Nagarajan, S., and Houde, J. 2013. What does motor efference copy represent? Evidence from speech production. *The Journal of Neuroscience* 33(41): 16110-16116.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13(1): 3-21.

Pinker, S. 1994. *The Language Instinct.* William Morrow.

Shamwell, J., Oates, T., Bhargava, P., Cox, M., Oh, U., Paisner, M., and Perlis, D. 2012. The robot baby and massive metacognition: Early steps via growing neural gas. Proceedings of IEEE International Conference on Development, Learning, and Epigetic Robotics.