Achieving Fluency through Perceptual-Symbol Practice in Human-Robot Collaboration

Guy Hoffman
MIT Media Laboratory
20 Ames Street E15-468
Cambridge, MA 02139
guy@media.mit.edu

Cynthia Breazeal
MIT Media Laboratory
20 Ames Street E15-468
Cambridge, MA 02139
cynthiab@media.mit.edu

ABSTRACT

We have developed a cognitive architecture for robotic teammates based on the neuro-psychological principles of perceptual symbols and simulation, with the aim of attaining increased fluency in human-robot teams. An instantiation of this architecture was implemented on a robotic desk lamp, performing in a human-robot collaborative task. This paper describes initial results from a human-subject study measuring team efficiency and team fluency, in which the robot works on a joint task with untrained subjects. We find significant differences in a number of efficiency and fluency metrics, when comparing our architecture to a purely reactive robot with similar capabilities.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Algorithms, Experimentation, Human Factors, Performance

Keywords

Anticipation, Cognitive Architectures, Fluency, Human Subject Studies, Human-Robot Interaction, Perceptual Symbols, Robotics, Teamwork

1. INTRODUCTION

We define *fluency* in joint action as the ethereal yet manifest quality existent when two agents perform together at high level of coordination and adaptation, in particular when they are well-accustomed to the task and to each other. This quality is often observed in human behavior, but virtually absent in human-robot interaction.

In the past, we have addressed the issue of human-robot fluency through the development of an anticipatory action system. Anticipation has been shown to play a role in perception [20, 21], and we have found that anticipation leads to improved best-case task efficiency and to a significant difference in the perceived commitment of the robot to the team and its contribution to the team's fluency and success [10].

We have since developed a perceptual symbol architecture based on principles of embodied cognition and simulation, ideas which are gaining ground in the neuroscientific literature([1, 16, 14, 19]; for a review, see: [9]).

In this paper we briefly describe the main components of this cognitive framework and its implementation on a human-robot collaborative task. We then discuss a human subject study conducted to evaluate the performance of the implemented system, and the effects it has on the efficiency and fluency of the task. We were particularly interested in how the system performs within the context of practice, in which the human and the robot repeat a fixed set of identical actions.

1.1 Related Work

Human-robot collaboration has been investigated in a number of previous works, although — to the best of our knowledge — the question of fluent action meshing or the improvement thereof through repetition has not received specific attention. Kimura *et al.* have studied a robotic arm assisting a human in an assembly task [13]. Their work addressed issues of vision and task representation, but is using a purely reactive approach and does not address practice. Other human-robot collaboration work, such as that of Fong *et al.* [5] or Jones and Rock [11] studies human-robot collaboration with an emphasis on dialog and control, aimed primarily at the teleoperation scenario.

Most work in shared-location human-robot collaboration has been concerned with the mechanical coordination and safety considerations of robots in shared tasks with humans [22, 12] . In contrast, we have previously presented work in shared-location human-robot teamwork addressing turn-taking and joint plans, but not anticipatory action or fluency [8]. Anticipatory action, without relation to a human collaborator has been investigated in navigation work, e.g. [4].

The idea of top-down processing has been utilized in computational systems in the past. Bregler demonstrated a convincing application of this concept to action recognition in computer vision [3]. Wren and Pentland created a robust human dynamic recognition and classification system by feeding likelihood data from high-level HMM procedures to pixel-level classifiers [24, 23]. Similarly, Hamdam *et al.* classified gesture sequences using Continuous Density Hidden Markov Models [6]. Ude *et al.* discuss similar top-down processing ideas for visual attention on a humanoid robot

HRI'08, March 12–15, 2008, Amsterdam, The Netherlands.

[17]. None of these works, however, model the top-down influences as *perceptual simulation* using the same pathways used for bottom-up processing, as is supported by the neuropsychological literature, and proposed in this paper.

Our own previous work in anticipatory action to support human-robot fluency was implemented on a simulated robot, using a discretized model framed as a stepwise MDP with simulated perception, and no perceptual simulation [10]. This paper significantly extends this work as it proposes a novel cognitive architecture based on the simulation and emulation of perceptual symbols to govern anticipation. Furthermore, it is implemented on a physical robot using noisy, continuous sensory input, acting in a situated interaction with a moving human.

2. COGNITIVE ARCHITECTURE

This section provides a brief outline of the principles guiding the perception-based cognitive architecture employed in the studies described below. It is by no means complete, as this paper is mainly concerned with experimental results obtained in the human subject study. A full description of the cognitive architecture is given separately [7].

In this work, we subscribe to the notion of perceptual simulation and top-down biasing of perceptual processes. We thus model concepts leading to actions as perceptual processes and biases, residing within the perceptual system rather than in an amodal, symbolic, and centralized semantic network.

Perceptions are processed in *modality streams* built of processing nodes. These nodes can correspond to a feature (such as a color) or property (such as the speed) of perceptual data, or to a higher-level concept describing a statistical analysis of features of properties, in the spirit of [15]. These modality streams are connected to an action network consisting of action nodes, which are activated in a similar manner as perceptual processing nodes. When — through practice or otherwise — higher-level concepts or motor actions are activated, top-down bias effects activation of lower-level process nodes enabling the robot to act sooner and more rapidly, thus emulating the idea of practice.

Two main subsystems are used in the experiments:

Anticipatory simulation — Based on recurrent perceptual information, the robot builds a statistical map of likely occurrences, similar to the anticipatory system described in [10]. This in turn triggers top-down simulation biasing perceptual processing and thus altering reaction times and action behavior on the robot's part.

Intermodal connection reinforcing — This subsystem reinforces perceptual connections that co-activate frequently and consistently, and decreases the weight of connections that are infrequent or inconsistent, thus creating embellished connections between perceptual processing nodes which activate simultaneously.

The principle of operation of each of these subsystem will be elucidated through their instantiation, as described in Section 5.

3. ROBOTIC PLATFORM

The robot employed in this evaluation was AUR, a robotic desk lamp. The lamp has a 5-degree-of-freedom arm, based on the design of a previous robot, RoCo [2]. It features a variable aperture that can change the light beam's width, and a LED lamp which can illuminate in a range of the red-green-blue color space. AUR is stationary and mounted on top of a steel and wood workbench locating its base at approximately 36 inches above the floor. Its processing is done on a 2x Dual 2.66GHz Intel processor machine located underneath the workbench.

4. TASK DESCRIPTION

In the human-robot collaboration used in our studies the human operated in a workspace as depicted in Figure 1 (a) and (b). The robot, from its tabletop vantage point, could direct its head to different locations around its own axis, and change the color of the light beam. When asked to "Go", "Come", or "Come here" the robot would move to the location of the person's hand, assuming the hand was relatively static. Additionally, the color changed in response to speech commands to one of three colors: "Blue", "Red", and "Green"'.

To complete the task, the person was asked to stand in front of the lamp. The workspace contained three locations (A,B,C). At each location there was a black music stand with a white cardboard square labeled with the location letter, and including four doors. Each door, when lifted, revealed the name of a color written underneath (Red, Blue, or Green).

The task was to complete a sequence of 8 actions, which was described in diagrammatical form on a sequence sheet as shown in Figure 1(c). This sequence was to be repeated 10 times, as quickly as possible.

Each action in the sequence specifies: a general location A, B, or C, and an indication of which of the four doors to open. The action is completed when the lamp shines the specified color of light at that location. This would result in the sound of a buzzer, indicating the person should move to the next action in the sequence. A different buzzer was sounded when a whole sequence was completed.

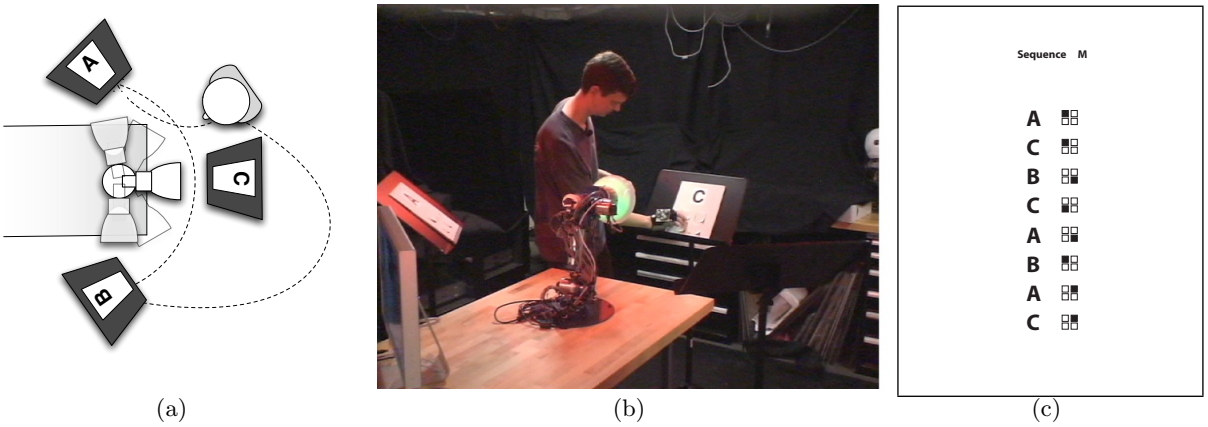
4.1 Sensing

The robot's sensory input device was a Vicon motion capture system which was used to identify and track the location and orientation of the human's right hand and head at a frequency of 10 times per second.

The system also takes input from Sphinx-4, an open-source, Java-based speech recognition system created by the Sphinx group at Carnegie Mellon University, in collaboration with Sun Microsystems Laboratories, Mitsubishi Electric Research Labs, and Hewlett Packard [18]. The commands recognized by the system in this task were: "Come", "Come Here", "Go", "Red", "Blue", "Green", and "Off".

5. COGNITIVE NETWORK

To solve the task described in this paper, we designed a cognitive network along the principles introduced in Section 2, made up of three modality streams: a visual, an auditory, and a proprioceptive. The action network includes five actions, three color changing nodes, a color "off" node, and a "goto position" node.



(a) (b) (c)

Figure 1: (a) Diagram, and (b) photograph of the collaborative lighting task workspace. (c) Sample experimental sequence.

5.1 Visual Modality

The visual modality stream's sensory input stems from the Vicon motion capture system, indicating the hand position in the workspace.

Two feature areas are downstream from that Vicon sensory node: in one, four workspace segmentation feature nodes activate proportionally to the proximity of the hand to each of the four corners of the workspace. In the other, a speed node detects the speed as a single-frame position derivative of the hand position. Downstream from the speed node, an inhibitory connection feeds the "Stillness" feature node, which simply detects the inverse of the speed node clamped between 0 and 1. The stillness node feeds into a property node indicating whether the hand has been still for some consecutive period of time. Three concept nodes represent the location of the hand near one of the task targets, A, B, or C.

The target concept nodes are connected to the "goto position" action node, as is a conjunction node combining the "stillness settled" and the "Go" speech feature node. A combination of a hand position near the target, an aggregate stillness of the hand, and a "Go" command will trigger this action, leading the robot to move its beam towards the appropriate position.

5.1.1 Simulation in the Visual Modality

Over time, as the robot correlates the location of the hand near a target, both the afferent and the efferent connections between the concept nodes and the workspace segmentation feature nodes, get reinforced. Then, in anticipation of the next game step as determined by the simulator, the appropriate concept node, as well as the "Go" action node get a simulated activation, as determined by the action and concept simulation coefficients and the anticipatory probability. At this point, the reinforced efferent connections to the feature nodes cause a bias towards the perception of the hand in the appropriate workspace regions.

During such simulation, the workspace segmentation feature nodes are partially activated, and as a result can reach full activation with sensory data that is further away from each of the target positions. This results in a pre-step bias of this perceptual stream leading to anticipatory action on the robot's part.

5.2 Auditory Modality

The auditory modality stream uses input from the Sphinx speech recognition system, parsing the auditory input into speech tokens. This sensory node feeds five speech feature nodes, which respond to the detection of a particular speech token in the auditory stream.

The four other speech feature nodes have afferent connections to the four light change action nodes. Note also that the "Go" action node causes the light to turn off to prevent light to shine while the lamp is in motion.

5.3 Proprioceptive Modality

Finally, the proprioceptive modality stream is based on the joint positions of the robot, thus representing the robot's sense of physical self.

This simple modality stream has a direction feature node, which calculates the lamp head direction from the joint positions. This node, in turn, feeds into left, right, and center property nodes, which classify the overall orientation of the robot.

5.3.1 Inter-modal connectivity

The orientation property nodes in this modality are connected to the speech feature nodes in the auditory modality, and managed with the connection reinforcement mechanism described above. The aim of these connections are to enable the robot to learn a perceptual correspondence between its orientation and a speech detection perceptual stimulus. Thus, while there is no direct connection between the proprioceptive modality and the action network, this modality indirectly affects the color actions through their connectivity to the auditory modality.

6. EMULATOR

To model practice in this network, we designed a Bayesian sequence learner along the same lines as described in [10]. Our Markov model used a 3-step sequence history.

At each step, after a few rounds of practice, the sequence learner estimates the probability of the appropriate hand-target concept, and "Go" action being expected. This triggers simulation of both the appropriate concept and the action node. The result of this emulation is that the workspace segmentation feature nodes simulate to the extent that they

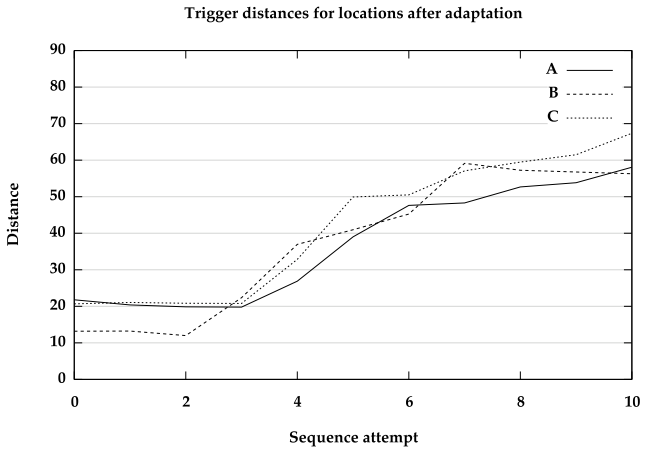


Figure 2: Effect of emulator on trigger distances for each of the three locations using a simple A-C-B sequence.

are correlated with the appropriate hand-target concept. Thus an increasing distance between the hand position and the correct target is adequate to trigger the appropriate response.

Figure 2 shows the the change in trigger area for each of the areas using a simple A-C-B sequence with ten iterations. The longer the practice session, the more 'primed' the perceptual stream is, and the further away the hand position can be to result in a concept and subsequent action activation.

7. INTER-MODAL SIMULATION

In addition to the visual emulation, we used inter-modal simulation between the robot's proprioceptive property nodes ("Left", "Right", and "Center") and the speech feature nodes (see: Section 5.3.1). This corresponds to a repeating stimulation of a certain word in the auditory stream when the robot has a certain position, resulting in the perceptual simulation of that speech segment every time the robot returns to that position.

If there is a consistent correlation between position and color, the robot will increasingly trigger the appropriate color without an explicit human command.

The effects of connection reinforcement on inter-modal simulation can be seen in Figure 3. The first graph shows the weights between the `Center` Property Node and each of the color speech feature nodes with a mostly consistent mapping between the two modality instances (there is some cross-occurrence towards the end with the `Red Speech` Feature node).

The second graph shows the effects of mostly alternating contingency between the proprioceptive node and the speech feature nodes. Across the experimental sequence, these reinforcements cancel each other out and result in a low simulation factor between these nodes.

8. EXPERIMENTAL DESIGN

We conducted a between-group controlled experiment with two conditions. The control (or `REACTIVE`) condition corresponds to the baseline condition in which both the Emulation/Simulation framework and the Intermodal Reinforcement were disabled. The remainder of the system, i.e. the perceptual network and all activation streams and thresholds, were identically retained. In the second (`FLUENCY`) condition, both the Emulation/Simulation framework and the Intermodal Reinforcement were active with fixed parameters.

To control for instruction bias, neither group was told whether the robot will adapt to their behavior. All participants were allowed to practice with the system before beginning the experiment. We recruited subjects from the campus and area communities, through email solicitation.

The experiment included two sequences, or "patterns". Pattern A was associated with a setup in which there were different colors under the different doors. In Pattern B there was a one-to-one mapping between location and color, i.e., the same color was hidden under all four doors in a single location. For example, all doors in location B hid the color "Blue", and all doors in location A hid the color "Green". Therefore both the human's memory and the robot's intermodal reinforcement process could more easily learn the correct association between spatial location and color.

This experimental protocol was reviewed and approved by the institutional review board of the Massachusetts Institute of Technology.

8.1 Subjects

We recruited 38 subject, who were arbitrarily designated to one of the two experimental conditions, and for each subject, they were arbitrarily assigned the order of sequences between Pattern A first, and Pattern B first. At the last day of the experiment we experienced an unrecoverable hardware failure, forcing us to release the last 5 subjects. We thus remained with 33 subjects (17 male), 15 in the `REACTIVE` condition, and 18 in the `FLUENCY` condition.

Two additional subjects in the `FLUENCY` condition experienced a mechanical failure. This was resolved in a short period of time, and the subjects continued the experiment. The failure affected the amount of data we were able to obtain from these subjects, a fact we addressed as described below.

Subjects were allowed to practice but usually elected not to practice for very long. As a result the first and second sequence attempt of most subjects was significantly slower than other rounds, leading us, for some metrics — for example when considering relative improvement over time — to only consider the second round each subject performed, assuming that at that point the subjects were familiar enough with the system that we measured only trial-related metrics.

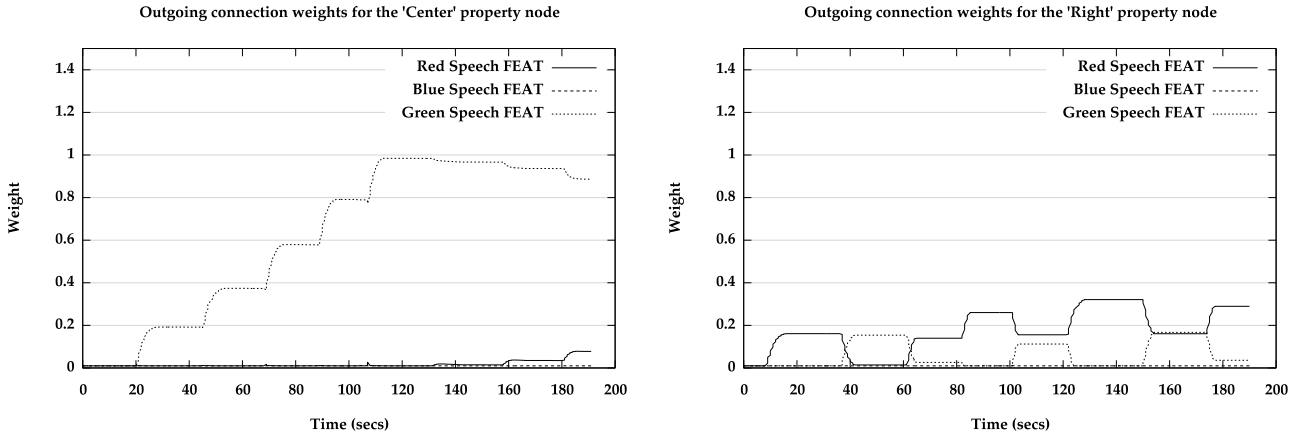
9. TERMINOLOGY

In the coming sections we will use the following terminology:

Turn is the time and actions occurring between two consecutive turn buzzers. These include a single event of correctly shining the right light onto the right board.

Sequence is a set of eight turns. There are ten **attempts** at a sequence.

Round is a set of ten attempts at a sequence. The *first round* is the round performed first; similarly for the phrase *second round*. Rounds can also be identifies by



(a) (b)

Figure 3: Connection reinforcement between the proprioceptive and auditory modality with (a) consistent contingency and (b) inconsistent contingency.

patterns. In this case we will refer to *Pattern A* and *Pattern B*. Note that for different subjects the order of patterns, i.e. the mapping between patterns and round numbers, is different.

Task is a set of two rounds, using both patterns.

10. RESULTS

The behavioral measures recorded in this experiment were elicited from the log files generated by the experiment software. The software running the robot logged a number of events, including the human's speech commands, the robot's action selection, the human's hand position, and the experimenter's buzzer times.

To account for a number of mechanical failures as mentioned above, as well as for mistakenly recorded turn and sequence buzzer events, the data has been automatically cleaned up, by eliminating the following sequences: (a) any sequence attempt that does not contain 7–9 turns was eliminated; (b) any sequence attempt that lasted for less than 25 seconds or more than 180 seconds was eliminated.

As a result, two subjects' data included only 8 sequences in one of the rounds, one subject's data remained with 9 sequences in both rounds, and two subjects' data remained with 9 sequences in one of their two rounds.

We have included the valid data from these subjects in our analyses, except in Hypothesis **H1** below. In **H1**, as well as in the graphs depicting sequence-by-sequence progress on the recorded behavioral measures, we included only data from trials containing 10 valid sequences.

10.1 Team Performance

Our first set of hypotheses were concerned with the performance of the human-robot team. We hypothesized the following metrics to be significantly lower in the `FLUENCY` condition compared to the `REACTIVE` condition:

H1 — The overall task completion time (both rounds)

H2 — Mean sequence attempt time

H3 — Mean sequence attempt time (second half)

H4 — Best sequence attempt time

H5 — Improvement (ratio between last and first attempt)[1]

Using a T-test with independent samples, we find a significant difference between the two conditions in all five hypotheses. All values are in seconds, except in H5, which is a fraction.

H1 — Total task time: `REACTIVE`: 1401.66 ± 162.90, `FLUENCY`: 1196.26 ± 226.83; t(24)=2.609, $p < 0.05$.

H2 — Mean sequence time: `REACTIVE`: 141.38 ± 17.38, `FLUENCY`: 116.21 ± 22.67; t(30)=3.487, $p < 0.01$.

H3 — Mean sequence time (2nd half): `REACTIVE`: 131.04 ± 17.76, `FLUENCY`: 97.93 ± 22.75; t(30)=4.544, $p < 0.001$.

H4 — Best sequence time: `REACTIVE`: 117.93 ± 16.11, `FLUENCY`: 83.87 ± 18.81; t(30)=5.461, $p < 0.001$.

H5 — Sequence time improvement: `REACTIVE`: 0.76 ± 0.18, `FLUENCY`: 0.50 ± 0.15; t(30)=4.718, $p < 0.001$.

All of our hypotheses were confirmed, demonstrating a significant improvement in team performance under the `FLUENCY` condition. Note that examining in particular the second half of each task round leads to an increase in difference between the two conditions, and an increase in significance. This phenomenon is further illustrated in the figures below.

Figure 4 shows the average sequence attempt time for both conditions, split by round. The notion of initial practice runs is evident in this figure, as the second round starts at a lower time than the first round. In both cases, the `FLUENCY` condition converges at slightly over 40 seconds, while the `REACTIVE` condition maintains an average over 60 seconds.

[1]We use only data from the second round under the assumption that, at this point, the subject is familiar with the task structure and robot, and we thus measure only the team's improvement and not the initial practice needed by the subject.

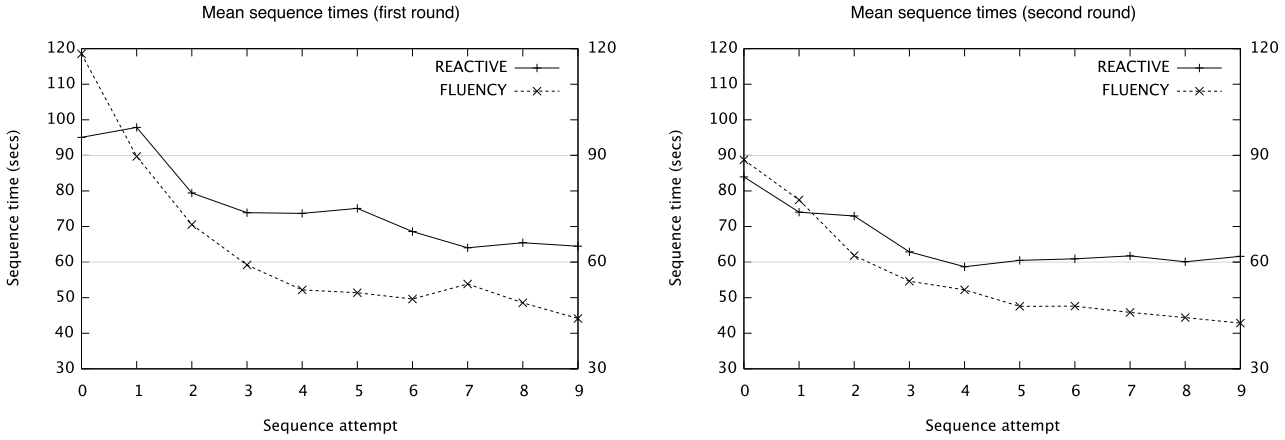


Figure 4: Mean sequence times — per round — over two ten-attempt practice sessions, comparing the `REACTIVE` and `FLUENCY` conditions.

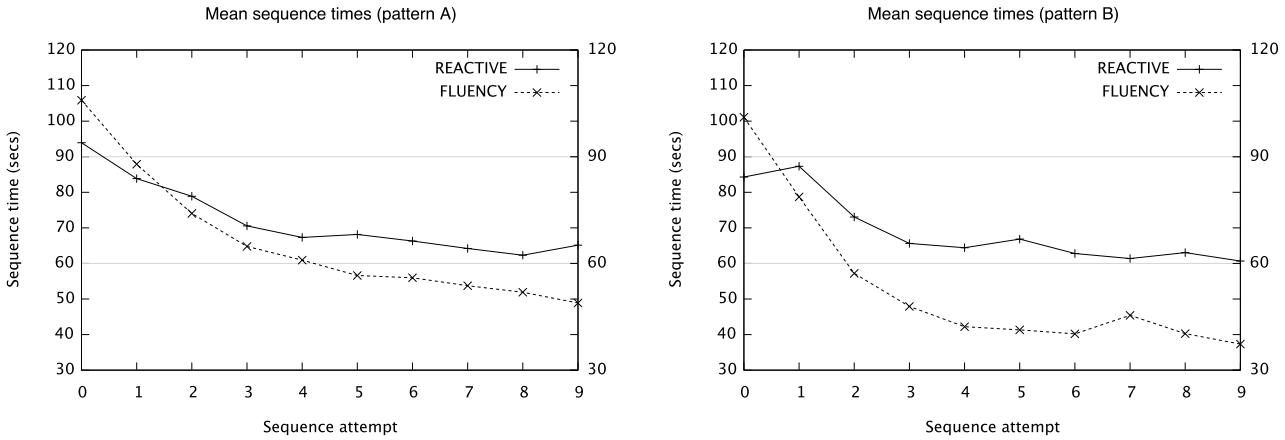


Figure 5: Mean sequence times — per pattern — over two ten-attempt practice sessions, comparing the `REACTIVE` and `FLUENCY` conditions.

Figure 5 shows data for sequence pattern A and B, respectively. Since the robot and the human "learn" the color sequences more fully in pattern B, we see a more dramatic improvement in the `FLUENCY` condition, converging on a below-40 second score in the final sequence attempt.

10.2 Fluency Metrics

The second set of hypotheses tested in this experiment relate to the fluency of the team. These are based on the fluency metrics set forth in [10]. The following metrics were elicited from the log files recorded by the experiment software:

`IDLE` — Human idle time. The percentage of time within each task round in which the human hand was stationary or move very little.

This metric is elicited from the human's hand position. A low-passed sensor with hysteresis is triggered every time the frame-by-frame distance of the hand position crosses a certain threshold.

`DELAY` — Robot functional delay. The time that passed from the beginning of a turn to the onset of the robot's movement.

We tested the hypotheses that the following metrics are significantly lower in the `FLUENCY` condition compared to the `REACTIVE` condition:

H6 — Mean human idle time

H7 — Mean robot functional delay

H8 — Mean robot functional delay (second half)

We found significant difference between the two conditions on these three hypotheses, as well. Using a T-test for independent samples, **H6** is a ratio, **H7** and **H8** are in seconds:

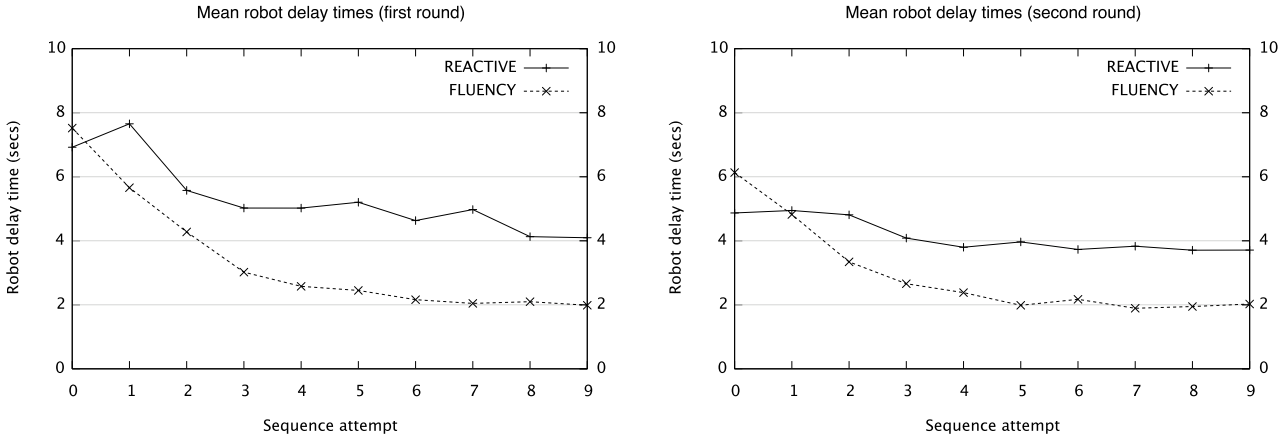


Figure 6: Mean functional delay times — per round — over two ten-attempt practice sessions, comparing the `REACTIVE` and `FLUENCY` conditions.

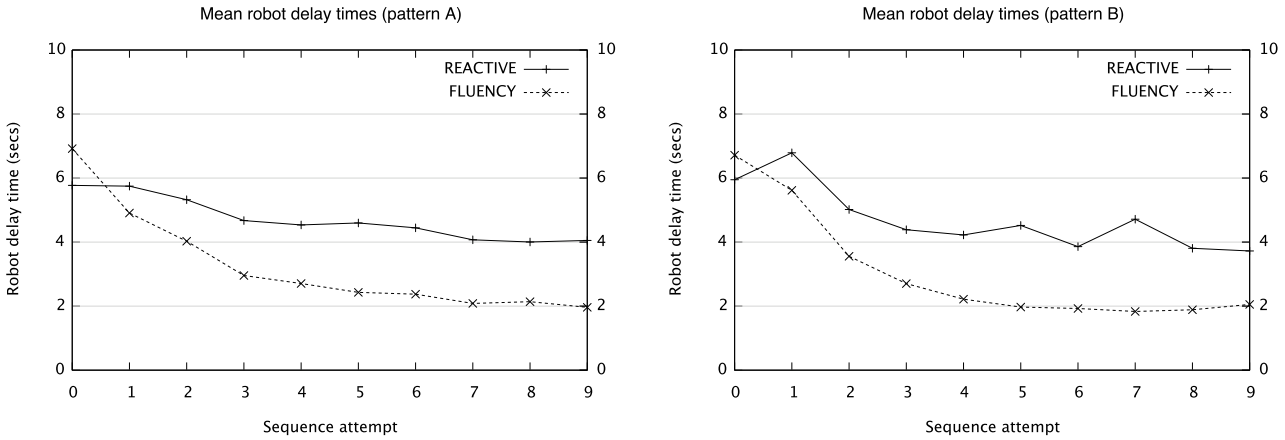


Figure 7: Mean functional delay times — per pattern — over two ten-attempt practice sessions, comparing the `REACTIVE` and `FLUENCY` conditions.

H6 — Human idle time: `REACTIVE`: 0.46 ± 0.08, `FLUENCY`: 0.364 ± 0.09; t(30)=3.001, $p < 0.01$.

H7 — Robot functional delay: `REACTIVE`: 4.81 ± 9.91, `FLUENCY`: 3.66 ± 15.72; t(30)=2.434, $p < 0.05$.

H8 — Robot functional delay (2nd half): `REACTIVE`: 4.07 ± 10.97, `FLUENCY`: 1.48 ± 6.14; t(30)=3.487, $p < 0.001$.

All of our hypotheses were confirmed, demonstrating a significant improvement in both fluency metrics under the `FLUENCY` condition. Again, examining the second half of each task round (hypothesis **H8**) shows an increase in difference between the two conditions, and an increase in significance.

Figures 6 and 7 shows the average change in robot functional delay time for both conditions, split by trial. It can be seen, in Figure 6 (b), that in the `REACTIVE` condition, the human teammate can do little to improve the robot's delay, after the initial "practice" period.

11. CONCLUSION

In this paper we introduce a novel cognitive architecture aimed at achieving fluency in human-robot joint action. This extends previous work which used anticipatory action selection for human-robot fluency. We describe a perceptual symbol system, which uses simulation and inter-modal reinforcement to allow for decreased reaction time through top-down biasing of perceptual processing.

We describe a human subject study evaluating the effects of this system, comparing it with a purely reactive system using the same bottom-up processing. We find significant differences in the task efficiency between the two conditions, as well as significant differences in a number of fluency metrics.

These results indicate that using a perception-based embodied cognitive architecture is a promising avenue to achieving efficiency and fluency in human-robot joint action.

12. REFERENCES

[1] L. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660, 1999.

[2] C. Breazeal, A. Wang, and R. Picard. Experiments with a robotic computer: body, affect and cognition interactions. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 153–160, New York, NY, USA, 2007. ACM Press.

[3] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, page 568. IEEE Computer Society, 1997.

[4] Y. Endo. Anticipatory and improvisational robot via recollection and exploitation of episodic memories. In *Proceedings of the AAAI Fall Symposium*, 2005.

[5] T. W. Fong, C. Thorpe, and C. Baur. Collaboration, dialogue, and human-robot interaction. In *Proceedings of the 10th International Symposium of Robotics Research, Lorne, Victoria, Australia*, London, November 2001. Springer-Verlag.

[6] R. Hamdan, F. Heitz, and L. Thoraval. Gesture localization and recognition using probabilistic visual learning. In *Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pages 2098–2103, Ft. Collins, CO, USA, June 1999.

[7] G. Hoffman. *Ensemble: Fluency and Embodiment for Robots Acting with Humans*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2007.

[8] G. Hoffman and C. Breazeal. Collaboration in human-robot teams. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference*, Chicago, IL, USA, September 2004. AIAA.

[9] G. Hoffman and C. Breazeal. Robotic partners' bodies and minds: An embodied approach to fluid human-robot collaboration. In *Fifth International Workshop on Cognitive Robotics, AAAI'06*, 2006.

[10] G. Hoffman and C. Breazeal. Cost-based anticipatory action-selection for human-robot fluency. *IEEE Transactions on Robotics and Automation (in press)*, 2007.

[11] H. Jones and S. Rock. Dialogue-based human-robot interaction for space construction teams. In *IEEE Aerospace Conference Proceedings*, volume 7, pages 3645–3653, 2002.

[12] O. Khatib, O. Brock, K. Chang, D. Ruspini, L. Sentis, and S. Viji. Human-centered robotics and interactive haptic simulation. *International Journal of Robotics Research*, 23(2):167–178, February 2004.

[13] H. Kimura, T. Horiuchi, and K. Ikeuchi. Task-model based human robot cooperation using vision. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS'99)*, pages 701–706, 1999.

[14] G. Lakoff and M. Johnson. *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. Basic Books, New York, 1999.

[15] K. Simmons and L. W. Barsalou. The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, 20:451–486, 2003.

[16] M. J. Spivey, D. C. Richardson, and M. Gonzalez-Marquez. On the perceptual-motor and image-schematic infrastructure of language. In D. Pecher and R. A. Zwaan, editors, *Grounding cognition: the role of perception and action in memory, language, and thinking*. Cambridge Univ. Press, Cambridge, UK, 2005.

[17] A. Ude, J. Moren, and G. Cheng. Visual attention and distributed processing of visual information for the control of humanoid robots. In M. Hackel, editor, *Humanoid Robots: Human-like Machines*, chapter 22, pages 423–436. Itech, 2007.

[18] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfe. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, November 2004.

[19] M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, December 2002.

[20] M. Wilson and G. Knoblich. The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131:460–473, 2005.

[21] S. Wilson, A. Saygin, M. Sereno, and M. Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7:701–702, 2004.

[22] H. Woern and T. Laengle. Cooperation between human beings and robot systems in an industrial environment. In *Proceedings of the Mechatronics and Robotics*, volume 1, pages 156–165, 2000.

[23] C. Wren, B. Clarkson, and A. Pentland. Understanding purposeful human motion. In *Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 378–383, 2000.

[24] C. R. Wren and A. Pentland. Dynaman: A recursive model of human motion. Technical Report VisMod 451, MIT Media Laboratory, 1997.