Anticipatory Perceptual Simulation for Human-Robot Joint Practice: Theory and Application Study

Guy Hoffman and **Cynthia Breazeal**
MIT Media Laboratory
20 Ames Street E15-468
Cambridge, MA 02142

Abstract

With the aim of fluency and efficiency in human-robot teams, we have developed a cognitive architecture based on the neuro-psychological principles of anticipation and perceptual simulation through top-down biasing. An instantiation of this architecture was implemented on a non-anthropomorphic robotic lamp, performing in a human-robot collaborative task.

In a human-subject study, in which the robot works on a joint task with untrained subjects, we find our approach to be significantly more efficient and fluent than in a comparable system without anticipatory perceptual simulation. We also show the robot and the human to be increasingly contributing at a similar rate. Through self-report, we find significant differences between the two conditions in the sense of team fluency, the team's improvement over time, and the robot's contribution to the efficiency and fluency. We also find difference in verbal attitudes towards the robot: most notably, subjects working with the anticipatory robot attribute more positive and more human qualities to the robot, but display increased self-blame and self-deprecation.

Introduction

Our goal is to design robots that can work fluently with a human partner in a physically situated setting. *Fluency* in joint action is the 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 observed in a variety of human behaviors, but is virtually absent in human-robot interaction.

Neurological and psychological evidence in humans indicates that anticipation and perceptual simulation plays a role in perception, in the perception of conspecifics, and in joint action (Wilson and Knoblich 2005; Sebanz, Bekkering, and Knoblich 2006). In simulated agents acting with humans, we have shown anticipation to lead to improved task efficiency and fluency, as well as a perceived commitment of a simulated robot to the team and its contribution to the team's fluency and success (Hoffman and Breazeal 2007).

Based on these findings, we believe that anticipation through perceptual simulation can provide a powerful model for robots acting jointly with humans if they are to collaborate fluently using multi-modal sensor data. To that end, we developed a cognitive architecture based on the principles of embodied cognition and top-down perceptual simulation, ideas which are gaining ground in the neuroscientific literature in recent years (Barsalou 1999; Wilson 2002).

In this paper we introduce some core concepts of our cognitive framework and its implementation on a non-anthropomorphic robot designed for human-robot collaboration. We discuss a controlled 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, as well as on the human subjects' perception of the robot and the team. We are particularly interested in how the system performs within the context of practice, in which the human and the robot repeat a set of identical actions.

Related Work

Human-robot collaboration has been investigated in a number of previous works, although the question of fluent action meshing or the improvement thereof through repetition has not received much attention. Kimura *et al.* have studied a robotic arm assisting a human in an assembly task (Kimura, Horiuchi, and Ikeuchi 1999). Their work addresses issues of vision and task representation, but does not address anticipation, fluency, or practice. Some work in shared-location human-robot collaboration has been concerned with the mechanical coordination and safety considerations of robots in shared tasks with humans (Woern and Laengle 2000; Khatib et al. 2004) . Other work addresses turn-taking and joint plans, but not anticipatory action or fluency (Hoffman and Breazeal 2004). Anticipatory action, without relation to a human collaborator has been investigated in navigation work, e.g. (Endo 2005).

The idea of top-down biasing has been utilized in computational systems in the past, e.g. in visual action recognition (Bregler 1997). 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 (Wren, Clarkson, and Pentland 2000). Ude *et al.* discuss similar top-down processing ideas for visual attention on a humanoid robot (Ude, Moren, and Cheng 2007). None of these works, however, model the top-down influences as *perceptual simulation* using the same pathways used for bottom-up processing, as supported by the neuro-psychological literature, and proposed in this paper.

Our own previous work in anticipatory action to support human-robot fluency was implemented on a simulated agent, using a discretized model framed as a stepwise MDP with simulated perception, and no perceptual simulation (Hoffman and Breazeal 2007). This paper significantly extends this work as it models anticipation through the simulation of perceptual symbols. Furthermore, it is implemented on a physical robot using noisy, continuous sensory input, acting in a situated interaction with a moving human.

Cognitive Architecture

We propose that fluency in joint action achieved through practice relies on two processes: *(a) anticipation* based on a model of repetitive past events, and *(b)* the modeling of the resulting anticipatory expectation as *perceptual simulation*, affecting a top-down bias of perceptual processes.

Modality Streams and Process Nodes

In this model, perceptions are processed in *modality streams* built of interconnected *process nodes*. These nodes can correspond to raw sensory input (such as a visual frame or a joint sensor), to a feature (such as the dominant color or orientation of a sensory data point), to a property (such as the speed of an object), or to a higher-level concept describing a statistical congruency of features of properties, in the spirit of the Convergence Zones in (Simmons and Barsalou 2003).

Modality streams are connected to an *action network* consisting of *action nodes*, which are activated in a similar manner as perceptual process nodes. An action node, in turn, leads to the performance of a motor action. Connections between nodes in a stream are not binary, but weighted according to the relative influence they exert on each other.

Importantly, activation flows in both directions, the *afferent*—from the sensory system to concepts and actions—and the opposite, *efferent*, direction.

Each node contains a floating-point activation value, α, which represents its excitatory state, may affect its internal processing, and is in turn forwarded (potentially altered by the node's processing) to the node's afferent connections.

A separate simulated activation value σ is also taken into account in the node's activation behavior and processing as follows: σ is added to the activation propagation when a node activates its afferent process nodes. Also, $\sigma + \alpha$ is used as a motor action trigger value in the action nodes. This allows us to model priming:

Priming

In humans, we observe the psychological phenomenon of "priming", or the bias (often measured as a decrease in response time) towards a previously triggered sensory or memory event. Such priming can occur through cross-modal activation, through previous activation, or from memory recall. Seen as a core element in fluent joint action, we can model priming through the efferent pathways in the modality streams: If a certain higher-level node n is activated through priming, the lower-level nodes that feed n are partially activated through the simulation value σ on the efferent pathway. As σ is added to the sensory-based activation α in the lower-level nodes, this top-down activation inherently lowers the perceptual activation necessary for the activation of those lower-level nodes, decreasing the real-world sensory-based activation threshold for action triggering. The result of this is reduced response time for anticipated sensory events, and increasingly automatic motor behavior.

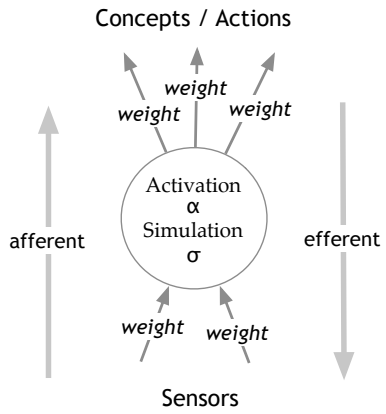


Figure 1: A process node within a modality stream. Weighted activation travels both up from sensory events to concepts an actions (the afferent pathway), and—through simulation—back downstream (the efferent pathway).

For example, let us assume a simple sensory activation stream which includes a sensor detecting the one-dimensional position $x \in [-1, 1]$ of an object of interest. This sensor feeds into two feature nodes, which correspond to the object being "left" or "right". In this example, the activation $\alpha \in [0, 1]$ of the "left" node would correspond to $max(-x, 0)$, and the activation of "right" to $max(x, 0)$. These two feature nodes feed, in turn, into a "left" action or a "right" action to be performed.

If the robot was primed toward a "left" perception (for example by a vocal command, through a related sensory or memory event, or by anticipation), the efferent connection to the "left" feature node would partially activate by receiving a simulation value σ. Then, even a slighter negative value of x would be sufficient to activate the feature node completely, resulting in an earlier appropriate action on the robot's part.

Practice Subsystems

In the proposed architecture, we use two subsystems to support practice.

History-based anticipatory simulation The first subsystem is a Markov-chain Bayesian predictor, building a probabilistic map of node activation based on recurring activation sequences during practice. This system is in the spirit of the anticipatory system described in (Hoffman and Breazeal 2007). It triggers high-level simulation, which—through the modality stream's efferent pathways—biases the activation of lower-level perceptual nodes.

If the subsequent sensory data supports these perceptual expectations, the robot's reaction times are shortened as de-

scribed above. In the case where the sensory data does not support the simulated perception, reaction time is longer and can, in some cases, lead to a short erroneous action, which is then corrected by the real-world sensory data. While slower, we believe that this "double-take" behavior, often mirrored by the human partner's motion, may contribute to the human's sense of similarity to the robot.

Inter-modal connection reinforcing An additional mechanism of practice is that of *weight reinforcement* on existing activation connections. While most node connections are fixed, some can be assigned to a connection reinforcement system, which will dynamically change the connection weights between the nodes. This system works according to the contingency principle, reinforcing connections that co-occur frequently and consistently, and decreasing the weight of connections that are infrequent or inconsistent.

This subsystem thus reinforces consistent coincidental activations, but inhibits competing reinforcements stemming from the same source node, leading to anticipated simulated perception of inter-modal perception nodes. This, again, triggers top-down biasing of lower-level perception nodes, shortening reaction times as described above.

Application

We have implemented an instantiation of the proposed architecture on a robotic system, which we subsequently used to evaluate our approach in a controlled human-subject study.

Robotic Platform

The robot employed in this evaluation was AUR, a robotic desk lamp, seen in Figure 2 (b). The lamp has a 5-degree-of-freedom arm 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 90 cm above the floor. Its processing is done on a 2x Dual 2.66GHz Intel processor machine located underneath the workbench.

The robot uses a Vicon motion capture system to identify and track the location and orientation of the human's right hand at a frequency of 10 times per second. This was made possible by a special glove with retroflective markers on it, worn on the human's right hand.

The system also takes input from Sphinx-4, an open-source speech recognition system created at Carnegie Mellon University, in collaboration with Sun Microsystems, Mitsubishi, and Hewlett Packard (Walker et al. 2004). The commands recognized by the system were: "Come", "Come Here", "Go", "Red", "Blue", "Green", and "Off".

Task Description

In the human-robot collaboration used in our studies, the human operates in a workspace as depicted in Figure 2 (a) and (b). The robot can direct its head to different locations, and change the color of its light beam. When asked to "Go", "Come", or "Come here" the robot would point towards the location of the person's hand, if the hand was relatively static. Additionally, the color changed in response to speech commands to one of three colors: blue, red, and green.

The workspace contained three locations (A,B,C). At each location there was a white cardboard square labeled with the location letter, and four doors cut into its surface. Each door, when opened, revealed the name of a color printed underneath.

The task was to complete a sequence of 8 actions, which was described in diagrammatical form on a sequence sheet as shown in Figure 2 (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. Neither the human nor the robot know the order of actions in the task sequences, or the names of the colors hidden behind the doors, at the beginning of the task.

History-based anticipatory simulation

Using a 3-step sequence history Markov model, the learner estimates the probability of the appropriate target board, and the "Go" action being expected. The probability of a certain concept to be triggered next is translated into a simulation value σ. This value is then propagated on the efferent pathway, biasing visual perceptual nodes. As a result, feature nodes simulate to the extent that they 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, and robot reaction time is decreased.

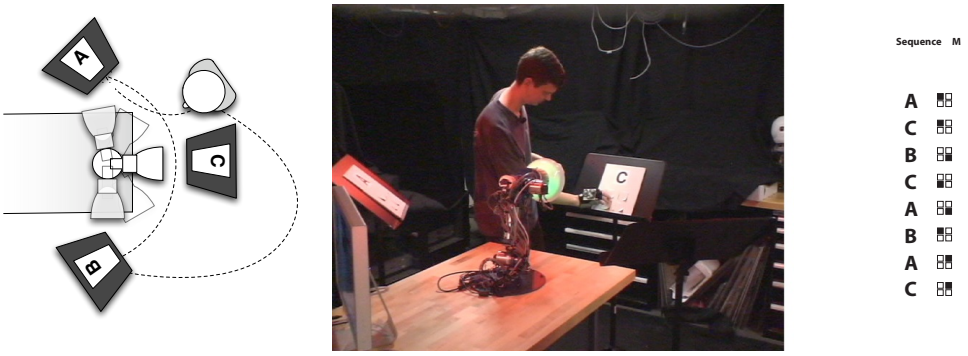
Inter-modal connection reinforcing

In addition, we used inter-modal simulation between the robot's proprioceptive property nodes, which sense the robot's joint configuration, and auditory feature nodes. Thus, certain physical configurations of the robot lead to the simulation of a certain word in the auditory stream, resulting in the perceptual simulation of that speech segment every time the robot reaches a certain position. If there is a consistent correlation between position and color, the robot will increasingly trigger the appropriate color without an explicit human command.

Experimental Design

To evaluate the validity of our approach to human-robot teamwork, we conducted a between-group controlled experiment with two conditions. The control (or `REACTIVE`) condition corresponds to the baseline condition in which no anticipatory simulation or cross-modal reinforcement occurred. The remainder of the system, i.e. the perceptual network and all activation streams and thresholds, were identically retained. In the second (`FLUENCY`) condition, the simulation subsystems were active with fixed parameters.

We recruited 38 subjects, who were arbitrarily designated to one of the two experimental conditions. At the last day of the experiment we experienced an unrecoverable hardware



(a) (b) (c)

Figure 2: (a) Diagram, and (b) photograph of the collaborative lighting task workspace. (c) Sample experimental sequence.

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.

In sections below 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. There are two rounds in a full **task** run.

Results

We have confirmed a number of behavioral hypotheses relating to the efficiency and fluency of the human-robot team, among them the following four metrics: a significant improvement in **task time** (REACTIVE: 1401.66 ± 162.90 secs, FLUENCY: 1196.26 ± 226.83 secs; t(24)=2.609, $p < 0.05$), a significant improvement in **mean sequence time** (REACTIVE: 141.38 ± 17.38 secs, FLUENCY: 116.21 ± 22.67 secs; t(30)=3.487, $p < 0.01$), in **human idle time** (REACTIVE: 0.46 ± 0.08 %, FLUENCY: 0.364 ± 0.09 %; t(30)=3.001, $p < 0.01$)[1], and in the robot's **functional delay** (REACTIVE: 4.81 ± 9.91 secs, FLUENCY: 3.66 ± 15.72 secs; t(30)=2.434, $p < 0.05$)[1]. These efficiency and fluency results are further elaborated upon in a separate publication (Hoffman and Breazeal 2008).

Relative contribution of human and robot

As both the human and the robotic team members undergo a learning curve of adapting to the collaborative task, we are interested in the relative contribution of each team member to the improvement of the team, comparing the learning rate of the human and the robot.

We estimated this measure as follows: For each sequence attempt, we compare two of the above-mentioned metrics to the value of the same metric in the first attempt of a given round. As the first round may include a few practice attempts, we only evaluate the second round for each subject.

Since the robot does not adapt or learn in the REACTIVE condition, we consider the improvement of the team in that group to be solely on behalf of the human. We call this "the human contribution" to the team's improvement. Subtracting the human contribution function from the improvement of the team in the FLUENCY condition, we obtain "the robot contribution" to the team's improvement.

Figure 3 (a) shows the relative contribution of the team members on the improvement in sequence time. We find that the rate of adaptation on the robot's part roughly matches that of the human, both contributing to about 20% of the reduction in sequence time over the course of a ten-attempt experimental round. We postulate that this phenomenon may contribute to an increased sense of partnership and "like-me" perception in human-robot teams.

In Figure 3 (b) we show the contribution to the robot's functional delay (a measure that has been shown to be related to team fluency). Again, we see a similar adaptation curve, but on this metric the robot's contribution converges on roughly twice that of the human, contributing to a circa 40% improvement compared to the human's circa 20%.

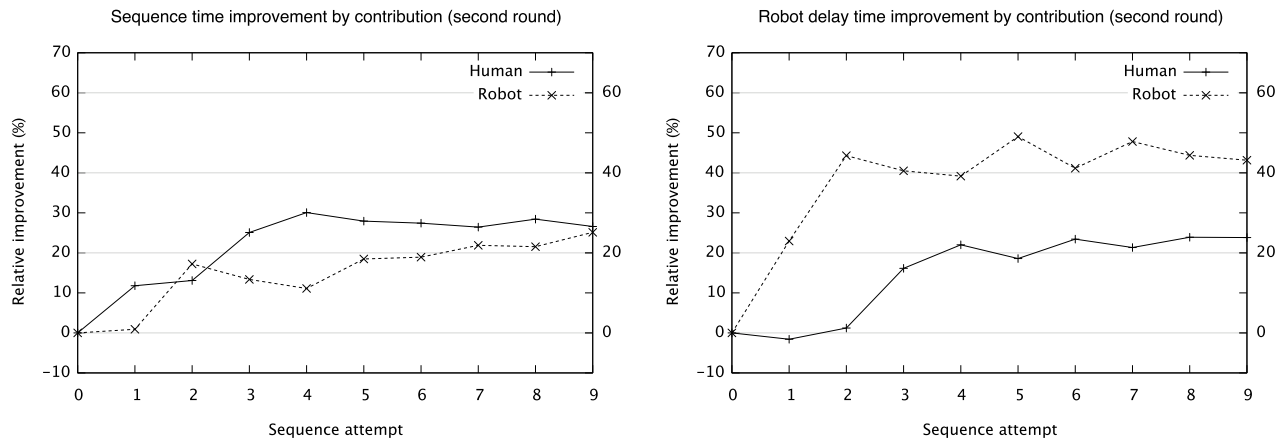
Self-Report Questionnaire

In addition to the behavioral metrics we have administered a self-report questionnaire including 41 questions. These questions were aimed to evaluate the human teammates' reaction to the robot with and and without perceptual simulation. 38 questions asked the subjects to rank agreement with a sentence on a 7-point Likert scale from "Strongly Disagree" (1) to "Strongly agree" (7). Three questions were open ended responses. We have compounded the questions into nine scales we propose to be valuable to evaluate fluent human-robot teamwork.

In this paper we will focus on the following four scales:

- FLUENCY — The sense of fluency in the teamwork;
- IMPROVE — The team's improvement over time;
- R-CONTRIB — The robot's contribution;
- R-TRUST — The human's trust in the robot;

[1] Adapted from (Hoffman and Breazeal 2007).



(a) (b)

Figure 3: Relative contribution of the team members on (a) sequence time, and (b) robot delay.

In addition we measure three individual questions:

- R-FLUENT — The robot's contribution to the fluency;
- H-COMMIT — The human's commitment to the task;
- R-ADAPT — The robot's adaptation to the human.

We hypothesized there to be a significant difference in these metrics between the two conditions, and specifically that these metrics be higher for the FLUENCY condition.

Table 1: Survey questionnaire results metrics. Values are $mean \pm s.d.$. on a 7-point Likert scale.

Metric	REACTIVE	FLUENCY	t(31)	
FLUENCY	4.98 ± 0.96	5.93 ± 0.98	2.80	**
IMPROVE	5.16 ± 0.96	6.17 ± 1.09	2.80	**
R-CONTRIB	2.85 ± 1.11	4.00 ± 1.32	2.69	*
R-TRUST	4.90 ± 1.25	5.42 ± 1.28	1.17	
R-FLUENT	4.73 ± 1.22	6.11 ± 1.18	3.28	**
H-COMMIT	6.40 ± 0.74	5.83 ± 1.10	1.70	
R-ADAPT	3.47 ± 1.46	5.94 ± 1.06	5.66	***

Table 1 shows the results for the questionnaire hypotheses, and reveals significant differences between subjects in the two experimental conditions with regard to the fluency scales in the questionnaire. Both the FLUENCY and the R-FLUENT measures are significantly different at $p < 0.01$. Additionally, subjects in the FLUENCY condition rated the robot's contribution to the team significantly higher than subjects in the REACTIVE condition, as well as the team's overall improvement. This supports our hypothesis that the proposed architecture contributes to the quality of fluency and collaboration in human-robot teams.

While these task-related scales differ significantly, we were not able to show a significant difference in the trust the human put in the robot, or in the human's commitment to the task—which was incidentally higher for the REACTIVE condition, if not significantly so. We believe that this is in part due to the low expectation people have of robots, which caused the evaluation of the REACTIVE robot to be high as a response to the robot's generally reliable functioning.

Open-ended responses The qualitative response of subjects in the open-ended responses of subjects in the FLUENCY condition was more favorable than that of subjects in the REACTIVE condition.

Positive comments in the FLUENCY condition included subjects reporting to be "highly impressed [with the robot's] learning", and a subject saying that the robot "worked well, and I felt a sense of relief/relaxation when it just did what I was about to tell it to do." Such positive comments were rare in the REACTIVE condition.

Several negative comments, in particular with regards to the robot's contribution as a team member, were found throughout the comments of subjects in the REACTIVE condition. These included "The robot was more of an assistance than an active team member", and "I felt like I was controlling the robot, rather than it being part of a team." In contrast, subjects in the FLUENCY condition remarked on the robot's contribution to the team, and referred to it several times as a teammate: "By the end of the first sequence I realized that he could learn and work as my teammate", and "my interaction with the robot was not that different than with a human teammate."

Self-deprecation in the FLUENCY condition A surprising result was that in the FLUENCY condition we found a high number of self-deprecating comments, and comments indicating worry or stress of fallible human performance in relation to the robot's strong performance. Several subjects in that condition remarked on stressful feelings that they weren't performing at an adequate level.

These remarks included "I would essentially forget the pair of colors I had [memorized] - this slowed me down", "The robot is better than me", "The performance could had been better if I didn't make those mistakes", "[I] worried

that I might slow my teammate down with any mistakes I might have made", and even "I am obsolete". There were no similar comments in the REACTIVE condition.

While it is beyond the scope of this work to further explore this aspect of our findings, it should be of interest to designers in the human-robot interaction field. The prevalence of this reaction may indicate a need for humans to feel more accomplished than the robot they are interacting with. Maintaining the balance of increased robot responsiveness, and the intimidation that might result is an overlooked aspect of HRI, which these results urge us to consider.

Lexical analysis We confirmed these anecdotal findings using an independent qualitative coding of the open question responses. Specifically, subjects in the FLUENCY condition commented on the robot more positively, and subjects in the REACTIVE condition commented on the robot more negatively. FLUENCY subjects attributed more human characteristics to the robot, although there is little difference in the emotional content of the comments.

Also, gender attributions, as well as attributions of intelligence occurred only in the FLUENCY condition, while subjects in the REACTIVE conditions tended to comment on the robot as being unintelligent. Finally, we did confirm the tendency to self-deprecating comments as more prevalent in the FLUENCY condition. A full description of the lexical analysis is available in a separate publication (Hoffman 2007).

Conclusion

For robots to act in fluently with a human partner, in a real-world situated teamwork scenario, they must overcome strict turn-taking behavior, which induces delays and inefficiencies, and can cause frustration. This is particularly true for a repetitive joint task, where the human teammate expects an increasingly meshed interaction with the robot.

We introduce a novel cognitive architecture aimed at achieving fluency in human-robot joint action. Based on neuro-psychological findings in humans, we propose a perceptual symbol system, which uses anticipatory simulation and inter-modal reinforcement to decrease reaction time through top-down biasing of perceptual processing.

We present a human subject study evaluating the effects of our approach, comparing it with a system using only bottom-up processing. We find significant differences in the task efficiency and fluency between the two conditions. Evaluating the relative contribution of the human and the robot, and find a similar learning curve, possible contributing to the human subjects' sense of similarity to the robot.

From self-report, we find significant differences in the perception of the team's fluency and the robot's contribution to that fluency, as well as in a number of other self-report metrics. Interestingly, we also find a tendency towards self-criticism in subjects collaborating with the anticipatory version of the robot.

References

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

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

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

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

Hoffman, G., and Breazeal, C. 2007. Cost-based anticipatory action-selection for human-robot fluency. *IEEE Transactions on Robotics and Automation* 23(5):952–961.

Hoffman, G., and Breazeal, C. 2008. Achieving fluency through perceptual-symbol practice in human-robot collaboration. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction (HRI'08)*. ACM Press.

Hoffman, G. 2007. *Ensemble: Fluency and Embodiment for Robots Acting with Humans*. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

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

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

Sebanz, N.; Bekkering, H.; and Knoblich, G. 2006. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences* 10(2):70–76.

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

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

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

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

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

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

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