Robotics: Science and Systems, Workshop on Human-Robot Collaboration. June 2013, Berlin, Germany.

Evaluating Fluency in Human-Robot Collaboration

Guy Hoffman
Media Innovation Lab, IDC Herzliya
P.O. Box 167, Herzliya 46150, Israel
Email: hoffman@idc.ac.il

***Abstract*—Collaborative fluency is the coordinated meshing of joint activities between members of a well-synchronized team. We aim to build robotic team members that can work side-by-side humans by displaying the kind of fluency that humans are accustomed to from each other. As part of this effort, we have developed a number of metrics to evaluate the level of fluency in human-robot shared-location teamwork. In this paper we discuss issues in measuring fluency, present both subjective and objective metrics that have been used to measure fluency between a human and robot, and report on findings along the proposed metrics.**

I. Introduction

When humans collaborate on a joint task, and especially when they are accustomed to the task and to each other, they can reach a high level of coordination, resulting in a well-synchronized meshing of their actions. Their timing is precise and efficient, they alter their plans and actions appropriately and dynamically, and this behavior emerges often without exchanging much verbal information.

We denote this quality of interaction the *fluency* of the joint activity, or in short, *collaborative fluency*, and in our research are interested in how robots could similarly perform more fluently with their human counterparts.

As it stands, most human-robot collaboration is structured in a stop-and-go fashion, inducing delays, and following a rigid command-and-response pattern. Collaboration with robots, where it occurs, holds little of the fluent quality which is part of a satisfying collaboration, the meshed "dance" that evokes both appreciation and confidence in a well-tuned human team.

We believe that for personal robots to play a long-term engaging role in untrained humans' lives, they must display a significantly more fluent coordination of their actions with that of their human counterparts.

The notion of fluency in human-robot collaboration is not well defined, and its meaning is not generally agreed upon. As can be seen by the description above, fluency is a somewhat vague and ephemeral notion. That said, we contend that fluency is a quality that can be positively assessed and recognized when compared to a non-fluent scenario. Moreover, we believe that tools for its evaluation are crucial for the design of successful robotic teammates.

In this paper we discuss various ways to measure the extent of fluency in a human-robot collaboration scenario, including subjective and objective metrics, and the relationship between the two. We also review recent work that has made use of these and other metrics in a number of shared-location human-robot collaborative task settings.

A. Related Work

The term "human-robot collaboration" has a number of meanings in the HRI literature. Some frame it in the context of mixed-initiative control and shared autonomy, arbitrating between a remote robot's autonomy and direct human control (e.g. [2]). In this work, however, we focus only on the collaboration between a human and an autonomous robot at a *shared location*, making use of the co-located partners' behavior to achieve a joint goal.

In early shared-location collaboration work, Kimura et al. [10] study a robotic arm assisting a human in an assembly task. Their work addresses issues of vision and task representation, but does not investigate timing or fluency. In our own earlier work, we investigate turn-taking and joint plans, mostly in the context of verbal and non-verbal dialog [7]. That work also does not include overlapping action or questions of fluency.

Sakita et al. [14] use a robot to assist a human in an assembly task. The robot intervenes in one of three ways: taking over for the human, disambiguating a situation, or executing an action simultaneously with a human. While relying on some nonverbal symbols, the interaction described is also strictly turn-based. More recent work in this vein [**?** 1] investigates mechanisms to coordinate joint activities, and in particular when a breakdown in the joint task coordination occurs. None of these deal directly with timing or the fluent meshing of the coordinated activity. Another body of research in shared-location human-robot collaboration is concerned with the mechanical coordination and safety considerations of robots in shared tasks with humans (e.g. [9]).

Work in rhythm-related HRI directly addresses the notion of timing. Weinberg and Driscoll [15] include nonverbal behavior and physically-based anticipation in their "Haile" robotic drummer project. Michalowski et al. [11] study the effects of rhythmic movement of a beat-tracking dancing robot. Neither, however, are directly related to the achievement of a joint task.

Examples of work specifically dealing with fluency of shared-workspace collaboration includes anticipatory action systems in shared-workspace MDPs [5], perceptual simulation in joint tasks [6], fluency of object handovers from a robot to a human [3], timing in multi-modal turn-taking interactions [4], and human-robot cross-training for shared learning in human-robot teams [12]. We discuss these works in detail in Section V.

II. Characteristics of Collaborative Fluency

A. Fluency vs Efficiency

Team fluency is related to task efficiency, defined simply as the inverse of the time it takes to complete identical tasks or subtasks. One would assume that a more fluent interaction should be more efficient. However, we have found that the two are not directly correlated.

Indeed, the need to separately measure the fluency of an interaction arose in the evaluation of a framework for human-robot collaboration, in which we found that participants rated their experience as significantly more fluent, even when there was no difference in efficiency of the task completion [5].

This finding suggests that collaborative fluency is a separate feature of the joint activity, requiring separate metrics.

B. Subjective vs Objective Fluency Metrics

To that end, we developed two types of fluency metrics for human-robot collaboration: *subjective* metrics, which are based on people's perception of the fluency of an interaction; and *objective* metrics, which can quantitatively estimate the degree of fluency in a given interaction.

Subjective fluency metrics include both direct measures of fluency that people attach to a collaboration, and downstream outcomes of the perceived fluency, such as the trust human collaborators put in the robot, or their sense that the robot is committed to the team.

C. Observer vs Participant Fluency Perception

When evaluating subjective fluency perception, we need to separate the fluency perceived by a bystander watching a collaborative interaction, and the fluency experienced by the human participant in a human-robot team.

We denote these two categories *observer* and *participant* fluency perception, respectively. In our own work we found anecdotally, that even when observers do not detect a difference in collaborative fluency between two interactions, participants do. This suggests that participation is more sensitive to fluency than observation.

In Section V, we review both work that evaluates observer fluency perception and work that evaluates participant fluency perception, although this distinction is not usually made explicit.

III. Subjective Fluency Metrics

Subjective fluency metrics assess how fluent people *perceive* the collaboration to be. We use questionnaires to rate agreement with fluency notions, including both single statements and composites of indicators related to the same measure.

In addition to directly evaluating fluency, we explore possible downstream outcomes of collaborative fluency. These outcomes can include the perceived intelligence of the robot, the perceived reliability of the robot, the trust humans put in it, or the contribution of the robot to the team.

It should be noted that there are currently no accepted practices, instruments, or measures to evaluate fluency in human-robot collaboration. This section presents a review of subjective measures that we and others have used in the past to measure aspects of fluency, as a basis for discussion towards future human-robot fluency studies.

A. Composite Measures

To evaluate people's sense of human-robot fluency, we have used the following composite measures. They include one direct measure of fluency, and several downstream measures.

Note that the measures are phrased for the participant fluency perception scenario, but can be adjusted for observer fluency perception, where necessary. We report Cronbach's alpha as measured in our most recent human-robot collaborative fluency study using these measures [6].

1) Human-Robot Fluency: This composite measure evaluates the overall fluency between the human and the robot, and consists of three indicators:

- "The human-robot team worked fluently together."
- "The human-robot team's fluency improved over time."[1]
- "The robot contributed to the fluency of the interaction."

Cronbach's alpha for this measure was found to be 0.801.

2) Robot Contribution: This composite downstream measure evaluates the robot's contribution to the team, and consists of two indicators:

- "I had to carry the weight to make the human-robot team better." (reverse scale)
- "The robot contributed equally to the team performance."
- "I was the most important team member on the team." (reverse scale)
- "The robot was the most important team member on the team."

Cronbach's alpha for this measure was found to be 0.785.

3) Trust in Robot: This composite downstream measure evaluates the trust the robot evokes, and consists of two indicators:

- "I trusted the robot to do the right thing at the right time."
- "The robot was trustworthy."

Cronbach's alpha for this measure was found to be 0.772.

4) Robot Teammate Traits: This composite downstream measure evaluates the robot's perceived character traits related to it being a team member, and consists of three indicators:

- "The robot was intelligent."
- "The robot was trustworthy."
- "The robot was committed to the task."

Cronbach's alpha for this measure was found to be 0.827.

5) Working Alliance for Human-Robot Teams: We have adapted an existing instrument, the "Working Alliance Index" (WAI) [8], measuring the quality of working alliance between humans, to the human-robot teamwork scenario. This downstream measure is made up of two sub-scales, the "bond" subscale and the "goal" sub-scale, in addition to one additional individual question.

The "bond" sub-scale consists of the following seven indicators:

- "I feel uncomfortable with the robot." (reverse scale)
- "The robot and I understand each other."
- "I believe the robot likes me."

[1]This question relates specifically to the adaptive aspect of fluency, and is only appropriate in a robot learning or adaptation scenario.

- "The robot and I respect each other."
- "I am confident in the robot's ability to help me."
- "I feel that the robot appreciates me."
- "The robot and I trust each other."

Cronbach's alpha for this measure was found to be 0.808.

The "goal" sub-scale consists of the following three indicators:

- "The robot perceives accurately what my goals are."
- "The robot does not understand what I am trying to accomplish." (reverse scale)
- "The robot and I are working towards mutually agreed upon goals."

Cronbach's alpha for this measure was found to be 0.794.

The complete composite measure additionally includes the following indicator:

- " I find what I am doing with the robot confusing." (reverse scale)

Cronbach's alpha for the overall WAI was found to be 0.843.

6) Improvement: This composite measure is only applicable for a learning and adaptation scenario, and consists of three indicators:

- "The human-robot team improved over time"
- "The human-robot team's fluency improved over time."
- "The robot's performance improved over time."

Cronbach's alpha for this measure was found to be 0.793.

B. Individual Measures

We have also found it useful to evaluate some of the above, and additional, indicators individually. Additional individual measures include:

- "The robot's performance was an important contribution to the success of the team."
- "It felt like the robot was committed to the success of the team."
- "I was committed to the success of the team."

C. Additional Indicators

As these measures have been validated only in a limited setting, we find it useful to also report on indicators that we have not found successful in the evaluation of fluency. Further study is merited to examine these measures with respect to the perceived fluency of the human-robot collaboration.

These indicators include:

- "The human-robot team did well on the task."
- "The robot performed well as part of the team."
- "The human-robot team felt well-tuned."
- "The robot did its part successfully."

IV. Objective Fluency Metrics

In addition to subjective measures, we want to attain objective measures that could serve as benchmarks to evaluate fluency in human-robot collaboration. We propose four measures relating to the fluency of an interaction, and which we have used to estimate the contribution to fluency of various learning and task collaboration algorithms. All of these measures are task-agnostic, and relate only to the periods of action. Also, they are generally understood as between a two-member team, with one human and one robot team member.

A. Robot Idle Time

The first measure is the rate of *robot idle time*. This corresponds to the percentage of the total task time that the robot was not active. Robot idle time occurs in situations in which the robot waits for additional input from the human, is processing input, is computing a decision, is waiting for additional sensory input, or is waiting for the human to complete an action.

B. Human Idle Time

The symmetric measure is the rate of the *human idle time*. This corresponds to the percentage of the total task time that the human was not active. As humans usually have more information in human-robot collaborative tasks, and faster perceptual processing, we found that—more often than not—human idle time is due to the human waiting for the robot to complete an action in order for them to do the next step of the collaboration.

In terms of the sense of fluency, human idle time can be perceived as boredom, time wasted, or an imbalance between team members.

C. Concurrent Activity

A third measure is the rate of *concurrent activity*. This corresponds to the percentage of time out of the total task time, during which both agents have been active at the same time. Another way to understand this measure is the amount of action overlap between the two agents.

D. Functional Delay

The forth measure is the rate of *functional delay* experienced by the agents. This is the accumulated time, as a ratio of total task time, between the completion of one agent's action, and the beginning of the other agent's action.

Note that this measure can be larger than 1, if the accumulated functional delay is longer than the total task time. This occurs if within a task of length t there are n actions by an agent, with a mean functional delay of d for each action, and $\frac{t}{n} < d < t$.

Functional delay can also be negative, in the case that actions are overlapping.

The functional delay can be calculated for both agents together, or for each agent separately. However, in our experience we have found that the functional delay imposed by the human is usually negligable, so that the total functional delay is equal to the functional delay imposed by the robot (i.e. the time between the end of the human's action and the onset of the robot's action). We therefore usually consider only this metric.

E. Examples

The four metrics laid out above are, of course, interrelated, as they are all a function of the amount and timing of each agent's action. However, they are not interchangeable. One measure can improve while another regresses.

To illustrate the interplay between the various measures in some common scenarios, we analyze three template scenarios.

Figure 1 shows a strictly alternating turn taking scenario, in which each action by the agent is immediately followed by the next action of the other agent. The imbalance in idle times is due to the fact that the robot starts the interaction. Strict alternation results in no functional delay and no concurrent action.

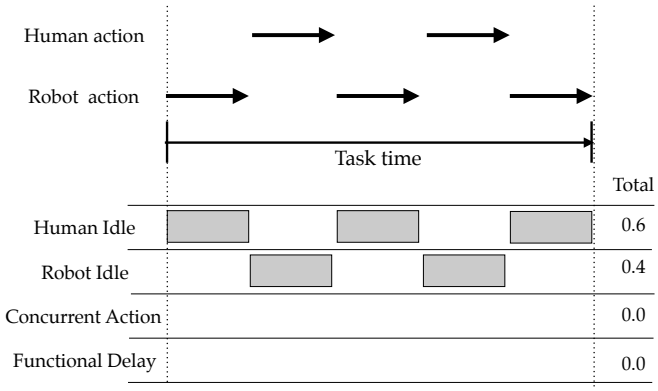

Fig. 1. Objective fluency metrics in a fully separated turn taking scenario with no processing delays induced by either agent.

Figure 2 shows a similar interaction to the previous example, with the exception that the human starts the task, and the robot has some processing time after the human's action is complete. In this example, the robot needs the full human action to complete before being able to process it and select its own action. A common example of this scenario is turn taking with perceptual delay, such as speech recognition.

On the one hand, the result is a more balanced idle time between the two agents, due to the increase in robot idle time, and the same human idle time as in the previous example. However, the robot's processing incurs a functional delay on the interaction. And, since this is still a strict turn-taking scenario, there is no concurrent action between the agents.

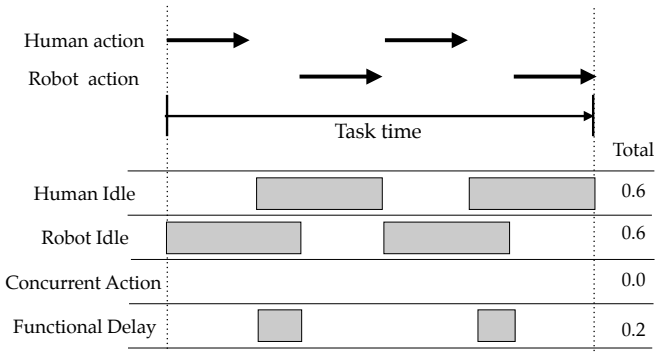

Fig. 2. Objective fluency metrics in a fully separated turn taking scenario in which the robot has a processing delay with respect to the human's fully completed action.

Finally, Figure 3 shows an interaction in which the human can start their part while the robot is still working on its last action. Again, the robot has a functional delay. In this case, the concurrent action measure is non-zero, and the functional delay slightly reduced. Both human and robot idle times are the same as in the first example.

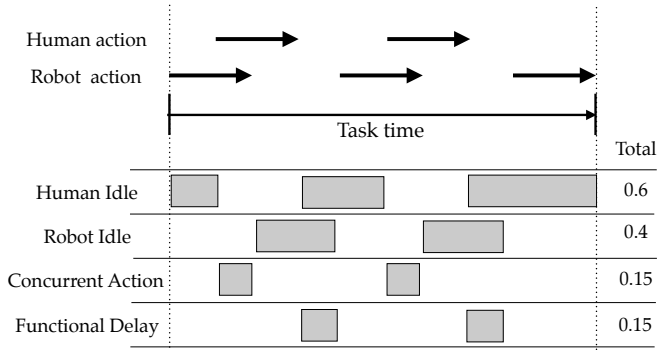

Fig. 3. Objective fluency metrics in a scenario in which the robot has a processing delay, but the human can start their action before the robot's action is completed.

F. Validating the Objective Metrics

We are currently conducting a large-scale study relating the objective fluency metrics to subjective notions of fluency. As part of this study, we have developed a simple human-robot collaborative scenario with flexible timing on both agents' part.

The scenario is a joint workspace (Figure 4), in which the human and the robot must transfer a number of objects from the right (human) end table of the workspace to the left (robot) end table. In order to do this, the human hands over the object to the robot by placing it on the shared (middle) table.

We are using this model in both an observer and a participant perception setup. In the observer perception study, participants watch videos of various collaborative scenarios controlled for the objective fluency metrics. We then measure their subjective fluency metrics and relate the two aspects of fluency. In the participant perception study, the participant controls the human behavior and the robot adapts according to a set number of behavior patterns, aimed at varying objective fluency metric outcomes. Again, we then relate the subjective and objective metrics in these interactions.

V. Usage of Fluency Metrics in Past Research

While the metrics proposed here should still be considered a work-in-progress, they have been used both in our work and in other studies.

A. Anticipatory Action in Collaborative MDP

Adaptive Anticipatory Action Selection is a method for meshing an agent's action with that of a human in a shared workspace collaborative task [5]. A human-subject study was conducted, evaluating the effects of this method when compared to a reactive (turn-taking) method. There was neither a significant difference in the mean task efficiency, nor in the final convergent task efficiency between the anticipatory and the reactive behavior.

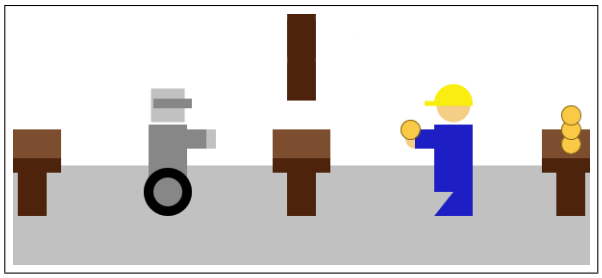

Fig. 4. Joint activity scenario modeling a simple timed handover task, used to evaluate the relation between objective and subjective fluency metrics.

Subjects were asked to rate a subset of five of the subjective metrics described above. We found significant differences in the rating of the following phrases: "The robot's performance was an important contribution to the success of the team"; "The robot contributed to the fluency of the interaction"; and "It felt like the robot was committed to the success of the team". No significant differences were present in the rating of the phrases "I was committed to the success of the team" (since removed from the fluency metrics); and "I trusted the robot to do the right thing at the right time".

In terms of objective metrics, the rate of concurrent motion was significantly higher in the anticipatory group, settling at about twice the rate compared to the reactive group. We also found a significantly lower functional delay in the anticipatory group, and especially as the interaction progressed. There was no difference in human idle time between the groups.

B. Perceptual Simulation for Joint Activities

Another study evaluated the effects of Anticipatory Perceptual Simulation, a computational cognitive framework that simulates priming for robots working with humans on a collaborative task [6].

In a human subject study, participants rated the interaction on a questionnaire made up of the composite measures described above, and additional composite measures not included in the fluency metric set. There were significant differences in human-robot fluency, the improvement of the team, the robot's contribution, and the WAI goal sub-scale. There were also significant differences on the individual measures "The robot contributed to the fluency of the interaction", and "The robot learned to adapt its actions to mine". We did not find significant differences in the composite measure of the trust in the robot, the robot's character, the WAI bond sub-scale, or the overall WAI scale. In addition, we did not find differences in the human's commitment to the task, a measure since removed from the set of subjective fluency metrics.

Objective task efficiency was measured and found to improve by using anticipatory perceptual simulation. In addition, two objective fluency metrics were measured: human idle time, and the functional delay incurred by the robot. Both were found to have been positively affected by the algorithm, with an increasing improvement of robot functional delay as the interaction progressed, indicating the robot's adaptation to the human's action timing.

C. Fluency of Handovers

Cakmak *et al.* have developed methods to enable more fluent hand-over of an object from a robot to a human [3]. They have specifically investigated the effects of spatial contrast—making the handover pose distinct from other poses—and temporal contrast—accentuating the timing of the handover gesture—on the fluency of the handover.

A survey was used to estimate the readability of handovers, and in an experimental human-subject study, two objective measures of fluency were evaluated across a factorial variable set. These metrics were the human functional delay, and the robot functional delay. The researchers have found that temporal contrast positively effects human functional delay in hand-over tasks.

D. Timed Petri Nets for Multi-modal Turn-taking

Chao and Thomaz designed a system based on timed petri nets to enable multi-modal turn-taking and joint action meshing. The system is designed for overlapping actions, both in the verbal and in the non-verbal modality, and it specifically aims to achieve fluency in a joint task.

A human-subject study compared a robot using the system to allow for action-interruption to an action-completing baseline robot, in a joint puzzle-solving interaction. Participants rated several subjective fluency metrics relating to the relative contribution, trust, and naturalness of the interaction. Participants in the interruption condition rated their mental contribution higher, and rated the interaction as less "awkward" than those in the baseline condition. Task efficiency was used as an objective metric of team fluency.

E. Cross-Training for Human-Robot Joint Learning

Nikolaidis and Shah have proposed human-robot cross-training to improve adaptation of human-robot teams [12]. Cross-training is a method used in human teams where team members switch roles to train on both sides of a shared plan.

The researchers used a human-subject study to compare the cross-training method with a standard reinforcement learning algorithm. The study used objective metrics to evaluate mental model similarity and convergence. It also used objective and subjective metrics to evaluate the fluency of the resulting interaction.

In terms of subjective fluency metrics, the study used both individual items from the "trust in robot" measure, and adapted the following two items from the WAI "goal" sub-scale: "[The robot] does not understand how I am trying to execute the task" (reverse scale); and "[The robot] perceives accurately what my preferences are". All four measures were found to be significantly higher for the cross-training condition, compared to the traditional machine learning condition.

The study also evaluated three objective fluency metrics: the rate of concurrent motion, the human idle time, and the robot idle time. The first two measures were coded by a single coder from video of the interaction, while the third was automatically gleaned from the robot's logs. The researchers found a significant improvement in all three objective metrics.

VI. Conclusion and Extensions

In this paper, we proposed a concept of human-robot collaborative fluency, the coordination and meshing of actions by team members. As part of the development of robots that display collaborative fluency, we presented metrics to evaluate fluency in human-robot collaboration. Subsets of these metrics have been used in the past years to evaluate fluency, both in our work, and in other work concerned with the meshing of actions between humans and robots working on a shared task.

We have presented composite subjective measures, made up of items we found internally valid, as well as individual indicators used in human-robot collaboration studies. Further, we presented four objective measures that provide benchmarks for evaluating the fluency of a collaborative interaction.

These metrics are an evolving work-in-progress. Over the years, we have added, refined, and removed some of these metrics from our inventory. We are currently in the process of relating the objective and subjective metrics to converge on a generally agreed set of measures.

There are aspects of collaborative fluency, which these metrics not yet address, and should be considered for future work. These include: how to take into account correct and incorrect actions of the robot and the human? Does the role relationship between human and robot (e.g. supervisor, subordinate, or peer—as proposed by **?**]) effect perceptions of fluency? How to account for corrections and repetitions of identical actions? And how to extend these measures to larger mixed teams than just one human and one robot?

The proposed metrics themselves also leave room for extension, for example the use of standard metrics for downstream measures, such as cognitive load [13] or trust, as well as the relative contribution of the different objective metrics to collaborative fluency.

In conclusion, we believe that a validated set of human-robot fluency metrics can greatly advance the goal of robotic teammates accepted for long-term collaboration with humans, be it in the workplace, school, or home.

References

[1] M Awais and D Henrich. Proactive premature intention estimation for intuitive human-robot collaboration. In *2012 IEEE/RSJ Intl Conference on Intelligent Robots and Systems (IROS)*, pages 4098—-4103, 2012.

[2] D J Bruemmer, D D Dudenhoeffer, and J Marble. Dynamic Autonomy for Urban Search and Rescue. In *2002 AAAI Mobile Robot Workshop*, Edmonton, Canada, August 2002.

[3] M Cakmak, S Srinivasa, M Kyung Lee, S Kiesler, and J Forlizzi. Using spatial and temporal contrast for fluent robot-human hand-overs. *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, page 489, 2011.

[4] C Chao and A L Thomaz. Timing in Multimodal Turn-Taking Interactions : Control and Analysis Using Timed Petri Nets. *Journal of Human-Robot Interaction*, 1(1): 4–25, 2012.

[5] G Hoffman and C Breazeal. Cost-Based Anticipatory Action-Selection for Human-Robot Fluency. *IEEE Transactions on Robotics and Automation*, 23(5):952–961, October 2007.

[6] G Hoffman and C Breazeal. Effects of anticipatory perceptual simulation on practiced human-robot tasks. *Autonomous Robots*, 28(4):403–423, December 2009.

[7] Guy Hoffman and Cynthia Breazeal. Collaboration in Human-Robot Teams. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference*, Chicago, IL, USA, September 2004. AIAA.

[8] A O Horvath and L S Greenberg. Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, 36(2):223–233, 1989.

[9] 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.

[10] H Kimura, T Horiuchi, and K Ikeuchi. Task-Model Based Human Robot Cooperation Using Vision. In *Proc of the IEEE International Conf on Intelligent Robots and Systems (IROS)*, pages 701–706, 1999.

[11] M Michalowski, S Sabanovic, and H Kozima. A Dancing Robot for Rhythmic Social Interaction. In *HRI '07: Proc of the ACM/IEEE Int'l Conf on Human-robot interaction*, pages 89–96, Arlington, Virginia, USA, March 2007.

[12] Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 33–40. IEEE Press, 2013.

[13] F Paas, J E Tuovinen, H Tabbers, and P W M Van Gerven. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38(1):63–71, March 2003.

[14] K. Sakita, K. Ogawara, S. Murakami, K. Kawamura, and K. Ikeuchi. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 846–851, 2004.

[15] G Weinberg and S Driscoll. Robot-Human Interaction with an Anthropomorphic Percussionist. In *Proc of the ACM Conf on Human Factors in Computing (CHI)*, pages 1229–1232, 2006.