Regulation and Entrainment in Human-Robot Interaction

Dr. Cynthia Breazeal
MIT Artificial Intelligence Lab
Cambridge, MA 02139 USA
cynthia@ai.mit.edu

Abstract:

Newly emerging robotics applications for domestic or entertainment purposes are slowly introducing autonomous robots into society at large. A critical capability of such robots is their ability to interact with humans, and in particular, untrained users. This paper explores the hypothesis that people will intuitively interact with robots in a natural social manner provided the robot can perceive, interpret, and appropriately respond with familiar human social cues. Two experiments are presented where naive human subjects interact with an anthropomorphic robot. Evidence for mutual regulation and entrainment of the interaction is presented, and how this benefits the interaction as a whole is discussed.

1. Introduction

New applications for domestic, health care related, or entertainment based robots motivate the development of robots that can socially interact with, learn from, and cooperate with people. One could argue that because humanoid robots share a similar morphology with humans, they are well suited for these purposes – capable of receiving, interpreting, and reciprocating familiar social cues in the natural communication modalities of humans.

However, is this the case? Although we can design robots capable of interacting with people through facial expression, body posture, gesture, gaze direction, and voice, the robotic analogs of these human capabilities are a crude approximation at best given limitations in sensory, motor, and computational resources. Will humans readily read, interpret, and respond to these cues in an intuitive and beneficial way?

Research in related fields suggests that this is the case for computers [1] and animated conversation agents [2]. The purpose of this paper is to explore this hypothesis in a robotic media. Several expressive face robots have been implemented in Japan, where the focus has been on mechanical engineering design, visual perception, and control. For instance, the robot in the upper left corner of figure 1 resembles a young Japanese woman (complete with silicone gel skin, teeth, and hair [3]. The robot's degrees of freedom mirror those of a human face, and novel actuators have been designed to accomplish this in the desired form factor. It can recognize six human facial expressions and can

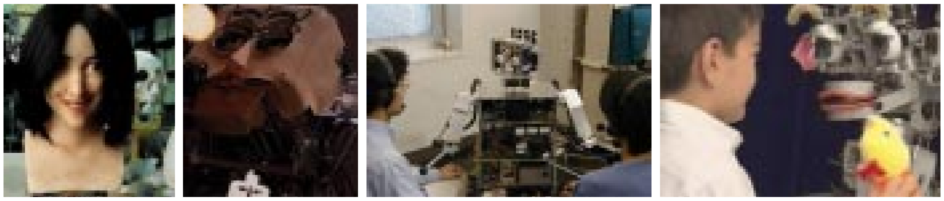


Figure 1. A sampling of robots designed to interact with people. The far left picture shows a realistic face robot designed at the Science University of Tokyo. The middle left picture shows *WE-3RII*, an expressive face robot developed at Waseda University. The middle right picture shows *Robita*, an upper-torso robot also developed at Waseda University to track speaking turns. The far right picture shows our expressive robot, *Kismet*, developed at MIT. The two leftmost photos are courtesy of Peter Menzel [6].

mimic them back to the person who displays them. In contrast, the robot shown in the upper right of corner of figure 1 resembles a mechanical cartoon [4]. The robot gives expressive responses to the proximity and intensity of a light source (such as withdrawing and narrowing its eyelids when the light is too bright). It also responds expressively to a limited number of scents (such as looking drunk when smelling alcohol, and looking annoyed when smoke is blown in its face). The lower right picture of figure 1, shows an upper-torso humanoid robot (with an expressionless face) that can direct its gaze to look at the appropriate person during a conversation by using sound localization and head pose of the speaker [5].

In contrast, the focus of our research has been to explore dynamic, expressive, pre-linguistic, and relatively unconstrained face to face social interaction between a human and an anthropomorphic robot called Kismet (see lower right of figure 1). For the past few years, we have been investigating this question in a variety domains through an assortment of experiments where naive human subjects interact with the robot. This paper summarizes our results with respect to two areas of study: the communication of affective intent and the dynamics of proto-dialog between human and robot. In each case we have adapted the theory underlying these human competencies to Kismet, and have experimentally studied how people consequently interact with the robot. Our data suggests that naive subjects naturally and intuitively read the robot's social cues and readily incorporate them into the exchange in interesting and beneficial ways. We discuss evidence of communicative efficacy and entrainment that results in an overall improved quality of interaction.

2. Communication of Affective Intent

Human speech provides a natural and intuitive interface for both communicating with humanoid robots as well as for teaching them. Towards this goal, we have explored the question of recognizing affective communicative intent in robot-directed speech. Developmental psycholinguists can tell us quite a lot about how preverbal infants achieve this, and how caregivers exploit it to regulate the infant's behavior. Infant-directed speech is typically quite exaggerated in the pitch and intensity (often called *motherese*). Moreover, mother's intuitively use selective prosodic contours to express different communicative intentions. Based on a series of cross-linguistic analyses, there appear to be at least four different pitch contours (approval, prohibition, comfort, and attentional bids), each associated with a different emotional state [7]. Figure 2 illustrates these four prosodic contours.

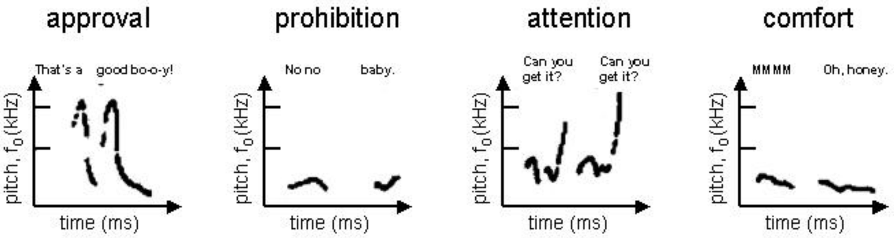


Figure 2. Fernald's prototypical prosodic contours for approval, attentional bid, prohibition, and soothing.

Mothers are more likely to use falling pitch contours than rising pitch contours when soothing a distressed infant [8], to use rising contours to elicit attention and to encourage a response [9], and to use bell shaped contours to maintain attention once it has been established [10]. Expressions of approval or praise, such as "Good girl!" are often spoken with an exaggerated rise-fall pitch contour with sustained intensity at the contour's peak. Expressions of prohibitions or warnings such as "Don't do that!" are spoken with low pitch and high intensity in staccato pitch contours. Fernald suggests that the pitch contours observed have been designed to directly influence the infant's emotive state, causing the child to relax or become more vigilant in certain situations, and to either avoid or approach objects that may be unfamiliar [7].

Inspired by these theories, we have implemented a recognizer for distinguishing the four distinct prosodic patterns that communicate praise, prohibition, attention, and comfort to preverbal infants from neutral speech. We have integrated this perceptual ability into our robot's *emotion system*, thereby allowing a human to directly manipulate the robot's affective state which is in turn reflected in the robot's expression.

2.1. The Classifier Implementation

We made recordings of two female adults who frequently interact with Kismet as caregivers. The speakers were asked to express all five communicative intents (approval, attentional bid, prohibition, soothing, and, neutral) during the interaction. Recordings were made using a wireless microphone whose output was sent to the speech processing system running on Linux. For each utterance, this phase produced a 16-bit single channel, 8 kHz signal (in a `.wav` format) as well as its corresponding pitch, percent periodicity, energy, and phoneme values. All recordings were performed in Kismet's usual environment to minimize variability in noise due to the environment.

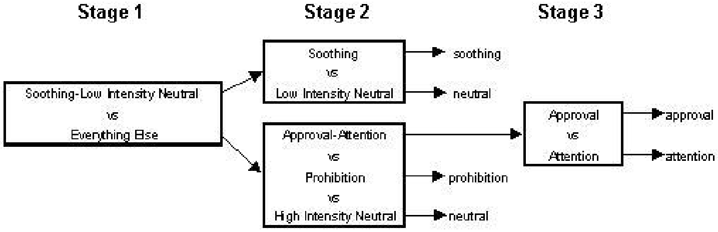

Stage 1
Stage 2
Stage 3
Soothing-Low Intensity Neutral vs Everything Else
Soothing vs Low Intensity Neutral
soothing
neutral
Approval-Attention vs Prohibition vs High Intensity Neutral
prohibition
neutral
Approval vs Attention
approval
attention

Figure 3. The classification stages.

The implemented classifier consists of several mini classifiers executing in stages (as shown in figure 3). In all training phases we modeled each class of data using the Gaussian mixture model, updated with the EM algorithm and a Kurtosis-based approach for dynamically deciding the appropriate number of kernels [11]. In the beginning stages, the classifier uses global pitch and energy features to separate the classes based on arousal measures (see fig 4). The remaining clustered classes were then passed to later classification stages that used features that carefully encoded the shape of the contours (as suggested by Fernald). These findings are consistent with Fernald's work and proved useful in separating the *difficult* classes. The classifier's structure follows logically from these observations.

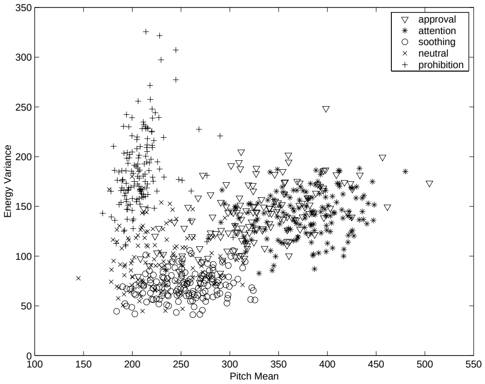


Figure 4. Feature space of all five classes.

The output of the recognizer is integrated into the rest of Kismet's synthetic nervous system as shown in figure 5. Due to space limitations, we leave the details to the interested reader as described in [12]. For our purposes here, the result of the classifier is passed to the robot's higher level perceptual system where it is combined with other contextual information. The result of the classifier can bias the robot's affective state by modulating the arousal and valence parameters of the robot's *emotion system*. The emotive responses are designed such that praise induces positive affect (a happy expression), prohibition induces negative affect (a sad expression), attentional bits enhance arousal (an alert expression), and soothing lowers arousal (a relaxed expression). The net affective/arousal state of the robot is displayed on its face and expressed through body posture [13], which serves as a critical feedback cue to the person who is trying to communicate with the robot. This expressive feedback serves to close the loop of the human-robot system.

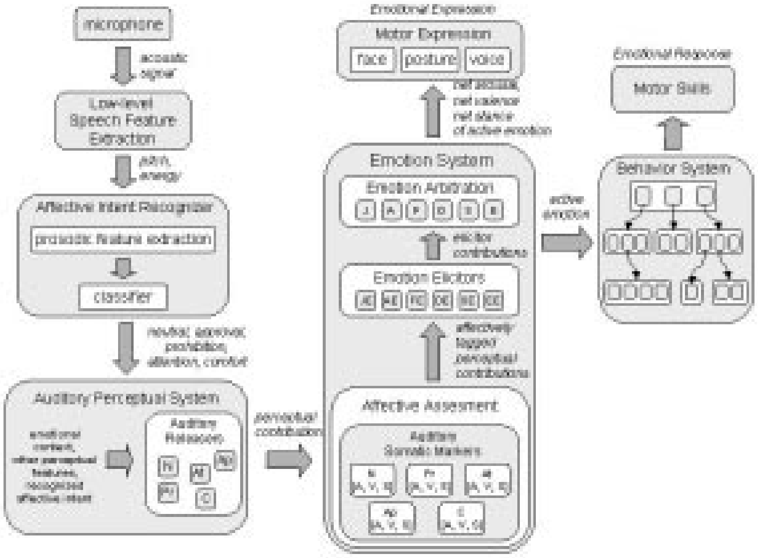
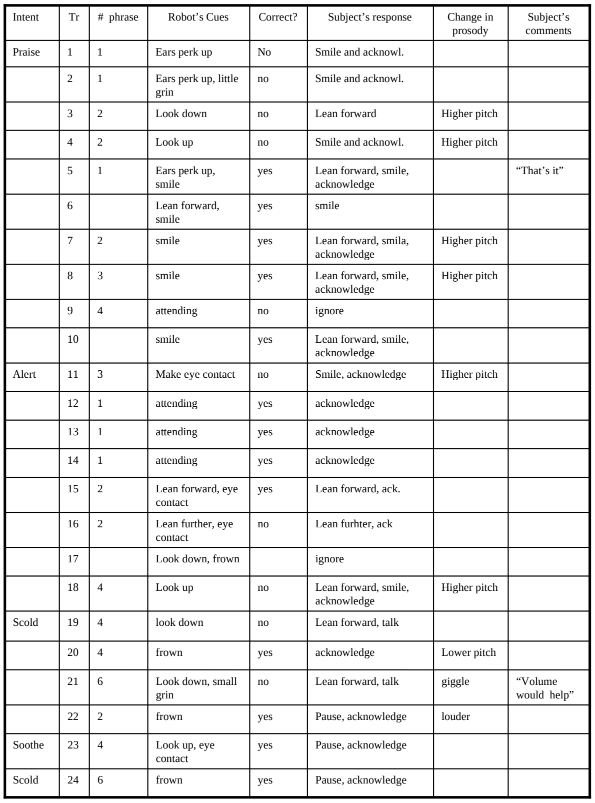


Figure 5. The output of the affective intent classifier is passed to the robot's *emotion system*, where it can influence the robot's affective state, its facial expression, and its behavior. The classifier output is first combined with other contextual information in the higher level perceptual system. These perceptions are then assessed for affective impact with respect to how they contribute to the robot's arousal, valence and stance parameters. This information is used to elicit the most relevant emotional response, that subsequently modulates the robot's expressive and behavioral response.

2.2. Affective Intent Experiment

Communicative efficacy has been tested with people very familiar with the robot as well as with naive subjects in multiple languages (French, German, English, Russian, and Indonesian). Female subjects ranging in age from 22 to

54 were asked to praise, scold, soothe, and to get the robot's attention. They were also asked to signal when they felt the robot "understood" them. All exchanges were video recorded for later analysis.



Intent	Tr	# phrase	Robot's Cues	Correct?	Subject's response	Change in prosody	Subject's comments
Praise	1	1	Ears perk up	No	Smile and acknowl.		
	2	1	Ears perk up, little grin	no	Smile and acknowl.		
	3	2	Look down	no	Lean forward	Higher pitch	
	4	2	Look up	no	Smile and acknowl.	Higher pitch	
	5	1	Ears perk up, smile	yes	Lean forward, smile, acknowledge		"That's it"
	6		Lean forward, smile	yes	smile		
	7	2	smile	yes	Lean forward, smila, acknowledge	Higher pitch	
	8	3	smile	yes	Lean forward, smile, acknowledge	Higher pitch	
	9	4	attending	no	ignore		
	10		smile	yes	Lean forward, smile, acknowledge		
Alert	11	3	Make eye contact	no	Smile, acknowledge	Higher pitch	
	12	1	attending	yes	acknowledge		
	13	1	attending	yes	acknowledge		
	14	1	attending	yes	acknowledge		
	15	2	Lean forward, eye contact	yes	Lean forward, ack.		
	16	2	Lean further, eye contact	no	Lean furhter, ack		
	17		Look down, frown		ignore		
	18	4	Look up	no	Lean forward, smile, acknowledge	Higher pitch	
Scold	19	4	look down	no	Lean forward, talk		
	20	4	frown	yes	acknowledge	Lower pitch	
	21	6	Look down, small grin	no	Lean forward, talk	giggle	"Volume would help"
	22	2	frown	yes	Pause, acknowledge	louder	
Soothe	23	4	Look up, eye contact	yes	Pause, acknowledge		
Scold	24	6	frown	yes	Pause, acknowledge		

Figure 6. Sample experiment session of a naive speaker, S3.

Figure 6 illustrates a sample event sequences that occurred during experiment sessions of a naive speaker. Each row represents a trial in which the subject attempts to communicate an affective intent to Kismet. For each trial, we recorded the number of utterances spoken, Kismet's cues, subject's responses and comments, as well as changes in prosody, if any.

2.3. Discussion

Recorded events show that subjects in the study made ready use of Kismet's expressive feedback to assess when the robot "understood" them. The robot's expressive repertoire is quite rich, including both facial expressions and shifts in body posture. The subjects varied in their sensitivity to the robot's expressive feedback, but all used facial expression, body posture, or a combination of both to determine when the utterance had been properly communicated to the robot. All subjects would reiterate their vocalizations with variations about a theme until they observed the appropriate change in facial expression. If the wrong facial expression appeared, they often used strongly exaggerated prosody to "correct" the "misunderstanding". In trial 20–22 of subject S3's experiment session, she giggled when kismet smiled despite her scolding, commented that volume would help, and thus spoke louder in the next trial. In general, the subjects used Kismet's expressive feedback to regulate their own behavior.

Kismet's expression through face and body posture becomes more intense as the activation level of the corresponding emotion process increases. For instance, small smiles verses large grins were often used to discern how "happy" the robot appeared. Small ear perks verses widened eyes with elevated ears and craning the neck forward were often used to discern growing levels of "interest" and "attention". The subjects could discern these intensity differences and several modulated their own speech to influence them. For example, in trials 1 and 2, Kismet responded to subject S3's praise by perking its ears and showing a small grin. In the next two trials the subject raised her pitch while praising Kismet to coax a stronger response. In trials 6–8 Kismet smiles broadly. We found that subjects often use Kismet's expressions to regulate their affective impact on the robot.

During course of the interaction, several interesting dynamic social phenomena arose. Often these occurred in the context of prohibiting the robot. For instance, several of the subjects reported experiencing a very strong emotional response immediately after "successfully" prohibiting the robot. In these cases, the robot's saddened face and body posture was enough to arouse a strong sense of empathy. The subject would often immediately stop and look to the experimenter with an anguished expression on her face, claiming to feel "terrible" or "guilty". In this emotional feedback cycle, the robot's own affective response to the subject's vocalizations evoked a strong and similar emotional response in the subject as well. This empathic response can be considered to be a form of entrainment.

Another interesting social dynamic we observed involved *affective mirroring* between robot and human. For instance, for another female subject (S2), she issued a medium strength prohibition to the robot, which caused it to dip its head. She responded by lowering her own head and reiterating the prohibition, this time a bit more foreboding. This caused the robot to dip its head even further and look more dejected. The cycle continues to increase in intensity until it bottoms out with both subject and robot having dramatic body postures and facial expressions that mirror the other. We see a similar pattern for subject S3 while issuing attentional bids. During trials 14–16 the subject mirrors the same alert posture as the robot. This technique was often employed to modulate the degree to which the strength of the message was "communicated" to the robot. This dynamic between robot and human is further evidence of entrainment.

3. Proto-Dialog

Achievement of adult-level conversation with a robot is a long term research goal. This involves overcoming challenges both with respect to the content of the exchange as well as to the delivery. The dynamics of turn-taking in adult conversation are flexible and robust. Well studied by discourse theorists, humans employ a variety of para-linguistic social cues, called *envelope displays*, to regulate the exchange of speaking turns [2]. Given that a robotic implementation is limited by perceptual, motor, and computational resources, could such cues be useful to regulate the turn-taking of humans and robots?

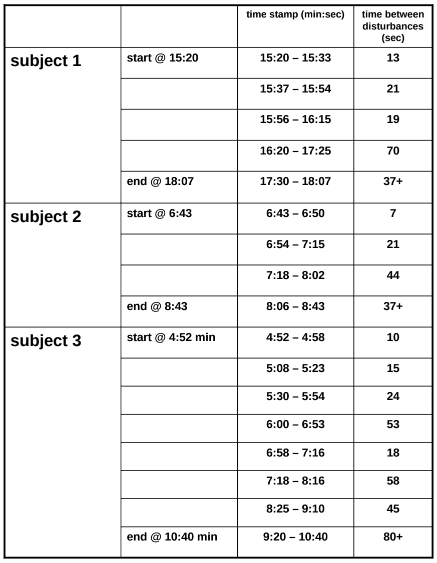
Kismet's turn-taking skills are supplemented with envelope displays as posited by discourse theorists. These paralinguistic social cues (such as raising of the brows at the end of a turn, or averting gaze at the start of a turn) are particularly important for Kismet because processing limitations force the robot to take-turns at a slower rate than is typical for human adults. However, humans seem to intuitively read Kismet's cues and use them to regulate the rate of exchange at a pace where both partners perform well.

3.1. Envelope Display Experiment

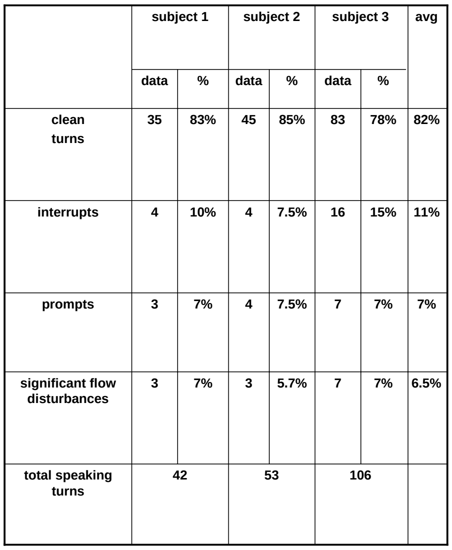
To investigate Kismet's turn-taking performance during proto-dialogs, we invited three naive subjects to interact with Kismet. Subjects ranged in age from 12 to 28 years old. Both male and female subjects participated. In each case, each subject was simply asked to carry a "play" conversation with the robot. The exchanges were video recorded for later analysis. The subjects were told that the robot did not speak or understand English, but would babble to them something like an infant.

Often the subjects begin the session by speaking longer phrases and only using the robot's vocal behavior to gauge their speaking turn. They also expect the robot to respond immediately after they finish talking. Within the first couple of exchanges, they may notice that the robot interrupts them, and they begin to adapt to Kismet's rate. They start to use shorter phrases, wait longer for the robot to respond, and more carefully watch the robot's turn taking cues. The robot prompts the other for their turn by craning its neck forward, raising its brows, and looking at the person's face when it's ready for them to speak. It will hold this posture for a few seconds until the person responds. Often, within a second of this display, the subject does so. The robot then leans back to a neutral posture, assumes a neutral expression, and tends to shift its gaze away from the person. This cue indicates that the robot is about to speak. The robot typically issues one utterance, but it may issue several. Nonetheless, as the exchange proceeds, the subjects tends to wait until prompted.

Before the subjects adapt their behavior to the robot's capabilities, the robot is more likely to interrupt them. There tend to be more frequent delays in the flow of "conversation" where the human prompts the robot again for a response. Often these "hiccups" in the flow appear in short clusters of mutual interruptions and pauses (often over 2 to 4 speaking turns) before the turns become coordinated and the flow smoothes out. However, by analyzing the video of these human-robot "conversations", there is evidence that people entrain



		time stamp (min:sec)	time between disturbances (sec)
subject 1	start @ 15:20	15:20 – 15:33	13
		15:37 – 15:54	21
		15:56 – 16:15	19
		16:20 – 17:25	70
	end @ 18:07	17:30 – 18:07	37+
subject 2	start @ 6:43	6:43 – 6:50	7
		6:54 – 7:15	21
		7:18 – 8:02	44
	end @ 8:43	8:06 – 8:43	37+
subject 3	start @ 4:52 min	4:52 – 4:58	10
		5:08 – 5:23	15
		5:30 – 5:54	24
		6:00 – 6:53	53
		6:58 – 7:16	18
		7:18 – 8:16	58
		8:25 – 9:10	45
	end @ 10:40 min	9:20 – 10:40	80+



	subject 1		subject 2		subject 3		avg
	data	%	data	%	data	%	
clean turns	35	83%	45	85%	83	78%	82%
interrupts	4	10%	4	7.5%	16	15%	11%
prompts	3	7%	4	7.5%	7	7%	7%
significant flow disturbances	3	7%	3	5.7%	7	7%	6.5%
total speaking turns	42		53		106		

Figure 7. The left table shows data illustrating evidence for entrainment of human to robot. The right table summarizes Kismet's turn taking performance during proto-dialog with three naive subjects. Significant disturbances are small clusters of pauses and interruptions between Kismet and the subject until turn-taking become coordinated again

to the robot (see the table to the left in figure 7). These "hiccups" become less frequent. The human and robot are able to carry on longer sequences of clean turn transitions. At this point the rate of vocal exchange is well matched to the robot's perceptual limitations. The vocal exchange is reasonably fluid. The table to the right in figure 7 shows that the robot is engaged in a smooth proto-dialog with the human partner the majority of the time (about 82%).

4. Conclusions

Experimental data from two distinct studies suggests that people do use the expressive cues of an anthropomorphic robot to improve the quality of interaction between them. Whether the subjects were communicating an affective intent to the robot, or engaging it in a play dialog, evidence for using the robot's expressive cues to regulate the interaction and to entrain to the robot were observed. This has the effect of improving the quality of the interaction as a whole. In the case of communicating affective intent, people used the robot's expressive displays to ensure the correct intent was understood to the appropriate intensity. In the case of proto-conversation, the subjects quickly used the robot's cues to regulate when they should exchange turns. As the result, the interaction becomes smoother over time with fewer interruptions or awkward pauses. These results signify that for social interactions with humans, expressive robotic faces are a benefit to both the robot and to the human who interacts with it.

5. Acknowledgements

Support for this research was provided by ONR and DARPA under MURI N00014–95–1–0600, by DARPA under contract DABT 63-99-1-0012, and by NTT.

References

[1] B. Reeves and C. Nass 1996, *The Media Equation.* CSLI Publications. Stanford, CA.

[2] J. Cassell 2000, "Nudge Nudge Wink Wink: Elements of face-to-face conversation for embodied conversational agents". In: J. Cassell, J. Sullivan, S. Prevost & E. Churchill (eds.) *Embodied Conversational Agents*, MIT Press, Cambridge, MA.

[3] F. Hara 1998, "Personality characterization of animate face robot through interactive communication with human". In: *Proceedings of IARP98.* Tsukuba, Japan. pp IV-1.

[4] H. Takanobu, A. Takanishi, S. Hirano, I. Kato, K. Sato, and T. Umetsu 1998, "Development of humanoid robot heads for natural human-robot communication". In: *Proceedings of HURO98.* Tokyo, Japan. pp 21–28.

[5] Y. Matsusaka and T. Kobayashi 1999, "Human interface of humanoid robot realizing group communication in real space". In: *Proceedings of HURO99.* Tokyo, Japan. pp. 188-193.

[6] P. Menzel and F. D'Alusio 2000, *Robosapiens.* MIT Press.

[7] A. Fernald 1985, "Four-month-old Infants Prefer to Listen to Motherese". In *Infant Behavior and Development, vol 8.* pp 181-195.

[8] Papousek, M., Papousek, H., Bornstein, M.H. 1985, The Naturalistic Vocal Environment of Young Infants: On the Significance of Homogeneity and Variability in Parental Speech. In: Field,T., Fox, N. (eds.) *Social Perception in Infants.* Ablex, Norwood NJ. 269–297.

[9] Ferrier, L.J. 1987, Intonation in Discourse: Talk Between 12-month-olds and Their Mothers. In: K. Nelson(Ed.) *Children's language, vol.5.* Erlbaum, Hillsdale NJ. 35–60.

[10] Stern, D.N., Spieker, S., MacKain, K. 1982, Intonation Contours as Signals in Maternal Speech to Prelinguistic Infants. *Developmental Psychology*, 18: 727-735.

[11] Vlassis, N., Likas, A. 1999, A Kurtosis-Based Dynamic Approach to Gaussian Mixture Modeling. In: *IEEE Trans. on Systems, Man, and Cybernetics. Part A: Systems and Humans*, Vol. 29: No.4.

[12] C. Breazeal & L. Aryananda 2000, "Recognition of Affective Communicative Intent in Robot-Directed Speech". In: *Proceedings of the 1st International Conference on Humanoid Robots (Humanoids 2000).* Cambridge, MA.

[13] C. Breazeal 2000, "Believability and Readability of Robot Faces". In: *Proceedings of the 8th International Symposium on Intelligent Robotic Systems (SIRS 2000).* Reading, UK, 247–256.