Cynthia Breazeal
MIT Media Lab
77 Massachusetts Ave NE18-5fl
Cambridge, MA 02139, USA
cynthiab@media.mit.edu

Regulation and Entrainment in Human–Robot Interaction

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. In this paper we explore 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. We present evidence for mutual regulation and entrainment of the interaction, and we discuss how this benefits the interaction as a whole.

KEY WORDS—human–robot interaction, sociable robots, emotion recognition, facial expression, vocal turn-taking

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. We 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 (Reeves and Nass 1996) and animated conversation agents (Cassell 2000). 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, Hara and Kobayashi (1997) at the Science University of Tokyo developed a realistic face robot that resembles a young Japanese woman—complete with silicone gel skin, teeth, and hair. 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 mimic them back to the person who displays them. In contrast, Takanobu and colleagues (1998) at Waseda University developed an expressive robot face that is more in the spirit of a mechanical cartoon. 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). Matsusaka and Kobayashi (1999) developed 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.

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 the far right picture 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; see Breazeal (2002) for an overview. These earlier experiments focused on having the robot regulate the intensity of the interaction by using familiar social cues, and thereby work with the human to establish an appropriate interaction where the robot is neither overwhelmed nor under stimulated. For instance, Breazeal et al. (2001) discuss the concept of *social amplification* by which the robot uses expressive displays to intuitively draw the human into an appropriate interaction

The International Journal of Robotics Research
Vol. 21, No. 10–11, October-November 2002, pp. 883-902,

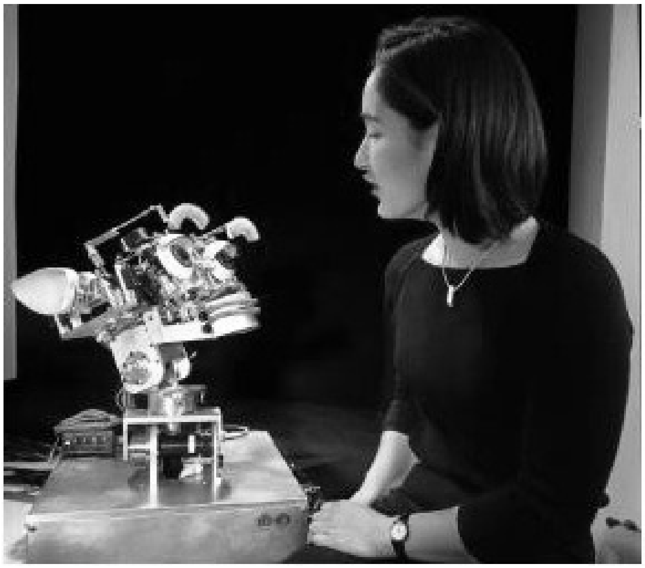

Fig. 1. A picture of our expressive robot, *Kismet*, developed at MIT.

distance that benefits the image processing limitations of the robot. Similarly, Breazeal (1998) describes the use of facial expressions (such as those analogous of fear, surprise, annoyance, and interest) to intuitively regulate how aggressively a person stimulates the robot with toys and gestures. As argued extensively in child-development literature, a caregiver diligently reads their child's expressive cues to maintain and establish a level of arousal that is suitable for learning, attention, and memory (Bullowa 1979). Given that our aim is to explore social learning scenarios akin to those that transpire between adults and very young children, the ability to establish and maintain a suitable learning environment for the robot is a critical skill.

In this paper we summarize our results with respect to two new 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 suggest that naive subjects naturally and intuitively read the robot's social cues and readily incorporate them into the exchange in interesting and beneficial ways. Specifically, during conversational turn-taking, they entrain (i.e., become more synchronized over time) to the tempo of Kismet's vocal turn-taking utterances. As a result, the number of interruptions (i.e., the human and robot speak at the same time), and awkward pauses (i.e., the robot or human misses a speaking turn) diminish over time. During the communication of affective intent we also see evidence of entrainment in body posture, head tilt, and facial expression. The subjects seem to exploit this affective synchrony to regulate the intensity of the robot's affective response to their praising, scolding, attention arousing, or soothing tones of voice. They also readily use the robot's expressive cues to determine if their affective intent was correctly communicated to the robot. Hence, it appears that entrainment and regulation are naturally exploited by people to improve their communicative efficacy with the robot in these two different studies.

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, mothers 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 (Fernald 1985).

Figure 2 illustrates these four prosodic contours. As shown, expressions of approval or praise, such as "That's a good bo-o-y!" 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 "No no, baby" are spoken with low pitch and high intensity in staccato pitch contours. Soothing tones such as "MMMM. Oh, honey" are low in both pitch and intensity, have a falling pitch contour, and are longer in duration. Finally, attentional bits such as "Can you get it?" tend to be higher in energy with a rising pitch to elicit attention and to encourage a response. Fernald (1985) 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.

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. A very detailed presentation of the recognizer and its performance assessment can be found in Breazeal and Aryananda (2001), so we present an abbreviated description here. 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. The focus of this paper is to explore the nature of the interaction that arises when a human tries to communicate different affects to the robot.

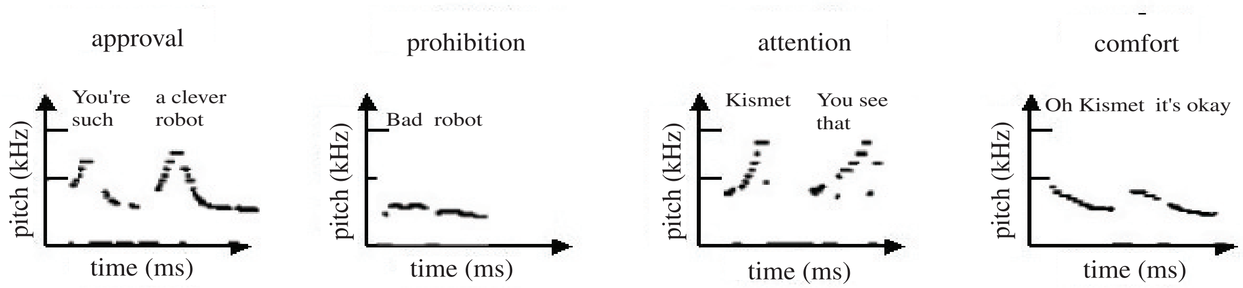

Fig. 2. Fernald's prototypical prosodic contours shown in robot directed speech for approval, attentional bid, prohibition, and soothing.

2.1. Recognition of Affective Intent

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. There were a total of 726 samples in the data set. Due to the limited set of training data, cross validation was applied for all classification processes (100 times per classifier).

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 (Vlassis and Likas 1999). The idea of the Gaussian mixture model is to represent the distribution of a data vector by a weighted mixture of component models, each one parametrized on its own set of parameters. Formally, the mixture density for the vector x assuming k components is

$$p(x) = \sum_{j=1}^{k} \pi_j f(x; \phi_j)$$

where $f(x; \phi_j)$ is the jth component model parametrized on ϕ_j, π_j are the mixing weights satisfying $\sum_{j=1}^{k} \pi_j = 1$, and $\pi_j \geq 0$.

In this algorithm, kurtosis is viewed as a measure of non-normality and is used to decide on the number of components in the Gaussian mixture problem. For a random vector x with mean m and covariance matrix S, the weighted kurtosis is defined as

$$\beta j = \sum_{i=1}^{n} P(j|x_i) \frac{((x_i - m_j)^T S_j^{-1}(x_i - m_j))^2}{\sum_{i=1}^{n} P(j|x_i)}.$$

Iteratively, EM steps are applied until convergence, and a new component is added dynamically until the test of normality $B = [\beta - d(d+2)]/\sqrt{[8d(d+2)]/n}$ indicates that $|B| \leq T$ for a predefined threshold, T.

Based on our recordings, the preprocessed pitch contours from the training set resemble Fernald's prototypical prosodic contours for approval, attention, prohibition, comfort/soothing, and neutral. Hence, we used Fernald's insights to select those features that would prove useful in distinguishing these five classes.

For the first classifier stage, global pitch and energy features (i.e., pitch mean and energy variance) partitioned the samples into useful intermediate classes. For instance, the prohibition samples are clustered in the low-pitch mean and high-energy variance region. The approval and attention classes form a cluster at the high-pitch mean and high-energy variance region. The soothing samples are clustered in the low-pitch mean and low-energy variance region. Finally, the neutral samples have low-pitch mean, but are divided into two regions in terms of their energy variance values. The structure of each of the mini-classifiers shown in Figure 3 follows logically from these observations. The features for each mini-classifier are outlined in Table 1.

The final classifier was evaluated using a new test set of 371 utterances from adult female speakers. Table 2 shows the resulting classification performance and compares it to the results from cross validation. The performance is reasonably high, and the failure modes are also reasonable. For instance, in those cases where the valence of an intent (i.e., positive verses negative affect) is misclassified, strongly valenced intents are misclassified as neutral rather than the opposite affect. All classes are sometimes misclassified as neutral. Approval and attentional bids are generally misclassified as one

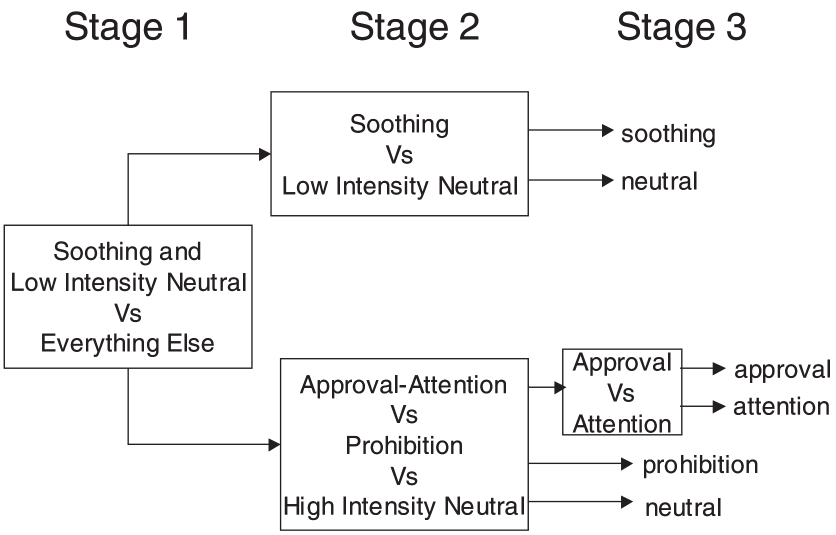
Stage 1
Stage 2
Stage 3
Soothing and Low Intensity Neutral Vs Everything Else
Soothing Vs Low Intensity Neutral
soothing
neutral
Approval-Attention Vs Prohibition Vs High Intensity Neutral
Approval Vs Attention
approval
attention
prohibition
neutral

Fig. 3. The classification stages.

Table 1. Features Used in Each Stage of the Multi-stage Classifier Model

Stage	Features	Cross Validation
Stage 1	Maximum energy Energy variance Energy range	93.6%
Stage 2A	Pitch segment average length Pitch segment minimum length Pitch contour slope Energy range Number of pitch segments	80.3%
Stage 2B	Pitch variance Energy variance Pitch mean	92.1%
Stage 3	Pitch variance Maximum rise–fall segment length	70.5%

or the other rather than mistaken for one of the other classes. This is not surprising given that so many of their prosodic features are shared in common. Overall, the performance of the system is encouraging.

2.2. Influencing the Robot's Affect

The output of the recognizer is integrated into the rest of Kismet's synthetic nervous system. As shown in Figure 4, the result of the classifier is passed to the robot's higher level perceptual system where it is combined with other contextual information in the form of perceptual releasers. As conceptualized in the field of ethology (Tingergen 1951), a releaser is a minimal set of perceptual features that have behavioral significance to the (robot) creature. In general, there are many different kinds of releasers defined for Kismet, each combining different contributions from a variety of perceptual and motivational systems. If the perceptual and motivational factors for a given releaser are active, then the output of the releaser can influence the rest of the system (e.g., contribute to the activation of an associated behavior in the behavior system).

The output of a releaser 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 bids enhance arousal (an alert expression), and soothing lowers arousal (a relaxed expression). The net

Table 2. Overall Classification Performance

Category	Test Size	Appr	Attn	Prohib	Comft	Ntrl	% Correct
Appr	84	64	15	0	5	0	76.2
Attn	77	21	55	0	0	1	74.3
Prohib	80	0	1	78	0	1	97.5
Comft	68	0	0	0	55	13	80.9
Ntrl	62	3	4	0	3	52	83.9
All	371						81.9

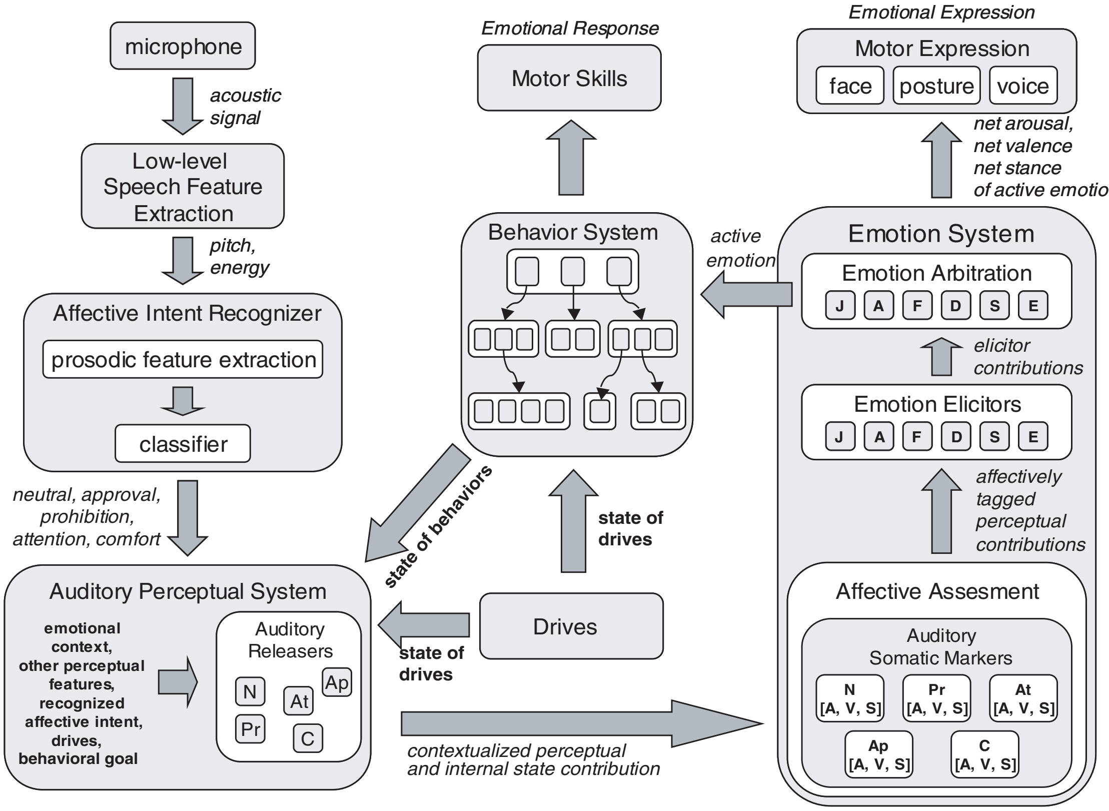

Fig. 4. System architecture for integrating vocal classifier input to Kismet's emotion system. In the higher level perceptual system, the classifier output is first combined with other contextual information (from the drives, past perceptual state and current behavioral goals) to form a set of auditory releasers. For the auditory releasers: N = neutral, Pr = prohibition, At = attention, Ap = approval, and C = comfort. These are marked with affective information in the affective assessment phase. The emotion elicitors use these affective tags to determine the most relevant emotion and activate it. In the emotion system: J represents "joy"; A, "anger"; F, "fear"; D, "disgust"; S, "sorrow"; and E, "excited/surprise". Once active, an emotion process can bias behavior and expression.

affective/arousal state of the robot is displayed on its face and expressed through body posture (Breazeal 2000), 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.

Within the emotion system, the output of each releaser must first pass through the affective assessment stage in order to influence emotional behavior. Within this assessment stage, each releaser is evaluated in affective terms and "tagged" with affective information (inspired by the somatic marker hypothesis proposed in Damasio (1994). There are three classes of tags that are used to affectively characterize its perceptual, motivational, and behavioral input. Each tag has an associated intensity that scales its contribution to the overall affective state. The *arousal* tag, A, specifies how energizing this percept is where positive values correspond to increasing arousal and negative values correspond to decreasing arousal. The *valence* tag, V, specifies how good or bad this percept is where positive values correspond to a pleasant stimulus and negative values correspond to an unpleasant stimulus. The *stance* tag, S, specifies how approachable the percept is where positive values correspond to advance whereas negative values correspond to retreat. Table 3 summarizes how each vocal affect releaser is somatically tagged. Because there are potentially many different kinds of factors that modulate the robot's affective state (e.g., behaviors, motivations, perceptions), this tagging process converts the myriad of factors into a common currency that can be combined to determine the net affective state.

For Kismet, the $[A, V, S]$ trio is the currency that the emotion system uses to determine which emotional response should be active. This occurs in two phases. First, all somatically marked inputs are passed to the *emotion elicitor* stage. Each emotion process has an elicitor associated with it that filters each of the incoming $[A, V, S]$ contributions. Only those contributions that satisfy the $[A, V, S]$ criteria for that emotion process are allowed to contribute to its activation. Within the *emotion arbitration* stage, the emotion processes compete for activation based on their activation level. There is an emotion process for each of Ekman's six basic emotions (Ekman 1992) corresponding to `joy`, `anger`, `disgust`, `fear`, `sorrow`, and `surprise`.

2.3. Displaying Expressive Feedback

If the activation level of the winning emotion process passes above threshold, it is allowed to influence the behavior system and the motor expression system through different pathways. By design, the expressive response leads the behavioral response of the robot. For instance, given that the caregiver makes an attentional bid, the robot will first exhibit an aroused and interested expression, then the orienting response ensues. By staging the response in this manner, the caregiver gets immediate expressive feedback that the robot understood her intent. The robot's expression also sets up the human's expectation of what behavior will soon follow, which gives the human a predictive cue as to what the robot is likely to do next. As a result, the human observing the robot can see its behavior, in addition to having an understanding of why the robot is behaving in that manner. In general, we have found emotive expression to be an important communication signal for the robot, lending richness to social interactions with humans and increasing people's level of engagement.

The emotive expression system is responsible for generating an expression that mirrors the robot's current affective state at the appropriate level of intensity. Kismet's facial expressions are generated using an interpolation-based technique over a three-dimensional affect space (see Figure 5). The current affective state (as determined by the emotion system) occupies a single point in this space at a time and moves within this space as the robot's affective state changes. There are nine *basis* postures that collectively span this space of emotive expressions as shown. Six of these postures sit at the extremes of each dimension and correspond to high arousal (i.e., surprise), low arousal (i.e., tired), negative valence (i.e., displeasure), positive valence (i.e., pleasure), open (i.e., accepting) stance, and closed (i.e., rejecting) stance. The remaining three postures are used to strongly distinguish the expressions for disgust, anger, and fear.

Expression, however, is not just conveyed through face, but through the entire body. Hence the expression system modifies the robot's body posture and speed of movement (Kismet moves more sluggishly when arousal is low, and in a more darting manner when highly aroused). There are six prototype body postures that also span the affect space. High arousal corresponds to an erect posture with a slight upward chin. Low arousal corresponds to a slouching posture where the neck lean and head tilt are lowered. The head tilts up a bit for positive valence, and down a bit for negative valence. An open stance corresponds to a forward lean movement, which suggests strong engagement. A closed stance corresponds to withdrawal, reminiscent of shrinking away from whatever the robot is looking at.

3. Entrainment During Affective Communication

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, comfort, 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.

Based on our implementation of the robot's emotion and expression systems, we can derive a list of quantifiable and

Table 3. Table Mapping [*A*, *V*, *S*] to Classified Affective Intents

Category	Arousal	Valence	Stance	Typical Expression
Approval	Medium high	High positive	Approach	Pleased
Prohibition	Low	High negative	Withdraw	Sad
Comfort	Low	Medium positive	Neutral	Content
Attention	High	Neutral	Approach	Interest
Neutral	Neutral	Neutral	Neutral	Calm

Note. Praise biases the robot to be “happy”, prohibition biases it to be “sad”, comfort evokes a “content, relaxed” state, and attention is “arousing”.

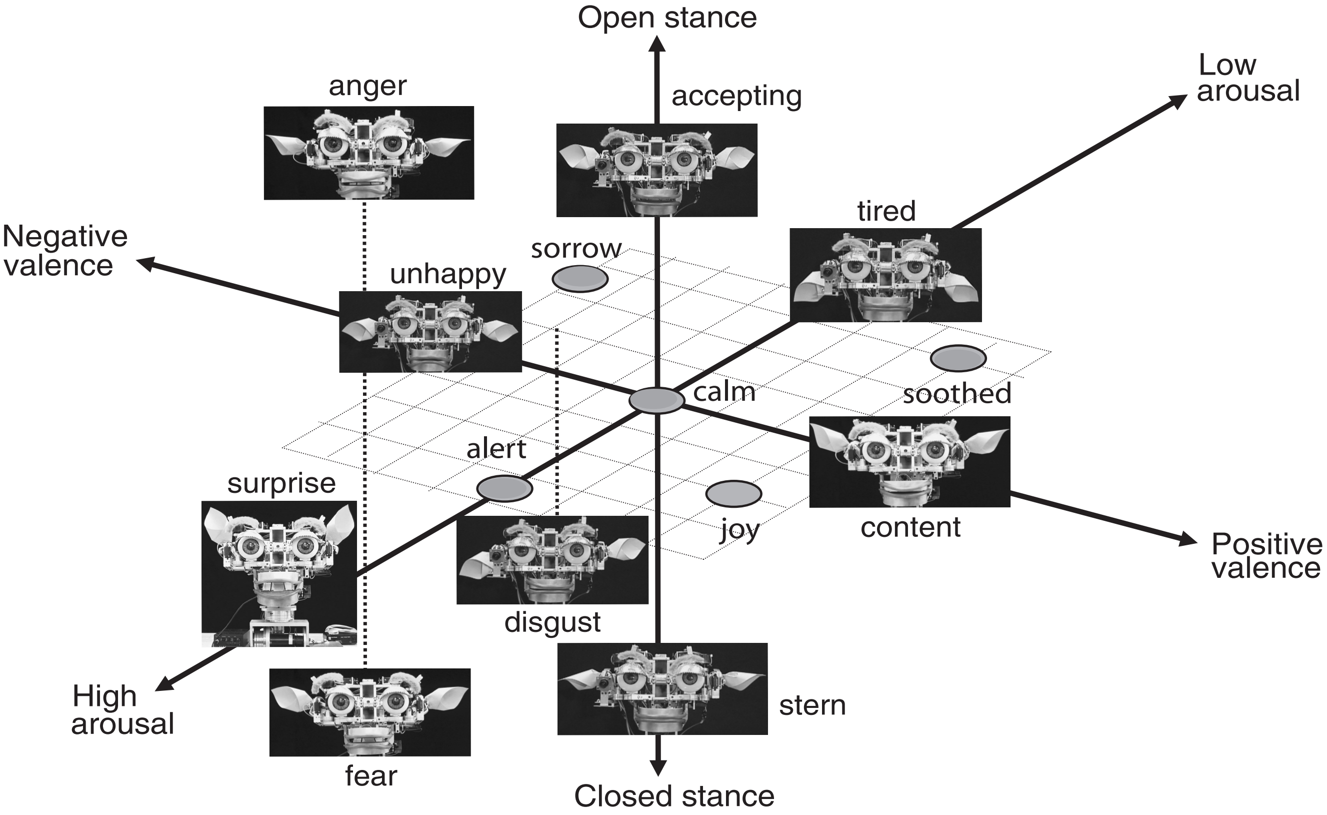

Fig. 5. The affect space consists of three dimensions. The extremes are: high arousal, low arousal, positive valence, negative valence, open stance, and closed stance. The emotional processes can be mapped to this space. The associated basis postures for expression are shown. They are blended using a localized weighted average based on the current affective state to generate the robot’s emotive facial expressions.

observable measures to assess the degree of entrainment, the communication of affective intent from human to robot, and the use of expressive feedback from robot to human. Since we are concerned with affective communication between the subject and the robot, these measures should be readily observable by the subject as well as by the experimenter. These are shown in Table 4. Video samples of Kismet recognizing these four classes of affective intent in different languages are presented in Extension 1.

The four tables provided in Appendix B (Tables 8–10) illustrate sample event sequences that occurred during experiment sessions with three naive speakers. 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, the subject's expressive cues, Kismet's expressive cues, and whether these cues were offered simultaneously along with the utterance, or in sequence.

Recorded events show that the 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, shifts in body posture and head pose, and changes in gaze direction as can be observed in Extension 1. The subjects varied in their sensitivity to the robot's expressive feedback, but all used facial expression, body posture, head pose, and gaze direction (or a combination of them) to determine when their intent had been properly communicated to the robot.

3.2. Themed Variations

All subjects 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 9–17 of Tables 8 and 9, the subject saw the robot frown after she issued a compliment (during trial 13) and recognized that this was an error (during trial 14). She immediately compensated in the following trial, adopting a very high and undulating pitch, with high energy. The robot responds in trial 15 with a positive and energetic expression. The subject then issues one more utterance with a strong praising intent, which serves to reinforce the robot's response and to strengthen it. Kismet smiles broadly. Another example can be seen in the third praising example of Extension 1. First, the subject issues an praising utterances to the robot, which it treats as an attentional bid. However, the subject quickly praises the robot again, and this time the robot understands correctly and responds by smiling. Once the subject is satisfied with the strength of Kismet's response, she acknowledges that she has been understood. Here, in general, the subjects used Kismet's expressive feedback to regulate their own behavior.

3.3. Sensitivity and Modulation of Intensity

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 versus large grins were often used to discern how "happy" the robot appeared. The "droopiness" of the ears, bowing of the head, and the downward curvature of the lips were used to discern whether or not the robot was sufficiently reprimanded. Perked ears versus 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 and expressive cues to influence them (e.g., trials 13–15 of the praising experiment above). Evidence of this can be observed during trials 10–12 and 13–15 of the prohibition experiment as shown in Table 10. In both cases, the robot responds to the first prohibitive utterance by bowing its head and drooping its ears. Although the subject responds to this, she issues another prohibition that strengthens the robot's expression: the ears become crestfallen, the robot's gaze is averted, and a frown appears on its face. At this point, she acknowledges that the robot understood her, although she does not always require this intensity of response. An example of this can be seen in the first scolding sample presented in Extension 1. In a similar example, the robot responds to the first scolding by lowering its head. However, as the subject continues to scold the robot, it intensifies its response. In general, we found that subjects often use Kismet's expressions to regulate their affective impact on the robot.

3.4. Empathic Reactions

During course of the interaction, subjects often displayed empathetic responses. 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". An example of this can be seen in the second scolding sample of Extension 1 after the subject successfully scolds the robot in German. Similarly, in trial 11 of Table 10, the subject apologizes to the robot after eliciting a strong reaction from scolding it. 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.

3.5. Affective Mirroring

Another interesting social dynamic we observed involved *affective mirroring* between robot and human. This is another

Table 4. Observable Expressive Measurements Demonstrated by Human and/or Robot During Affective Intent Experiments

Observable Measures for Communication of Affective Intent		
Cue	Reading	Annotation
Utterance	*utterance*	"*utter*"
Prosody	*pitch, energy, tempo*	Pr:
Body Posture	*neutral, erect, forward, away*	Bd:
Head Tilt	*neutral, up, down*	Hd:
Gaze Direction	*eye contact, glance/stare-down, glance/stare-up, glance/stare-right, glance/stare-left*	Gz:
Facial Expr	*neutral, relax, happy, sad alert, comforting, other*	Fc:
Ear Pose	*neutral, perk up, droop, fallen*	Er:
Lip Shape	*neutral, rounded, smile, frown*	Lp:
Acknowledge		*ack*
Sequential (across turns)		$\Longrightarrow, \Longleftarrow$
Sequential (within turn)		$\rightarrow$
Simultaneous		$\Longleftrightarrow$

form of entrainment that we observed frequently (and can be seen throughout the samples in Extension 1). For instance, during the prohibition experiment, we see the subject issue a medium strength prohibition to the robot, which causes it to dip its head. She responds by lowering her own head and reiterating the prohibition, this time a bit more foreboding. This causes 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. During the praising experiment, such as trial 16 with Subject *A* or trial 7 with Subject *B* in Table 9, the subject mirrors the same uplifted head pose, body lean, and facial expression as the robot (and vice versa). In the first part of trial 5 with Subject *B*, Kismet's forward body lean and smiling expression follows the subject's body posture and face. However, in the second part of the trial the subject sits back after following Kismet's lead.

3.6. Summary

In these studies, we found that the subjects relied on the robot's expressive feedback to determine whether or not the robot understood them. They also used the robot's expressive feedback to gauge their next response—to re-issue the same intent or not, and to issue it with an appropriate degree of intensity. During the exchanges, the continual small adjustments in body posture, gaze direction, head pose, and facial expression served to synchronize the movements of the robot and the human, bringing both into a similar expressive state. Whether consciously employed or not, this entrainment-based technique was often used by the human to modulate the degree to which the strength of her message was "communicated" to the robot, and to bring the robot and the human into a state of affective synchrony.

4. Entrainment During Proto-Dialogs

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 is flexible and robust. Well studied by discourse theorists, humans employ a variety of paralinguistic social cues, called *envelope displays*, to regulate the exchange of speaking turns (Cassell 2000). 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 one's brows and establishing eye contact to relinquish one's speaking turn, or looking aside to hold one's speaking turn even when speech is paused) are particularly important for Kismet because processing limitations force the robot to exchange speaking turns at a slower rate than is typical for human adults (for humans, this takes place after a 0.25 s pause once speech has ended. However, Kismet does so after a minimum of a 0.5 s pause). However, humans seem to intuitively read Kismet's cues and use them to regulate the rate of exchange at a pace where both

Table 5. Annotations for Proto-dialog Experiments

Annotations for Proto-dialog Experiment		
Type	Option	Annotation
Listener	Human	H
Speaker	Robot	R
Turn Phase	Acquire Floor	Aq
	Start Speech	St
	Stop Speech	Sp
	Hold Floor	Hd
	Relinquish Floor	Rq
Cue	Avert gaze	
	Eye contact	
	Elevate brows	
	Lean forward	
	Lean back	
	Blink	
	"*Utterance*"	
Turns	Clean turn	#
	Interrupt	I
	Missed	M
	Pause	P

partners perform well. Kismet's envelope displays are summarized below. To avoid a "canned" performance, Kismet does not exhibit displays according to a rigid schedule, but rather applies them more as a rule-of-thumb. These envelope displays can be observed in Extension 2 where the robot engages in a proto-dialog with two people. Note that we refer to this vocal exchange as a proto-dialog because although the human's utterances are spoken in natural language (i.e., in English), the robot uses a Kismet-esque babble for its speaking turn. Hence, the video is intended to demonstrate the use of envelope displays to regulate the dynamics of interaction during the exchange of speaking turns, rather than focus on the content of what is said.

- To acquire the floor: break eye contact and/or lean back a bit.
- To start its speaking turn: vocalize a Kismet-esque babble.
- To stop its speaking turn: stop vocalizing and re-establish eye contact. Blinking tends to occur at the end of a vocalization.
- To hold the floor: look to the side.
- To relinquish the floor: raise brows and/or lean forward a bit.

To investigate Kismet's turn-taking performance during proto-dialogs, we invited four naive subjects to interact with Kismet. Subjects ranged in age from 12 to 28 years old. Two male and two female subjects participated. In each case, the subject was simply asked to carry out a "play" conversation with the robot. The exchanges were video recorded for later analysis and annotated according to Table 5. The subjects were told that the robot does not speak nor understand English, but babbles in a characteristic manner. The proto-dialog carried out by Subject 3 (a female subject) is presented in Appendix B in Tables 11–14. The time codes are those that appear on the video tape used to record the sessions. A turn is defined with respect to the speaker who holds the floor and consists of four phases: acquire the floor, start the utterance, end the utterance, and relinquish the floor. The speaker may also hold the floor (i.e., maintain their speaking role during a silent pause).

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. Before the subjects adapt their behavior to the robot's capabilities, the robot is more likely to interrupt them. For instance, it is often the case that the robot interrupts them within the first couple of exchanges (e.g., turns 4 or 5 in Table 11). In general, there tends 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 two to four utterances of the speaker) before the turn phases become coordinated and the flow of the exchange of speaking turns smoothes out. We call these clusters *significant flow disturbances*. In Extension 3, we see what happens when a person tries to dominate the proto-dialog. The robot has a limit on the length of time a person

Table 6. Data Illustrating Evidence for Entrainment of Human to Robot

		Time Stamp (min:s)	Clean Turns Between Disturbances (s)
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 6:47	6:47–6:54	3
		6:55–7:21	7
		7:22–7:57	11
	end 8:44	8:03–8:44	16
Subject 4	start 4:52	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	9:20–10:40	80+

As time progresses there are increasing number of clean turns before a “hiccup” in the flow occurs.

can speak before the robot’s speech buffer fills up. When this happens, the robot immediately processes the utterance and responds. As can be seen in the video, this causes the robot to interrupt the person. Note, however, that the robot’s visual behavior keeps the interaction lively, and clearly signals that the robot is attending to and is interested in the person’s behavior. After a few utterances, the person starts using short phrases with longer pauses between them, and a smooth exchange of turns resumes.

However, by analyzing the video of these human–robot “conversations”, there is evidence that people entrain to the robot (see Table 6 and those in Appendix B). They often start to use shorter phrases, wait longer for the robot to respond, and more carefully watch the robot’s turn-taking cues. For instance, the robot prompts the person to take their speaking turn by either craning its neck forward, raising its brows, or establishing eye contact when it is ready for them to speak (e.g., turns 1, 6, or 8). It will hold this posture for a few seconds until the person responds. Often, within a second of this display, the subject does so. When the subject stops speaking, Kismet tends to lean back to a neutral posture, assume a neutral expression, and perhaps shift its gaze away from the person (e.g., turns 1, 3, or 10). 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 are more likely to wait until prompted by the relinquish turn display.

As the subjects seem to adjust their behavior according to Kismet’s envelope displays, these “hiccups” within speaking turns become less frequent. As can be seen in Table 6, for each subject there are progressively longer runs of cleanly exchanged turns as time progresses. This suggests that the flow of communication becomes smoother (e.g., fewer interruptions, pauses, and significant flow disturbances) as people read and entrain to Kismet’s envelope displays. At this point the rate of vocal exchange is well matched to the robot’s perceptual limitations. The table to the right in Figure 6 shows that the robot is engaged in a smooth proto-dialog with the human partner the majority of the time (about 82.5%).

5. Conclusions

Experimental data from two distinct studies suggest 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

Table 7. Kismet's Turn-taking Performance During Proto-dialog with Four Naive Subjects

	Sub 1		Sub 2		Sub 3		Sub 4		Average
	Data	%	Data	%	Data	%	Data	%	%
Clean Turns	35	83	45	85	38	84	83	78	82.5
Interrupts	4	10	4	7.5	5	11	16	15	10.9
Pauses	3	7	4	7.5	2	4	7	7	6.3
Significant Flow Distrb.	3	7	3	5.7	2	4	7	7	6
Total Speaking Turns	42		53		45		106		

Note. Significant disturbances are small clusters of pauses and interruptions between Kismet and the subject until turn-taking becomes coordinated again.

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. They also used the expressive displays to regulate the intensity of the robot's affective response. In the case of proto-conversation, the subjects quickly used the robot's cues to regulate when they should exchange turns. As a 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.

Appendix A: Index to Multimedia Extensions

The multimedia extension page is found at http://www.ijrr.org.

Table of Multimedia Extensions

Extension	Type	Description
1	Video	Samples of Kismet recognizing praise, prohibition, attentional bids, and soothing utterances through prosody.
2	Video	Kismet engaging in a proto-dialog with two people.
3	Video	Kismet engaging in a proto-dialog with one person.

Appendix B

To provide a detailed glimpse into social exchanges with Kismet, in this appendix we present a few of the annotated video sessions. Tables 8–10 highlight a portion of two of the praising experiments with Subjects *A* and *B* and one of the scolding experiments with Subject *C* (see Section 3 for a discussion and Table 4 for a description of the annotations). Similarly, Tables 11–14 annotate a segment of the proto-dialog experiment with Subject 3 (see Section 4 for a discussion and Table 5 for a description of the annotations).

Acknowledgments

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.

Table 8. Expressive Exchanges Between the Robot and Subject *A* While Communicating a Praising Affective Intent

			Praising Intent Issued by Subject A		
Trial	Utterance		Human		Robot
1	"Great job Kismet"	⟺	Bd:fwd, Fc:happy, Hd:up Gz:eye-ct	⟺	Er:perk-up, Fc:neu, Hd/Bd:neu Gz:eye-ct
2	"Way to go"			⟹	Gz:glance-up, Er:perk-up
3	"You're a smart robot"		Fc:happy	⟹	Gz:eye-ct, Lp:grin → Fc:relax → Fc:happy
4	*ack*				
5	"You're a cute robot"	⟺	Bd:fwd, Fc:happy, Hd:up Gz:eye-ct	⟹	Lp:curve-up Bd/Hd:neu Gz:eye-ct
	(*small gasp*)	⟺	Bd:erect	⟸	
6	"You're so smart"	⟺	Bd:fwd, Lp:smile Hd:up	⟹	Lp:curve-up, Gz:glance-up Hd:up
7	"What beautiful eyes"	⟺	Fc:happy Hd:up	⟺	Fc:happy, Hd:up
8	*ack*				
9	"Good job"	⟺	Bd:fwd Hd:up, Gz:eye-ct Fc:happy	⟹	Fc:neu, Hd/Bd:neu, Gz:look-right
10	"Good job"	⟹	Gz:stare-down		
11	"That was ingenious"	⟺	Bd:far-fwd, Gz:eye-ct	⟺	Bd:fwd Gz:eye-ct
12	"What are you looking at? Great"	⟺	Body:fwd Hd:up	⟺	Head:up Er:perk-up
			Bd:sit-back	⟹	Gz:eye-ct
13	"Who's the pretty robot"	⟺	Bd:fwd	⟹	Fc:sad, Hd:down
14	"Oh no"	⟺	Bd:sit-back, Pr:soft,low Fc:neu	⟸	
15	"You're great"	⟺	Pr:excited,high, Bd:far-fwd, Hd:up	⟹	Hd:up, Lp:smile Er:perk-up
			Bd:sit-back	⟸	

Table 9. Continuation of the Praising Intent Experiment with Subject *A*, Followed By a Praising Intent Session with Subject *B*, Then Followed By a Scolding Intent Session with Subject *C*

Praising Intent Issued by Subject A...continued					
Trial	Utterance		Human		Robot
16	“You make me laugh”	⟺	Hd/Bd:neu, Lp:smile	⟹	Fc:happy Hd:up
			Hd:up, Fc:happy Bd:fwd	⟸	
17	*ack*				
Praising Intent Issued by Subject B					
1	“You’re so cute”	⟺	Bd:fwd Fc:happy	⟺	Lp:rounded, Er:perk-up → Fc:happy
			Bd:neu	⟸	Bd:neu
2	*ack*				
3	“That’s a good job”	⟺	Hd/Bd:neu Fc:happy	⟹	Er:perk, Hd:up Fc:happy
4	*ack*				
5	“Kismet, great”	⟺	Bd:fwd, Fc:happy	⟹	Bd:fwd Fc:happy→
			Bd:sit-back	⟸	Bd:sit-back
6	*ack*				
7	“You’re fast at this “	⟺	Head/Body:neu, Lp:smile	⟹	Er:perk-up, Lp:smile, Body:neut, Hd:up
			Hd:up	⟸	
8	*ack*				
9	“You seem alive”	⟺	Body:neu Hd:up Fc:smile	⟹	Bd/hd:neu Fc:happy
			Fc:happy	⟸	
10	*ack*				
Prohibition Intent Issued by Subject C					
1	“Kismet”	⟺	Pr:deep,loud, Bd/Hd:neu, Fc:stern, Gz:eye-ct	⟺	Bd/Hd:neu, Fc:neu→ Bd:erect
2	“Kismet what did you do”	⟺	Pr:deeper, Hd:down	⟺	Hd:down, Gz:glance-dwn

Table 10. Continuation of the Scolding Intent Session with Subject *C*

Trial	Utterance		Human		Robot
3	“Bad robot”	⟺	Bd:fwd, Hd:far-dwn	⟹	Hd:far-dwn, Gz:hold-dwn, Bd:fwd
4	*ack*				
5	“Bad job”	⟺	Fc:stern, Hd:dwn, Gz:eye-ct, Bd:neu,	⟹	Bd:lean-back, Fc/Hd:neu, Gz:eye-ct
			Bd:sit-back	⟸	
6	“Don’t take over the humans”			⟹	Er:drop, Hd:dwn
7	*ack*				
8	“No Kismet”	⟺	Pr:loud,deep	⟹	Hd:dwn, Gz:dwn, Er:drop → Fc:sad
9	*ack*				
10	“That was so bad of you”	⟺	Bd:fwd, Fc:stern, Gz:eye-ct	⟺	Hd:bit-dwn, Gz:eye-ct, Er:bit-dwn
11	“Did you make potty on the rug”	⟺	Bd:far-fwd, Hd:dwn, Pr:deeper,louder	⟺	Gz:dwn, Hd:dwn→ Ers:fallen,
	(“Oh sorry”)			⟸	Face:sad
12	*ack*				
13	“Kismet”	⟺	Face:stern, Hd:dwn	⟺	Gz:eye-ct, Hd:dwn, Er:drop
14	“Bad robot”	⟺	Hd:dwn, Pr:deeper, Hd:dwn	⟺	Gz:dwn, Fc:sad, Hd:dwn
15	*ack*				
16	“Dont do that to me”	⟺	Hd:dwn	⟹	Hd:dwn
17	“Dont look at me like that”	⟺	Pr:deeper	⟹	Hd:up→ Hd:dwn
			Hd:dwn	⟸	Gz:dwn, Er:drop
	ack				

Table 11. Envelope Displays During a Proto-dialog Exchange Between Kismet and a Human Subject

Time	Speaker			Listener		
Code	S	Ph	Cue	L	Cue	Turns
06:47:20	R	Aq	avert gaze	H	eye contact	1
		St	*babble*			
06:48:14		Sp	eye contact			
06:50:01		Rq	elev brows, lean forward			
06:50:04	H	Aq	eye contact	R	eye contact	2
		St	“What’d you say?”	R	elev brows, lean forward	
06:50:10		Sp:Rq				
06:51:23	R	Aq	avert gaze	H	eye contact	3
			lean back			
06:52:23		Sp	*babble*			
06:53:19		St	eye contact			
			blink			
06:54:10		Rq	elev. brows			
06:55:17	H	Aq	eye contact	R	eye contact	4
		St	“My name is Carol.”		*babble*	I
06:56:16		Sp				
06:58:02		St	eye contact		eye contact	5
			“My name is Carol!”		*babble*	I
06:58:30		Sp				
07:00:29		St	“Carol.”			
07:01:12		Sp:Rq				
07:01:21	R	Aq	eye contact	H	eye contact	6
		St	*babble*			
07:02:17		Sp	eye contact			
07:02:45		Rq	elev brows			
07:03:01	H	Aq:St	“Carol.”	R	eye contact	7
07:03:09		Sp:Rq				
07:04:17	R	Aq	avert gaze	H	eye contact	8
07:05:04		St	*babble*			
07:06:00		Sp	eye contact			
07:07:01		Rq	lean forward, elev. brows			
07:07:09	H	Aq:St	“Carol!”	R	eye contact	9
07:07:19		Sp:Rq				
07:09:03	R	Aq	lean back	H	eye contact	10
07:10:24		St	*babble*			
07:11:23		Sp	eye contact			
07:12:05		Rq	lean forward			

Table 12. Continued...Envelope Displays During a Proto-dialog Exchange Between Kismet and a Human Subject

Time	Speaker			Listener		
Code	S	Ph	Cue	L	Cue	Turns
07:13:05	H	Aq	eye contact	R	eye contact	11
		St	"Did you ask me how I am? I'm fine. How are you?"			
07:14:25		Sp:Rq				
07:17:09	R	Aq	avert gaze	H		12
07:17:10		St	*babble*			
07:18:03		Sp	eye contact			
07:20:05		Hd	avert gaze			
07:21:24		Rq	eye contact raise brows			
07:22:23	H	Aq		R	eye contact	13
		St	"Are you speaking another language, Kismet?"	*babble*		I
07:24:23		Sp:Rq				14
07:24:06	R	Aq:St	*babble*	H		15
07:25:04		Sp	blink			
		Rq	elev brows			
07:25:14	H	Aq:St	"Sounds like you're speaking Chinese."	R	eye contact	16
07 27:10		St:Rq				
07:27:20	R	Aq	lean forward	H		17
07:27:45		St	*babble*			
07:28:03		Sp	eye contact			
07:28:25		Rq	elev brows			
07:30:08	H	Aq:St	"Hey!"	R	avert gaze	18
07:30:15		Sp:Rq	lean forward		eye contact	
07:31:08	R	Aq:St	*babble*	H	eye contact	19
07:33:01		Sp	blink eye contact			
07:33:30		Rq	elev brows			
07:34:01	H	Aq:St	"What are you saying?"	R	eye contact	20
07:34:26		Sp:Rq				
07:36:04	R	Aq:St	*babble*	H	eye contact	21
07:37:00		Sp	blink			
07:38:19		Rq	lean forward, elev brows, eye contact		lean forward nod head	
07:40:00		Aq	lean back, avert gaze			

Table 13. Continued...Envelope Displays During a Proto-dialog Exchange Between Kismet and a Human Subject

Time	Speaker			Listener		
Code	S	Ph	Cue	L	Cue	Turns
07:41:13		St	*babble*			
07:42:11		Sp:Rq	eye contact			
07:45:05	H	Aq		R	eye contact	22
		St	“Did you know that you look like a gremlin?”			
07:47:05		Sp:Rq				
07:47:26	R	Aq	avert gaze	H	eye contact	23
07:49:12		St	*babble*			
07:50:25		Sp:Rq	eye contact			
07:52:22	H	Aq:St	“All right...”	R	eye contact,	24
07:53:05		Sp			eye contact	
07:54:18		St	“What are you going to do the rest of the day?”			
07:55:29		Sp:Rq				
07:56:14	R	Aq:St	*babble*	H	eye contact	25
07:57:29		Sp:Rq	blink		avert gaze	
			eye contact			
08:03:01	H	Aq:St	“My name is Carol. You have to remember that I’m Carol. ”	R	eye contact *babble*	26 I
08:05:25		Sp:Rq	(pause)			P
08:06:31		St	“If you see me again, I’m Carol.”		eye contact	27
08:07:17		Sp:Rq	(pause)			P
08:08:26		St	“Hello!”			28
08:09:21		Sp			blink	
		Rq	lean forward			
08:10:13	R	Aq	avert gaze	H	lean back (laugh)	29
08:10:40		St	*babble*			
08:11:17		Sp	eye contact, blink			
08:11:45		Rq	lean forward			
08:12:19	H	Aq:St	“Hello!”	R		30
08:12:54		Sp:Rq				
08:13:23	R	Aq:St	*babble*	H		31
08:14:25		St:Rq	eye contact, elev brows			
08:15:05	H	Aq:St	“Hello!”	R		32
08:15:35		St:Rq				

Table 14. Continued...Envelope Displays During a Proto-dialog Exchange Between Kismet and a Human Subject

Time	Speaker			Listener		
Code	S	Ph	Cue	L	Cue	Turns
08:16:10	R	Aq	lean back	H		33
08:16:31		St	*babble*			
08:17:05		Sp	eye contact			
		Rq	elev brows			
08:17:55	H	Aq:St	“Hello.”	R	eye contact	34
08:18:15		Sp:Rq				
08:18:20	R	Aq:St	*babble*	H	eye contact	35
08:19:10		Sp	eye contact			
08:19:46		Rq	lean forward			
08:20:26	H	Aq:St	“Are we having a conversation?”	R	eye contact	36
08:21:23		Sp:Rq				
08:22:17	R	Aq	avert gaze	H	eye contact	37
		St	*babble*			
08:23:00		Sp	eye contact			
08:23:26		Rq	blink			
08:24:12	H	Aq:St	“Is that right?”	R	eye contact	38
08:24:35		Sp:Rq				
08:24:45	R	Aq:St	*babble*	H	eye contact	39
08:25:01		Sp	blink			
		Rq	elev brows			
08:25:10	H	Aq:St	That’s right!”	R	eye contact	40
08:26:11		Sp:Rq	nod (laugh)			
08:29:08	R	Aq:St	*babble*	H	(laughing)	41
08:30:07		Sp	eye contact			
08:32:03		Hd	lean back			
08:33:13		Hd	avert gaze			
08:35:17		Hd	eye contact			
08:36:10		St	*babble*			
08:37:08		Sp	blink			
08:37:48		Rq	lean forward, elev brows			
08:38:00	H	Aq:St	“Is that right?”	R	eye contact	42
08:38:15		Sp:Rq				
08:38:20	R	Aq	(pause)	H	eye contact	43
		Rq	lean forward elev brows			
08:39:03	H	Aq:St	“Is that right!”	R	eye contact	44
08:39:06		Sp:Rq				
08:39:26	R	Aq:St	*babble*	H	“and then...”	45
						I
08:40:23		Sp:Rq	lean forward			

References

Breazeal, C. 1998. A motivational system for regulating human–robot interaction. In *Proc. 15th National Conference on Artificial Intelligence (AAAI98)*, Madison, WI, pp. 54–61.

Breazeal, C. 2000. Believability and Readability of Robot Faces. In *Proc. 8th Int. Symp. on Intelligent Robotic Systems (SIRS 2000)*, Reading, UK, pp. 247–256.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Breazeal, C., and Aryananda, L. 2001. Recognizing affective intent in robot directed speech. *Autonomous Robots* 12(1):83–104.

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. 2001. Active vision for sociable robots. *IEEE Trans. Syst. Man Cybern.* 31(5):443–453.

Bullowa, M., ed. 1979. *Before Speech: The Beginning of Interpersonal Communication*. Cambridge, UK: Cambridge University Press.

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

Damasio, A. 1994. *Descartes Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam's Sons.

Ekman, P. 1992. Are there basic emotions? *Psychological Review* 99(3):550–553.

Fernald, A. 1985. Four-month-old Infants Prefer to Listen to Motherese. In *Infant Behavior and Development* Vol. 8, pp. 181–195.

Hara, F. and Kobayashi, H. 1997. State of the art in component technology for an animated face robot: its component technology development for interactive communication with humans. *Advanced Robotics* 11(66):585–604.

Matsusaka, Y., and Kobayashi, T. 1999. Human interface of humanoid robot realizing group communication in real space. *Proc. HURO99*, Tokyo, Japan, pp. 188–193.

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

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

Tingergen, N. 1951. *The Study of Instinct*. New York: Oxford University Press.

Vlassis, N., and Likas, A. 1999. A Kurtosis-Based Dynamic Approach to Gaussian Mixture Modeling. *IEEE Trans. Syst. Man Cybern.* 29(4):393–399.