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 [5]. 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

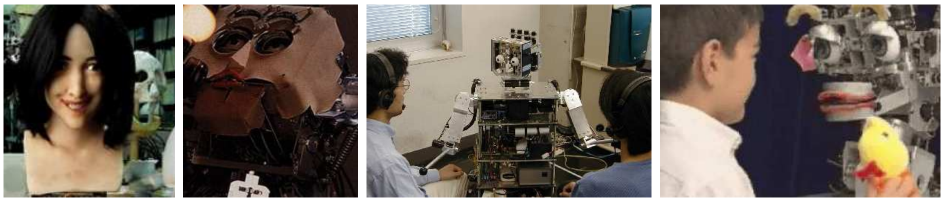

Figure 1. A sampling of robots designed to interact with people. The far left picture shows a realistic face robot designed at the Science University of Tokyo. The middle left picture shows *WE-3RII*, an expressive face robot developed at Waseda University. The middle right picture shows *Robita*, an 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 [8].

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 [6]. 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 [7].

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 [9]. Figure 2 illustrates these four prosodic contours.

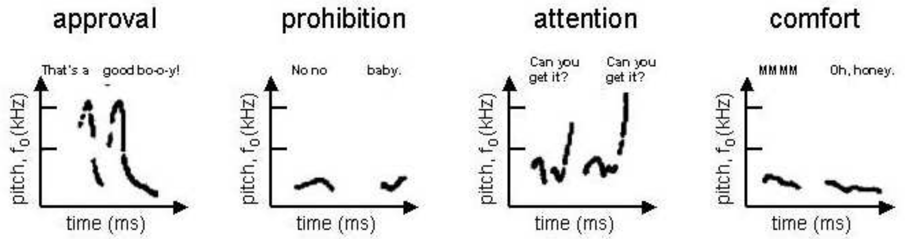

Figure 2. Fernald's prototypical prosodic contours for approval, attentional bid, prohibition, and soothing.

Mothers are more likely to use falling pitch contours than rising pitch contours when soothing a distressed infant [10], to use rising contours to elicit attention and to encourage a response [11], and to use bell shaped contours to maintain attention once it has been established [12]. 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 [9].

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

As shown in figure 3, the affective speech recognizer receives robot-directed speech as input. The speech signal is analyzed by the low-level speech processing system, producing time-stamped pitch (Hz), percent periodicity (a measure of how likely a frame is a voiced segment), energy (dB), and phoneme values[1]

[1]This auditory processing code is provided by the Spoken Language Systems Group at MIT. For now, the phoneme information is not used in the recognizer.

in real-time. The next module performs filtering and pre-processing to reduce the amount of noise in the data. The pitch value of a frame is simply set to 0 if the corresponding percent periodicity indicates that the frame is more likely to correspond to unvoiced speech. The resulting pitch and energy data are then passed through the feature extractor, which calculates a set of selected features (F_1 to F_n). Finally, based on the trained model, the classifier determines whether the computed features are derived from an approval, an attentional bid, a prohibition, soothing speech, or a neutral utterance.

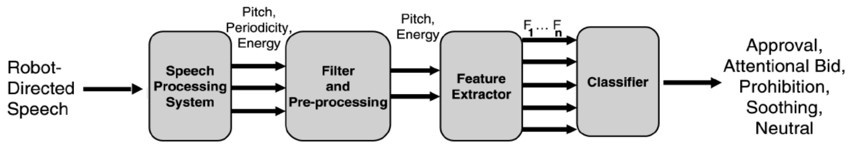

Figure 3. The spoken affective intent recognizer.

2.1.1. Training the System

Two female adults who frequently interact with Kismet as caregivers were recorded. The speakers were asked to express all five affective intents (approval, attentional bid, prohibition, comfort, and neutral) during the interaction. Recordings were made using a wireless microphone, and the output signal was sent to the low-level 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 real-time pitch, percent periodicity, energy, and phoneme values. All recordings were performed in Kismet's usual environment to minimize variability of environment-specific noise. Samples containing extremely loud noises (door slams, etc.) were eliminated, and the remaining data set were labeled according to the speakers' affective intents during the interaction. There were a total of 726 utterances in the final data set — approximately 145 utterances per class.

2.1.2. Data Preprocessing

The pitch value of a frame was set to 0 if the corresponding percent periodicity was lower than a threshold value. This indicates that the frame is more likely to correspond to unvoiced speech. Even after this procedure, observation of the resulting pitch contours still indicated the presence of substantial noise. Specifically, a significant number of errors were discovered in the high pitch value region (above 500 Hz). Therefore, additional preprocessing was performed on all pitch data. For each pitch contour, a histogram of ten regions was constructed. Using the heuristic that the pitch contour was relatively smooth, it was determined that if only a few pitch values were located in the high region while the rest were much lower (and none resided in between), then the high values were likely to be noise. Note that this process did not eliminate high but smooth pitch contour since pitch values would be distributed evenly across

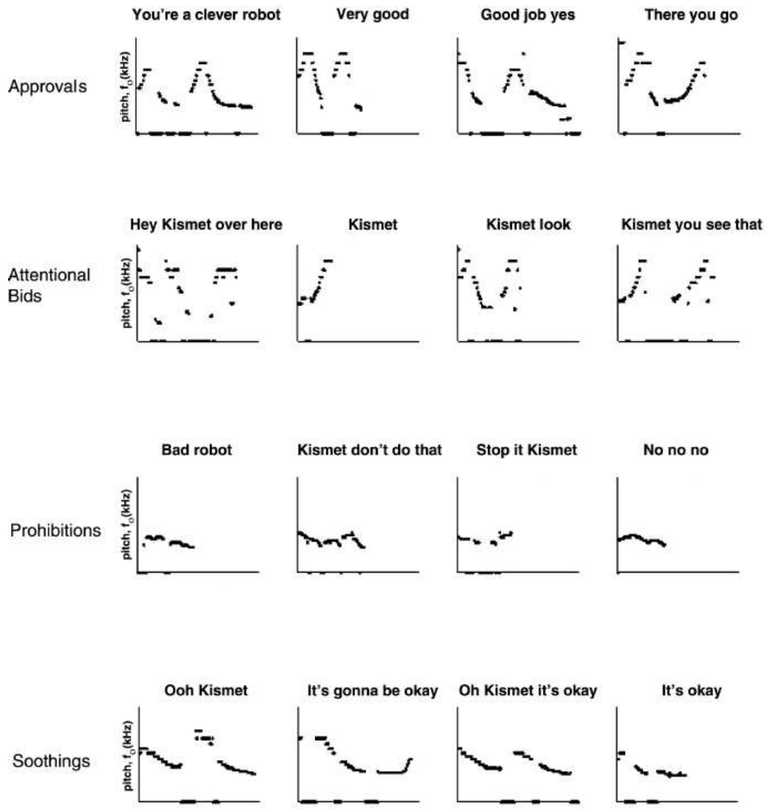

Figure 4. Fernald's prototypical prosodic contours found in the preprocessed data set. Notice the similarity to those shown in figure 2.

nearby regions.

2.1.3. Classification Method

In all training phases each class of data was modeled using a Gaussian mixture model, updated with the EM algorithm and a Kurtosis-based approach for dynamically deciding the appropriate number of kernels [13]. Due to the limited set of training data, cross-validation in all classification processes was performed. Specifically, a subset of data was set aside to train a classifier using the remaining data. The classifier's performance was then tested on the held-out test set. This process was repeated 100 times per classifier. The mean and variance of the percentage of correctly classified test data were calculated to estimate the classifier's performance.

Feature	*Description*
F_1	Pitch mean
F_2	Ptich Variance
F_3	Maximum Pitch
F_4	Minimum Pitch
F_5	Pitch Range
F_6	Delta Pitch Mean
F_7	Absolute Delta Pitch Mean
F_8	Energy Mean
F_9	Energy Variance
F_{10}	Energy Range
F_{11}	Maximum Energy
F_{12}	Minimum Energy

Table 1. Features extracted in the first-stage classifier. These features are measured over the non-zero values throughout the entire utterance. Feature F_6 measures the steepness of the slope of the pitch contour.

2.1.4. Feature Selection

As shown in figure 4, the preprocessed pitch contour in the labeled data resembles Fernald's prototypical prosodic contours for approval, attention, prohibition, and comfort/soothing. A set of global pitch and energy related features (see table 1) were used to recognize these proposed patterns. All pitch features were measured using only non-zero pitch values. Using this feature set, a sequential forward feature selection process was applied to construct an optimal classifier. Each possible feature pair's classification performance was measured and sorted from highest to lowest. Successively, a feature pair from the sorted list was added into the selected feature set to determine the best n features for an optimal classifier. Table 2 shows the results of the classifiers constructed using the best eight feature pairs. Classification performance increases as more features are added, reaches maximum (78.77 percent) with five features in the set, and levels off above 60 percent with six or more features. It was found that global pitch and energy measures were useful in roughly separating the proposed patterns based on arousal (largely distinguished by energy measures) and valence (largely distinguished by pitch measures). However, further processing was required to distinguish each of the five classes distinctly.

Accordingly, the classifier consists of several mini-classifiers executing in stages. In the beginning stages, the classifier uses global pitch and energy features to separate some of the classes into pairs (in this case, clusters of soothing along with low-energy neutral, prohibition along with high-energy neutral, and attention along with approval were formed). These clustered classes were then passed to additional classification stages for further refinement. New features had to be considered to build these additional classifiers. Using prior information, a new set of features encoding the shape of the pitch contour was included, which proved useful in further separating the classes.

Feature pair		Feature set	Performance mean (%)	Performance variance	% error approval	% error attention	% error prohibition	% error soothing	% error neutral
F1	F9	F1 F9	72.09	0.08	48.67	24.45	8.70	15.58	42.13
F1	F10	F1 F9 F10	75.17	0.12	41.67	25.67	9.65	13.15	33.98
F1	F11	F1 F9 F10 F11	78.13	0.08	29.85	27.20	8.80	10.63	32.90
F2	F9	F1 F2 F9 F10 F11	78.77	0.11	29.15	22.23	8.53	12.55	33.68
F1	F2								
F3	F9	F1 F2 F3 F9 F10 F11	61.52	1.16	63.87	43.03	9.08	23.05	53.35
F1	F8	F1 F2 F3 F8 F9 F10 F11	62.27	1.81	60.58	39.60	16.40	24.18	47.90
F5	F9	F1 F2 F3 F5 F8 F9 F10 F11	65.93	0.72	57.03	32.15	12.13	19.73	49.35

Table 2. The performance (the percent correctly classified) is shown for the best pair-wise set having up to eight features. The pair-wise performance was ranked for the best seven pairs. As each successive feature was added, performance peaks with five features (78.8%), but then drops off.

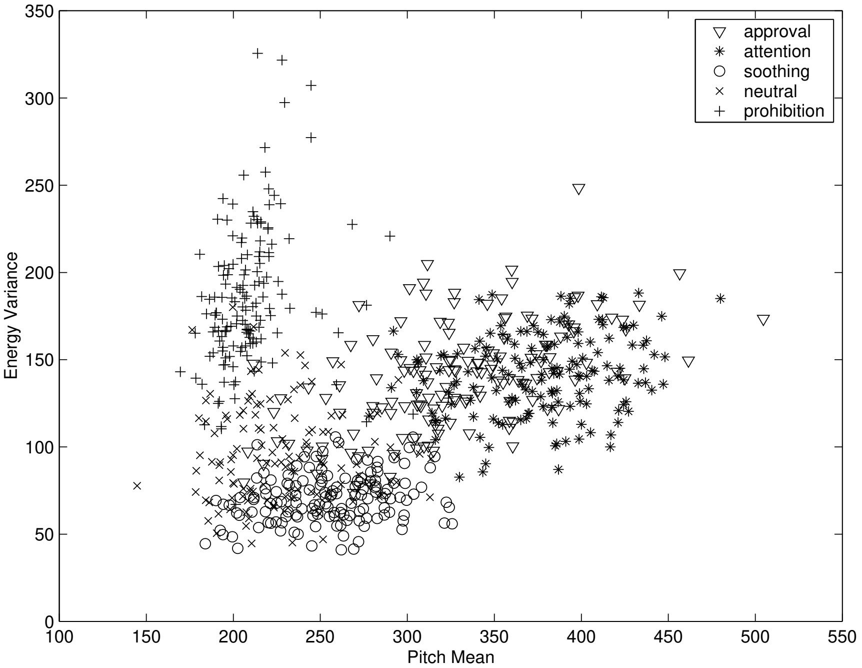

Figure 5. Feature space of all five classes with respect to energy variance, F_9, and pitch mean, F_1. There are three distinguishable clusters for prohibition, soothing and neutral, and approval and attention.

To select the best features for the initial classification stage, the seven feature pairs listed in table 2 were examined. All feature pairs worked better in separating prohibition and soothing than other classes. The F_1-F_9 pair generates the highest overall performance and the least number of errors in classifying prohibition. Several observations can be made from the feature space of this classifier(see figure 5). 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. The neutral samples have low pitch mean and are divided into two regions in terms of their energy variance values. The neutral samples with high energy variance are clustered separately from the rest of the classes (in between prohibition and soothing), while the ones with lower energy variance are clustered within the soothing class. These findings are consistent with the proposed prior knowledge. Approval, attention, and prohibition are associated with high intensity while soothing exhibits much lower intensity. Neutral samples span from low to medium intensity, which makes sense because the neutral class includes a wide variety of utterances.

Based on this observation, the first classification stage uses energy-related features to classify soothing and low-intensity neutral with from the other higher intensity classes (see figure 6). In the second stage, if the utterance had a low intensity level, another classifier decides whether it is soothing or neutral. If the utterance exhibited high intensity, the $F_1 - F_9$ pair is used to classify among prohibition, the approval-attention cluster, and high intensity neutral. An additional stage is required to classify between approval and attention if the utterance happened to fall within the approval-attention cluster.

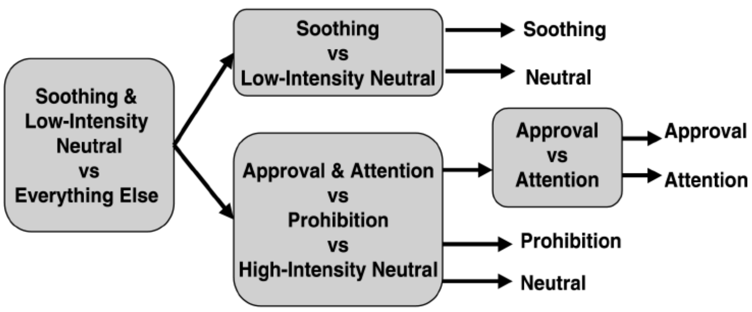

Figure 6. The classification stages of the multi-stage classifier.

2.1.5. Stage 1: Soothing — Low-Intensity Neutral versus Everything Else

The first two columns in table 3 show the classification performance of the top four feature pairs (sorted according to how well each pair classifies soothing and

Feature Pair	*Pair Perf. Mean (%)*	*Feature Set*	*Perf.* Mean (%)
F_9, F_{11}	93.0	F_9F_{11}	93.0
F_{10}, F_{11}	91.8	$F_9F_{10}F_{11}$	93.6
F_2, F_9	91.7	$F_2F_9F_{10}F_{11}$	93.3
F_7, F_9	91.3	$F_2F_7F_9F_{10}F_{11}$	91.6

Table 3. Classification results in stage 1.

low-intensity neutral against other classes). The last two columns illustrate the classification results as each pair is added sequentially into the feature set. The final classifier was constructed using the best feature set (energy variance, maximum energy, and energy range), with an average performance of 93.6 percent.

2.1.6. Stage 2A: Soothing versus Low-Intensity Neutral

Since the global and energy features were not sufficient in separating these two classes, new features were introduced into the classifier. Fernald's prototypical prosodic patterns for soothing suggest looking for a smooth pitch contour exhibiting a frequency down-sweep. Visual observations of the neutral samples in the data set indicated that neutral speech generated flatter and choppier pitch contours as well as less-modulated energy contours. Based on these postulations, a classifier using five features (number of pitch segments, average length of pitch segments, minimum length of pitch segments, slope of pitch contour, and energy range) was constructed. The slope of the pitch contour indicated whether the contour contained a down-sweep segment. It was calculated by performing a linear fit on the contour segment starting at the maximum peak. This classifier's average performance is 80.3 percent.

2.1.7. Stage 2B: Approval-Attention versus Prohibition versus High-Intensity Neutral

A combination of pitch mean and energy variance works well in this stage. The resulting classifier's average performance is 90.0 percent. Based on Fernald's prototypical prosodic patterns, it was speculated that pitch variance would be a useful feature for distinguishing between prohibition and the approval-attention cluster. Adding pitch variance into the feature set increased the classifier's average performance to 92.1 percent.

2.1.8. Stage 3: Approval versus Attention

Since the approval class and attention class span the same region in the global pitch versus energy feature space, prior knowledge (provided by Fernald's prototypical prosodic contours) gave the basis to introduce a new feature. As mentioned above, approvals are characterized by an exaggerated rise-fall pitch contour. This particular pitch pattern proved useful in distinguishing between the two classes. First, a three-degree polynomial fit was performed on each pitch segment. Each segment's slope sequence was analyzed for a positive slope followed by a negative slope with magnitudes higher than a threshold value.

	Class	Test Size	Classification Result					% Correctly Classified
			Approval	Attention	Prohibition	Soothing	Neutral	
First Pass	Approval	40	27	9	0	0	4	67.5
	Attention	40	11	29	0	0	0	72.5
	Prohibition	40	0	0	39	0	1	97.5
	Soothing	40	1	0	0	30	9	75
	Neutral	40	0	0	4	5	31	77.5
	All	200						78
Second Pass	Approval	84	64	15	0	5	0	76.19
	Attention	77	21	55	0	0	1	74.32
	Prohibition	80	0	1	78	0	1	97.5
	Soothing	68	0	0	0	55	13	80.88
	Neutral	62	3	4	0	3	52	83.87
	All	371						81.94

Table 4. Overall classification performance.

The longest pitch segment that contributed to the rise-fall pattern (which was 0 if the pattern was non-existent) was recorded. This feature, together with pitch variance, was used in the final classifier and generated an average performance of 70.5 percent. Approval and attention are the most difficult to classify because both classes exhibit high pitch and intensity. Although the shape of the pitch contour helped to distinguish between the two classes, it is very difficult to achieve high classification performance without looking at the linguistic content of the utterance.

2.1.9. Overall Performance

The final classifier was evaluated using a new test set generated by the same female speakers, containing 371 utterances. Because each mini-classifier was trained using different portions of the original database (for the single-stage classifier), a new data set was gathered to ensure that no mini-classifier stage was tested on data used to train it. Table 4 shows the resulting classification performance and compares it to an instance of the cross-validation results of the best single-stage five-way classifier obtained using the five features described in section 2.1.4. Both classifiers perform very well on prohibition utterances. The multi-stage classifier performs significantly better in classifying the *difficult* classes, i.e., approval versus attention and soothing versus neutral. This verifies that the features encoding the shape of the pitch contours (derived from prior knowledge provided by Fernald's prototypical prosodic patterns) were very useful.

It is important to note that both classifiers produce acceptable failure modes (i.e., strongly valenced intents are incorrectly classified as neutrally valenced intents and not as oppositely valenced ones). All classes are sometimes incorrectly classified as neutral. Approval and attentional bids are generally classified as one or the other. Approval utterances are occasionally confused for soothing and *vice versa*. Only one prohibition utterance was incorrectly classified as an attentional bid, which is acceptable. The single-stage classifier made one unacceptable error of confusing a neutral utterance as a prohibition. In the multi-stage classifier, some neutral utterances are classified as approval, attention, and soothing. This makes sense because the neutral class covers a wide variety of utterances.

3. Integration with the Emotion System

The output of the recognizer is integrated into the rest of Kismet's synthetic nervous system as shown in figure 7. The entry point for the classifier's result is at the auditory perceptual system. Here, it is fed into an associated releaser process. In general, there are many different kinds of releasers defined for Kismet, each combining different contributions from a variety of perceptual and motivational systems. Here, I only discuss those releasers related to the input from the vocal classifier. The output of each vocal affect releaser represents its perceptual contribution to the rest of the SNS. Each releaser combines the incoming recognizer signal with contextual information (such as the current "emotional" state) and computes its level of activation according to the magnitude of its inputs. If its activation passes above threshold, it passes its output on to the emotion system.

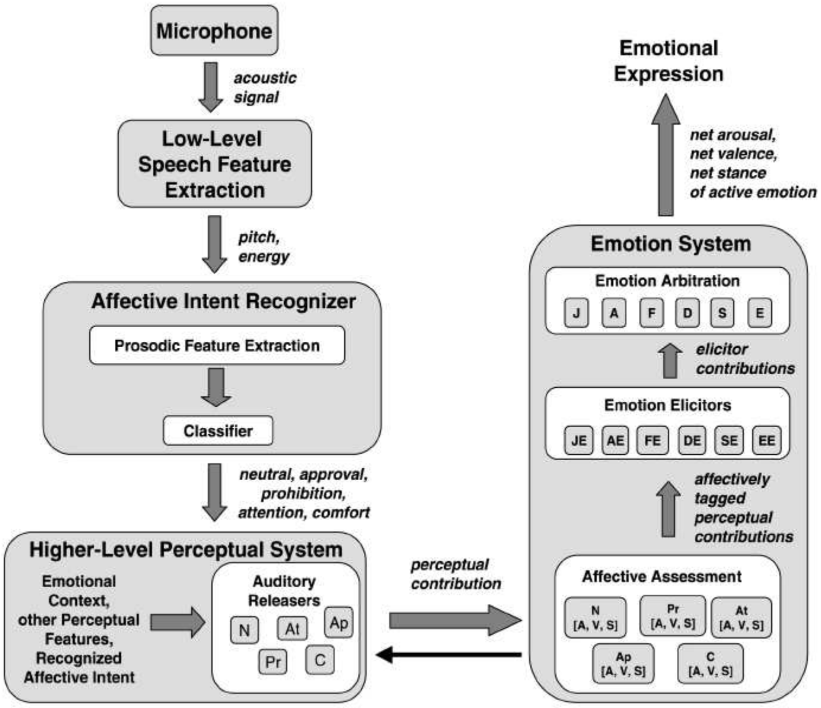

Figure 7. System architecture for integrating vocal classifier input to Kismet's emotion system.

Within the emotion system, the output of each releaser must first pass through the affective assessment subsystem in order to influence emotional behavior. Within this assessment subsystem, each releaser is evaluated in affective terms by an associated *somatic marker* (SM) process. This mechanism is inspired by the *Somatic Marker Hypothesis* of [3] where incoming perceptual

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	aproach	interest
Neutral	neutral	neutral	neutral	calm

Table 5. Table mapping $[A, V, S]$ to classified affective intents. Praise biases the robot to be "happy," prohibition biases it to be "sad," comfort evokes a "content, relaxed" state, and attention is "arousing".

information is "tagged" with affective information. Table 5 summarizes how each vocal affect releaser is somatically tagged.

There are three classes of tags that the affective assessment phase uses 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 arousing this percept is to the emotional system. Positive values correspond to a high arousal stimulus whereas negative values correspond to a low arousal stimulus. The *valence* tag, V, specifies how good or bad this percept is to the emotional system. Positive values correspond to a pleasant stimulus whereas negative values correspond to an unpleasant stimulus. The *stance* tag, S, specifies how approachable the percept is. Positive values correspond to advance whereas negative values correspond to retreat. 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 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. This filtering is done independently for each class of affective tag. For instance, a valence contribution with a large negative value will not only contribute to the `sorrow` emotion process, but to the `fear`, `anger`, and `distress` processes as well. Given all these factors, each elicitor computes its net $[A, V, S]$ contribution and activation level, and passes them to the associated emotion process within the *emotion arbitration* subsystem. In the second stage, the emotion processes within the emotion arbitration subsystem compete for activation based on their activation level. There is an emotion process for each of Ekman's six basic emotions [4]. Ekman posits that these six emotions are innate in humans, and all others are acquired through experience. The "Ekman six" encompass `joy`, `anger`, `disgust`, `fear`, `sorrow`, and `surprise`.

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. There are actually two threshold levels, one for expression and one for behavior. The expression threshold is lower than the behavior threshold; this allows the facial expression to *lead* the behavioral response. This enhances the readability and interpretation of the robot's behavior for the human observer. For instance, given that the caregiver makes an attentional bid, the robot's face 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. For Kismet, this feedback can come in a combination of facial expression, tone of voice, or posture. The robot's facial expression also sets up the human's expectation of what behavior will soon follow. 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. As I have argued previously, readability is an important issue for social interaction with humans.

3.1. 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.

Figure 8 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.

3.2. 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

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 8. Sample experiment session of a naive speaker, S3.

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.

4. 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.

4.1. 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.

		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 9. 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

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 to the robot (see the table to the left in figure 9). 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 9 shows that the robot is engaged in a smooth proto-dialog with the human partner the majority of the time (about 82%).

5. 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.

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

[4] P. Ekman 1992, "Are there basic emotions?", *Psychological Review* **99**(3), pp 550-553.

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

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.