Autonomous Robots 12, 83–104, 2002

Recognition of Affective Communicative Intent in Robot-Directed Speech

CYNTHIA BREAZEAL AND LIJIN ARYANANDA
Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 200 Technology Square rm 938, 936 Cambridge, MA 02139, USA
cynthia@ai.mit.edu
lijin@ai.mit.edu

Abstract. Human speech provides a natural and intuitive interface for both communicating with humanoid robots as well as for teaching them. In general, the acoustic pattern of speech contains three kinds of information: who the speaker is, what the speaker said, and how the speaker said it. This paper focuses on the question of recognizing affective communicative intent in robot-directed speech without looking into the linguistic content. We present an approach for recognizing four distinct prosodic patterns that communicate praise, prohibition, attention, and comfort to preverbal infants. These communicative intents are well matched to teaching a robot since praise, prohibition, and directing the robot's attention to relevant aspects of a task, could be used by a human instructor to intuitively facilitate the robot's learning process. We integrate this perceptual ability into our robot's "emotion" system, thereby allowing a human to directly manipulate the robot's affective state. This has a powerful organizing influence on the robot's behavior, and will ultimately be used to socially communicate affective reinforcement. Communicative efficacy has been tested with people very familiar with the robot as well as with naïve subjects.

Keywords: affective computing, human computer interaction, humanoid robots, sociable robots, speech recognition

1. Introduction

As robots take on an increasingly ubiquitous role in society, they must be easy for the average citizen to use and interact with. They must also appeal to persons of different age, gender, income, education, and so forth. This raises the important question of how to properly interface untrained humans with these sophisticated technologies in a manner that is intuitive, efficient, and enjoyable to use.

From the large body of human-technology research, we take as a working assumption that technological attempts to foster human-technology relationships will be accepted by a majority of people *if* the technological gadget displays rich social behavior (Reeves and Nass, 1996; Cassell et al., 1994). According to Reeves and Nass (1996), a social interface may very well be a universal interface because humans have evolved to be experts in social interaction. Similarity of morphology and sensing modalities makes humanoid robots one form of technology particularly well suited to this.

If Reeves and Nass findings hold true for humanoid robots, then those that participate in rich human-style social exchange with their users offer a number of advantages. First, people would find working with them more enjoyable and they would feel more competent. Second, communicating with them would not require any additional training since humans are already experts in social interaction. Third, if the robot could engage in various forms of social learning (imitation, emulation, tutelage, etc.), then it would be easier for the user to teach new tasks. Ideally, the user could teach the robot just as they would another person. Our group is particularly interested in this socially situated form of learning for humanoid robots, and we have argued for the many advantages social cues and skills could offer

robots that learn from people (Breazeal and Scassellati, 2000).

As one might imagine, a humanoid robot that could actually interact with people in a human-like way and be able to interpret, respond, and deliver human-style social cues (even at the level of a human infant) is quite a sophisticated machine. Over the past three years, we have been building infant-level social competencies into our robot, Kismet, so that we might explore social development and socially-situated learning between a robot and its human caregiver.

This paper explores one such competence: the ability to recognize affective communicative intent in robot-directed speech. Kismet has a fully integrated synthetic nervous system (SNS) that encompasses perceptual, attentional, motivational, behavioral, and motor capabilities (Breazeal, 1998). Within the motivational system are homeostatic regulation processes and emotional processes (Breazeal, 1999). As a whole, the motivation system provides affective information to the rest of the synthetic nervous system to influence behavior. Previous work has demonstrated how such systems can be used to bias learning both at goal-directed and affective levels (Blumberg, 1996; Velasquez, 1998; Yoon et al., 2000).

We are working towards implementing similar learning mechanisms on Kismet but with an added twist: the ability of the human caregiver to directly modulate the robot's affective state through verbal communication. This provides the human caregiver with natural and intuitive means for shaping the robot's behavior and for influencing what the robot learns. Particularly salient forms of vocal feedback include praise (positive reinforcement), prohibition (negative reinforcement), attentional bids (to direct the robot's attention to the important aspects of the task), and encouragement (to keep the robot motivated to try different things). Often these types of information are communicated affectively as well as linguistically in human speech.

In the rest of this paper we discuss previous work in recognizing emotion and affective intent in human speech. We discuss Fernald's work in depth to highlight the important insights it provides in terms of which cues are most useful for recognition of affective communicative intent as well as how it may be used by human infants to organize their behavior. We then outline a series of design issues particular to integrating this competence into our robot, Kismet. We present a detailed description of our approach and how we have integrated it into Kismet's affective circuitry. The performance of the system is evaluated with naïve subjects as well as the robot's caregivers. We discuss our results, suggest future work, and summarize our findings.

2. Emotion Recognition in Speech

There has been an increasing amount of work in identifying those acoustic features that vary with the speaker's affective state (Murray and Arnott, 1993). Changes in the speaker's autonomic nervous system can account for some of the most significant changes where the sympathetic and parasympathetic subsystems regulate arousal in opposition. For instance, when a subject is in a state of fear, anger, or joy, the sympathetic nervous system is aroused. This induces an increased heart rate, higher blood pressure, changes in depth of respiratory movements, greater subglottal pressure, dryness of the mouth, and occasional muscle tremor. The resulting speech is faster, louder, and more precisely enunciated with strong high frequency energy, a higher average pitch, and wider pitch range. In contrast, when a subject is tired, bored, or sad, the parasympathetic nervous system is more active. This causes a decreased heart rate, lower blood pressure, and increased salavation. The resulting speech is typically slower, lower-pitched, more slurred, and with little high frequency energy. Hence, the effects of emotion in speech tend to alter the pitch, timing, voice quality, and articulation of the speech signal (Cahn, 1990). However, several of these features are also modulated by the prosodic effects that the speaker uses to communicate grammatical structure and lexical correlates. For recognition tasks, this makes isolating those feature characteristics modulated by emotion challenging.

There have been a number of vocal emotion recognition systems developed in the past few years that use different variations and combinations of those acoustic features with different types of learning algorithms (Dellaert et al., 1996; Nakatsu et al., 1999). To give a rough sense of performance, a five-way classifier operating at approximately 80% is considered state of the art. This is impressive considering that humans cannot reliably discern a speaker's emotional state from speech alone. Some have attempted to use multimodal cues (facial expression with expressive speech) to improve recognition performance (Chen and Huang, 1998).

3. Affective Speech and Communicative Intent

However, for the purposes of training a robot, the raw emotional content of the speaker's voice is only part of the message. It tells us little about the intent of the message. A few researchers have developed recognition systems that can recognize speaker approval versus speaker disapproval from child-directed speech (Roy and Pentland, 1996), or recognize praise, prohibition, and attentional bids from infant-directed speech (Slaney and McRoberts, 1998).

However, 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* (Snow, 1972)). 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 (Grieser and Kuhl, 1988; Fernald, 1985; Mc Roberts et al., in press) (see Fig. 1). Mothers are more likely to use falling pitch contours than rising pitch contours when soothing a distressed infant (Papousek et al., 1985), to use rising contours to elicit attention and encourage a response (Ferrier, 1987), and to use bell shaped contours to maintain attention once it has been established (Stern et al., 1982). 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.

4. Affective Intent in Robot-Directed Speech

Inspired by this work, we have implemented a five-way recognizer that can distinguish Fernald's prototypical prosodic contours for praise, prohibition, comfort, attentional bids, and neutral speech. There are several design issues that must be addressed to successfully integrate Fernald's ideas into a robot like Kismet. As we have argued previously, this could provide a human caregiver with a natural and intuitive means for communicating with and training a robotic creature. The initial communication is at an affective level, where the caregiver socially manipulates the robot's affective state. For Kismet, the affective channel provides a powerful means for modulating the robot's behavior.

4.1. Robot Aesthetics

As discussed above, the perceptual task of recognizing communicative intent is significantly easier in infant-directed speech than in adult-directed speech. Even human adults have a difficult time recognizing intent from adult-directed speech without the linguistic information. However, it is a ways off before robots have natural language, but we can extract the affective content of the vocalization from prosody. This places a constraint on how the robot appears physically, how it moves, and how it expresses itself. If the robot looks and behaves as a very young creature, people will be more likely to treat it as such and naturally exaggerate their prosody when addressing the robot. This manner of robot-directed speech would be spontaneous and seems quite appropriate.

4.2. Real-Time Performance

Another design constraint is that the robot must be able to interpret the vocalization and respond to it at natural interactive rates. The human can tolerate small delays

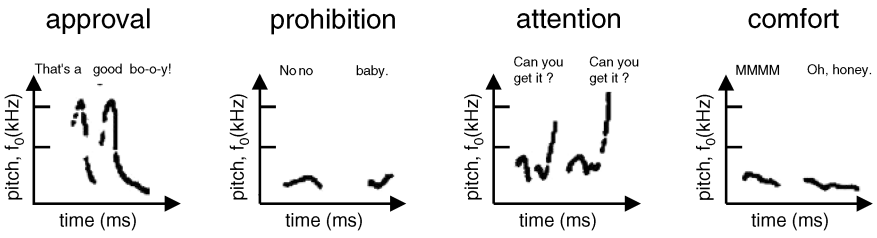

Figure 1. Fernald's prototypical prosodic contours for approval, attentional bid, prohibition, and soothing.

(perhaps a second or so), but long delays will break the natural flow of the interaction. Long delays also interfere with the caregiver's ability to use the vocalization as a reinforcement signal. Given that the reinforcement should be used to mark a specific event as good or bad, long delays could cause the wrong action to be reinforced and confuse the training process.

4.3. Voice as Training Signal

People should be able to use their voice as a natural and intuitive training signal for the robot. The human voice is quite flexible and can be used to convey many different meanings, affective or otherwise. The robot should be able to recognize when it is being praised and associate it with positive reinforcement. Similarly, the robot should recognize scolding and associate it with negative reinforcement. The caregiver should be able to acquire and direct the robot's attention with attentional bids to the relevant aspects of the task. Comforting speech should be soothing for the robot if it is in a distressed state, and encouraging otherwise.

4.4. Voice as Saliency Marker

This raises a related issue, which is the caregiver's ability to use their affective speech as a means of marking a particular event as salient. This implies that the robot should *only* recognize a vocalization as having affective content in the cases where the caregiver specifically intends to praise, prohibit, soothe, or get the attention of the robot. The robot should be able to recognize neutral robot-directed speech, even if it is somewhat tender or friendly in nature (as is often the case with motherese).

4.5. Acceptable vs Unacceptable Misclassification

Given that humans are not perfect at recognizing the affective content in speech, chances are the robot will make mistakes as well. However, some failure modes are more acceptable than others. For a teaching task, confusing strongly valenced intent for neutrally valenced intent is better than confusing oppositely valenced intents. For instance, confusing approval for an attentional bid, or prohibition for neutral speech, is better than interpreting a prohibition for praise. Ideally, the recognizer's failure modes will minimize these sorts of errors.

4.6. Expressive Feedback

Nonetheless, mistakes in communication will be made. This motivates the need for feedback from the robot back to the caregiver. Fundamentally, the caregiver is trying to communicate his/her intent to the robot. The caregiver has no idea whether or not the robot interpreted the intent correctly without some form of feedback. By interfacing the output of the recognizer to Kismet's emotional models, the robot's ability to express itself through facial expression, voice quality, and body posture will convey the robot's affective interpretation of the message to the caregiver. This enables people to reiterate themselves until they believe they have been properly understood. It also enables the caregiver to reiterate the message until the intent is communicated strongly enough ("What the robot just did was very good, and I want the robot to be really happy about it").

4.7. Speaker Dependence vs Independence

An interesting question is whether the recognizer should be speaker dependent or speaker independent. There are obviously advantages and disadvantages to both, and the appropriate choice depends on the application. Typically, it is easier to get higher recognition performance from a speaker dependent system than a speaker independent system. In the case of a personal robot, this is a good alternative since the robot should be personalized to a particular human over time, and should not be preferentially tuned to others. If the robot must interact with a wide variety of people, then the speaker independent system is preferable. The underlying question in both cases is what level of performance is necessary for people to feel that the robot is responsive and understands them well enough so that it is not challenging or frustrating to communicate with it and train it.

5. Robotic Physicality

Kismet is an expressive robotic creature with perceptual and motor modalities tailored to natural human communication channels (see Fig. 2). Kismet has three degrees of freedom to control gaze direction, three degrees of freedom to control its neck, and fifteen degrees of freedom in other expressive components of the face (such as ears, eyebrows, lips, and eyelids). Kismet is

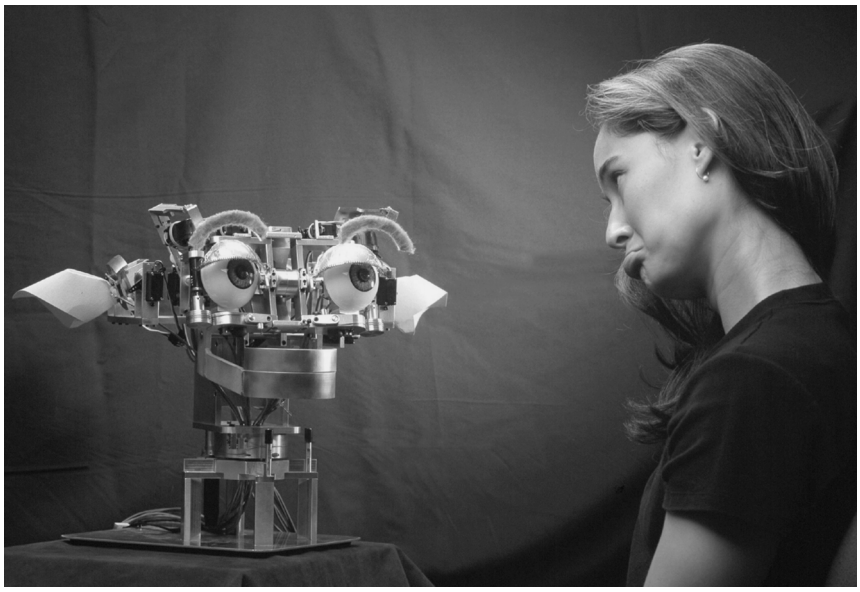

Figure 2. Kismet is an expressive robotic creature designed for natural social interaction with people.

able to display a wide assortment of facial expressions which mirror its affective state, as well as produce numerous facial displays for other communicative purposes (Breazeal and Scassellati, 1999).

To perceive its caregiver, Kismet uses a unobtrusive wireless microphone (worn by the human) and four color CCD cameras. Two wide field of view (fov) cameras are mounted centrally and move with respect to the head. They are used to direct the robot's attention toward people or toys and to compute a distance estimate. There is also a camera mounted within the pupil of each eye. These foveal cameras are used for higher resolution post-attentional processing, such as eye detection. The positions of the neck and eyes are important both for expressive postures and for directing the cameras towards behaviorally relevant stimuli. We have found that the manner in which the robot moves eyes and directs its gaze has profound social consequences when engaging people, beyond just steering its cameras to look at interesting things (Breazeal et al., 2000).

Aesthetically, Kismet is designed to have an infant-like appearance of a fanciful robotic creature. The key set of features that evoke nurturing responses of human adults has been studied across many different cultures (Eibl-Eibelsfeld, 1970), and these features have been explicitly incorporated into Kismet's design (Breazeal and Foerst, 1999). As a result, people tend to intuitively treat Kismet as a very young creature, and modify their behavior in characteristic baby-directed ways (Bullowa, 1979). One important implication of this is the *natural* use of "motherese" in Kismet-directed speech. Even the naïve subjects (male and female) use exaggerated prosody to address the robot. This allows us to readily exploit Fernald's affective communicative intent contours that she found to exist in infant-directed speech.

Our hardware and software control architectures have been designed to meet the challenge of real-time processing of visual signals (approaching 30 Hz) and auditory signals (frame size of 10 ms) with minimal latencies (<500 ms). Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system. Kismet's emotion, behavior, and expressive systems run on a collection of four Motorola 68332 processors. The affective speech recognition systems runs on Windows NT, and the low level speech processing software[1] runs on Linux. Even more so than Kismet's physical form, the control network is rapidly evolving as new behaviors and sensory modalities come on line.

6. The Algorithm

6.1. The Algorithmic Flow

As shown in Fig. 3, the affective speech recognizer receives robot directed speech as input. The speech signal

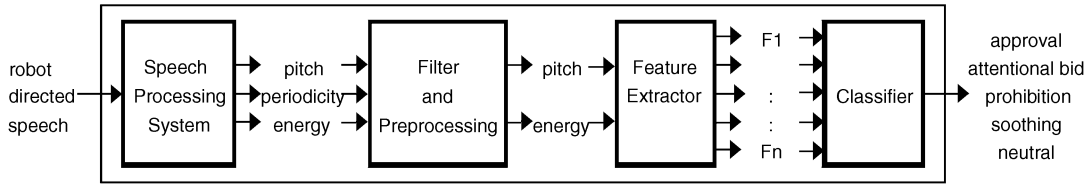

Figure 3. The affective speech recognition system.

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 all in real time.[2] 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 zero if the corresponding percent periodicity indicates that the frame is more likely to be unvoiced. 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, a soothing, or a neutral utterance.

6.2. *Training the System*

Data Collection. 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. We then eliminated samples containing extremely loud noises and labeled the remaining data set according to the speakers' communicative intents during the interaction. There were a total of 726 samples in the final data set.

Data Preprocessing. As mentioned above, the pitch value of a frame was set to zero if the corresponding percent periodicity was lower than a threshold value, indicating that the frame was more likely to be unvoiced. Even after this procedure, observation of the resulting pitch contours still indicated a lot of noise. Specifically, a significant number of errors were discovered in the high pitch value region (above 500 Hz). Therefore, additional preprocessing was performed to all pitch data. For each pitch contour, a histogram of ten regions was constructed. Using the heuristic that pitch contour was relatively smooth, we 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 a high but smooth pitch contour since pitch values would be distributed evenly across nearby regions.

Classification Method. 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 j-th component model parametrized on ϕ_j and π_j are the mixing weights satisfying $\sum_{j=1}^{k} \pi_j = 1$, $\pi_j \geq 0$. In this algorithm, kurtosis is viewed as measure of non-normality and 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 $\beta_j = \sum_{i=1}^{n} P(j \mid x_i)[(x_i - m_j)^T S_j^{-1}(x_i - m_j)]^2 / \sum_{i=1}^{n} P(j \mid x_i)$. 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 $|B| <$ a predefined threshold.

Due to the limited set of training data, we performed cross-validation in all classification processes.

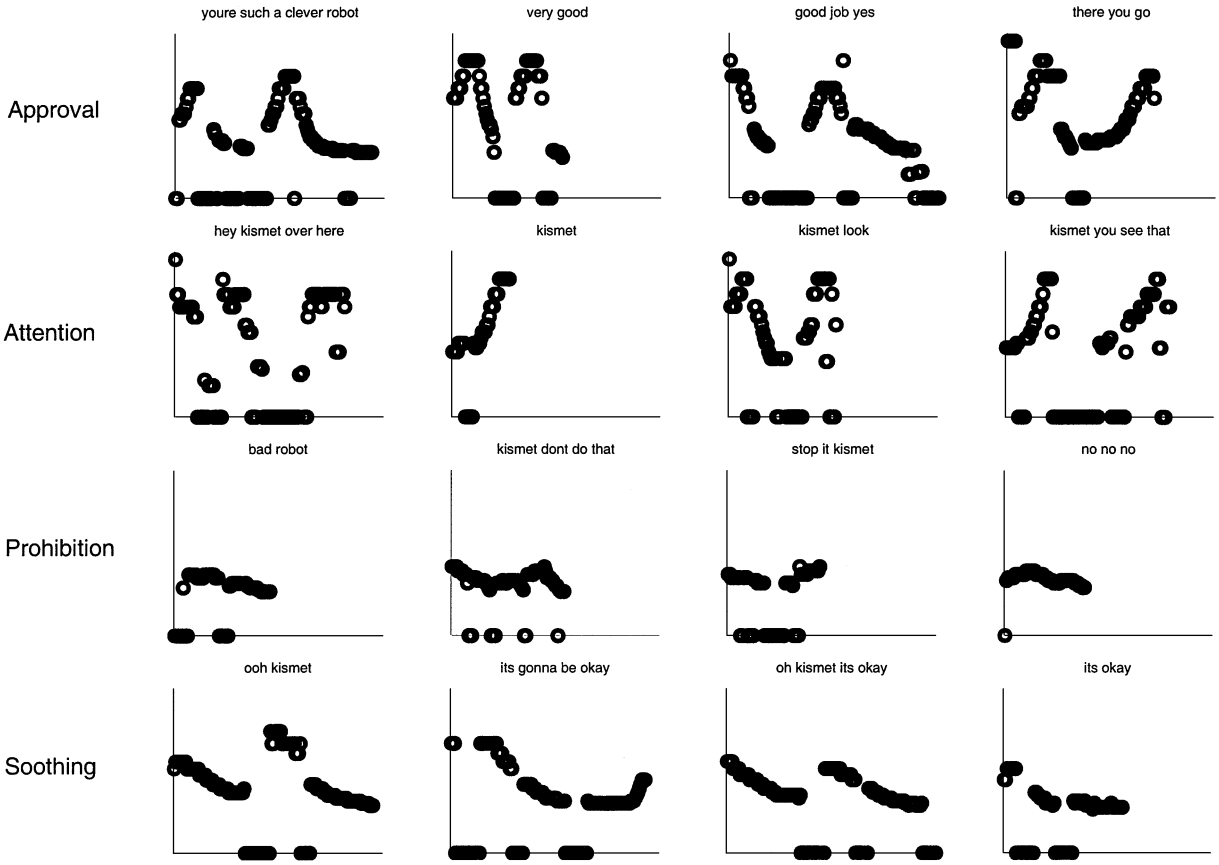

Figure 4. Fernald's prototypical prosodic contours found in the preprocessed data set.

Essentially, we held out a subset of data and built a classifier using the remaining training data, which was then tested on the held out test set. This process was repeated 100 times per classifier. Mean and variance of the percentage of correctly classified test data were calculated to estimate the classifier's performance.

Feature Selection. As shown in Fig. 4, the preprocessed pitch contour in the labeled data resembles Fernald's prototypical prosodic contours for approval, attention, prohibition, and comfort/soothing. In the first pass of training, we attempted to recognize these proposed patterns by using a set of global pitch and energy related features (see Table 1). All pitch features were measured using only non-zero pitch values. Using this feature set, we applied a sequential forward feature selection process 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 in order to determine the best n features for an optimal classifier. Table 2 shows results of the classifiers constructed using the best eight feature pairs. Classification performance increases as more features are added, reaches maximum (78.77%) with five features in the set, and levels off above 60% with six or more features. The number of misclassified samples in each class indicates that the global pitch and energy features were useful for separating prohibition from the other classes, but not sufficient for constructing a high performance 5-way classifier.

In the second pass of training, instead of having one optimal classifier that simultaneously classifies all five classes, we implemented several mini classifiers executing in stages. In the beginning stages, the classifier would use global pitch and energy features to separate some classes as well as they could. The remaining clustered classes were then passed to additional classification stages. Obviously, we had to consider new features in order to build these additional classifiers.

Table 1. Features extracted in the first pass.

	Feature description
F1	Pitch mean
F2	Pitch variance
F3	Maximum pitch
F4	Minimum pitch
F5	Pitch range
F6	Delta pitch mean
F7	Absolute delta pitch mean
F8	Energy mean
F9	Energy variance
F10	Energy range
F11	Maximum energy
F12	Minimum energy

Utilizing prior information, we included a new set of features encoding the shape of the pitch contour, which turned out to be useful in separating the *difficult* classes.

In order to select the best features for the initial classification stage, we observed the classification results of the best ten feature pairs obtained in the first pass (see Table 3). It is clear that all feature pairs work 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. We then carefully looked at the feature space of this classifier (see Fig. 5) and made several additional observations. 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, we concluded that in the first classification stage, we would use energy-related features to classify soothing and neutral with low intensity from the other higher intensity classes (see Fig. 6). In the second stage, if the utterance had a low intensity level, we would execute another classifier to decide whether it is soothing or neutral. If the utterance exhibited high intensity, we would use the F_1–F_9 pair to classify among prohibition, approval-attention cluster, and high intensity neutral. An additional stage would be required to classify between approval and attention if the utterance happened to fall within the approval-attention cluster.

Results

Stage 1: Soothing-Low Intensity Neutral vs Everything Else. The first two columns in Table 4 show classification performances of the top 4 feature pairs which are sorted based on how well each pair classifies soothing and 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

Table 2. First pass classification results.

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 3. Classification results of the best ten feature pairs.

Feature pair	Performance mean (%)	Performance variance	% error approval	% error attention	% error prohibition	% error soothing	% error neutral
F1 F9	72.09	0.08	48.675	24.45	8.7	15.575	42.125
F1 F10	70.96	0.08	41.95	26.625	15.1	15.15	46.4
F1 F11	70.03	0.08	29.525	29.275	19.05	14.75	57.275
F2 F9	68.79	0.096	45.675	33.75	13.75	13.85	49
F1 F2	65.47	0.1	41.625	18.275	24.075	25.875	62.8
F3 F9	64.04	0.2	68.75	37	13.775	18.325	41.925
F1 F8	63.6	0.13	44.55	27.2	21.675	27.15	61.425
F5 F9	63.49	0.11	38.575	57.075	20.625	18.375	47.9
F4 F9	63.42	0.11	52.125	45.275	25.675	17.15	42.675
F2 F11	63.28	0.09	35.325	39.525	20.05	17.625	71.075

(energy variance, maximum energy, and energy range), with an average performance of 93.57%. The resulting feature space is shown in Fig. 7.

Stage 2A: Soothing vs Low Intensity Neutral. Since the global and energy features were not sufficient in separating these two classes, we had to introduce new features into the classifier. Fernald's prototypical prosodic patterns for soothing proposed smooth pitch contours exhibiting a frequency downsweep. Visual observations of the neutral samples in the data set indicated that neutral speech generated flatter and

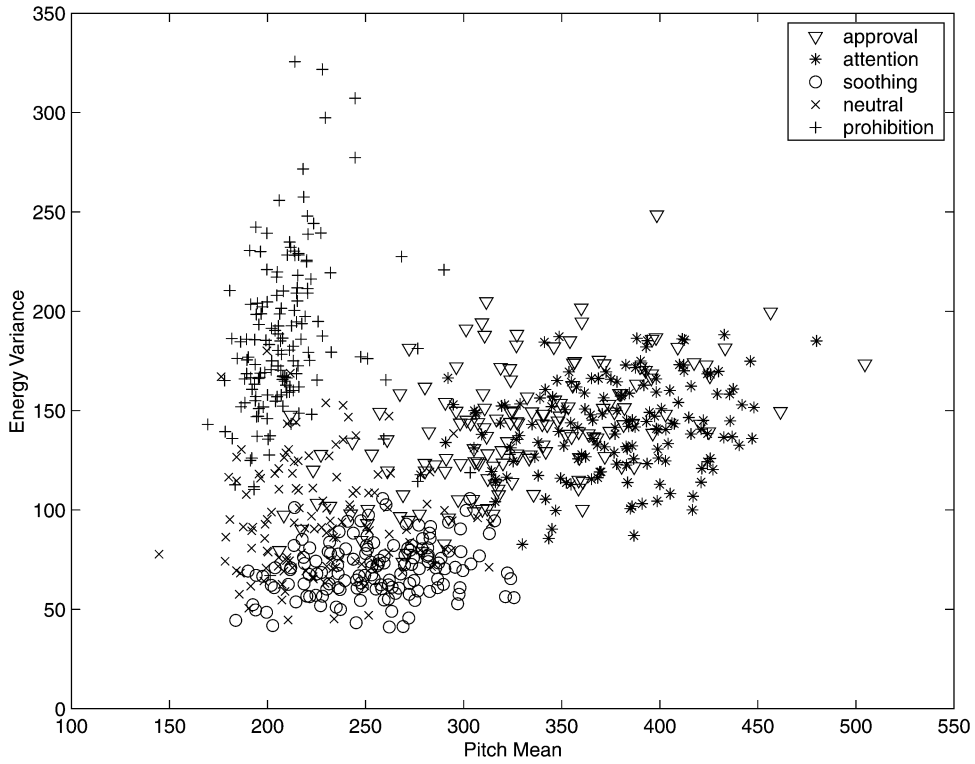

Figure 5. Feature space of all five classes.

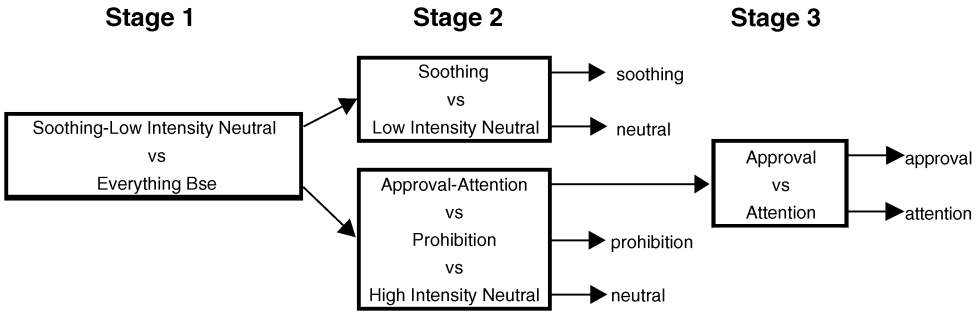

Figure 6. The classification stages.

coarse pitch contours as well as less modulated energy contours. Based on these postulations, we constructed a classifier using five features, i.e., number of pitch segments, average length of pitch segments, minimum length of pitch segments, slope of pitch contour, and energy range. The slope of pitch contour indicated whether or not the contour contained a downsweep segment. It was calculated by performing a 1-degree polynomial fitting on the remaining segment of the contour after the maximum peak. This classifier's average performance is 80.29%.

Stage 2B: Approval-Attention vs Prohibition vs High Intensity Neutral. We have discovered that a combination of pitch mean and energy variance works well in this stage. The resulting classifier's average performance is 89.99%. Based on Fernald's prototypical prosodic patterns and the feature space shown in Fig. 8, we speculated that pitch variance would be a useful feature for distinguishing between prohibition and approval-attention cluster. Adding pitch variance into the feature set increases classifier's average performance to 92.13%.

Table 4. Classification results in stage 1.

Feature pair	Pair performance mean (%)
F9 F11	93.00
F10 F11	91.82
F2 F9	91.7
F7 F9	91.34
Feature set	**Performance mean (%)**
F9 F11	93.00s
F9 F10 F11	93.57
F2 F9 F10 F11	93.28
F2 F7 F9 F10 F11	91.58

Stage 3: Approval vs Attention. Since approval and attention classes span across the same region in the global pitch and energy feature space, we utilized prior knowledge provided by Fernald's prototypical prosodic contours to introduce a new feature. As mentioned above, approvals are characterized by an exaggerated rise-fall pitch contour. We hypothesized that the existence of this particular pitch pattern would be a useful feature in distinguishing between the two classes. We first performed a 3-degree polynomial fitting on each pitch segment. We then analyzed each segment's slope sequence and looked for a positive slope followed by a negative slope with magnitudes higher than a threshold value. We recorded the maximum length of pitch segment contributing to the rise-fall pattern which was zero if the pattern was non-existent. This feature, together with pitch variance, was used in the final classifier and generated an average performance of 70.5%. This classifier's feature space is shown in Fig. 9. 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.

Overall Performance. The final classifier was evaluated using a new test set generated from the same speakers, containing 371 utterances. Table 5 shows the resulting classification performance and compares it to an instance of the cross-validation results of the best classifier obtained in the first pass. Both classifiers perform very well on prohibition utterances. The second pass classifier performs significantly better in classifying the *difficult* classes, i.e., approval vs attention and soothing vs neutral, thereby verifying that features encoding the shape of pitch contours derived based on prior knowledge provided

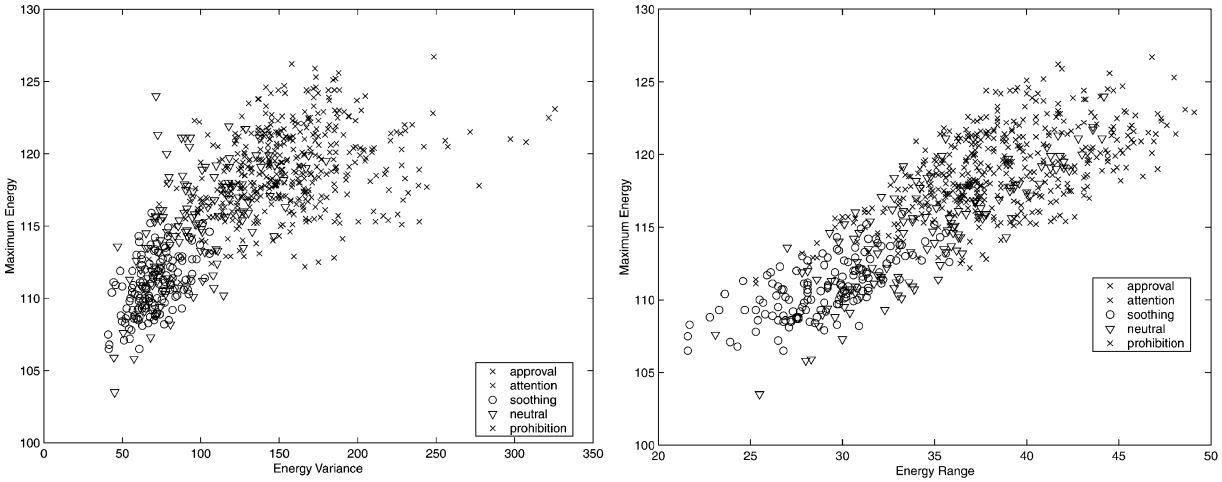

Figure 7. Feature space: Soothing vs neutral vs rest.

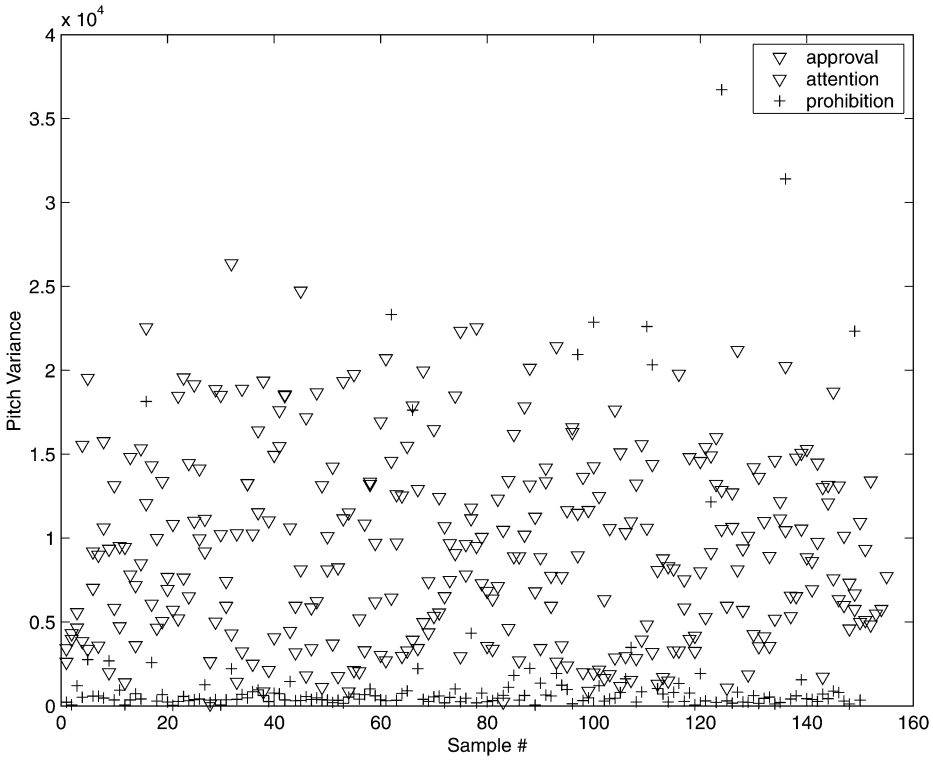

Figure 8. Feature space: Approval-attention vs prohibition.

by Fernald's prototypical prosodic patterns are very useful.

It is important to note that both classifiers produce acceptable failure modes, i.e., strongly valenced intents are misclassified as neutrally valenced intents and not as oppositely valenced ones. All classes are sometimes misclassified 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 probihition utterance was misclassified as an attentional bid, which is acceptable. The first pass made one unacceptable error of confusing

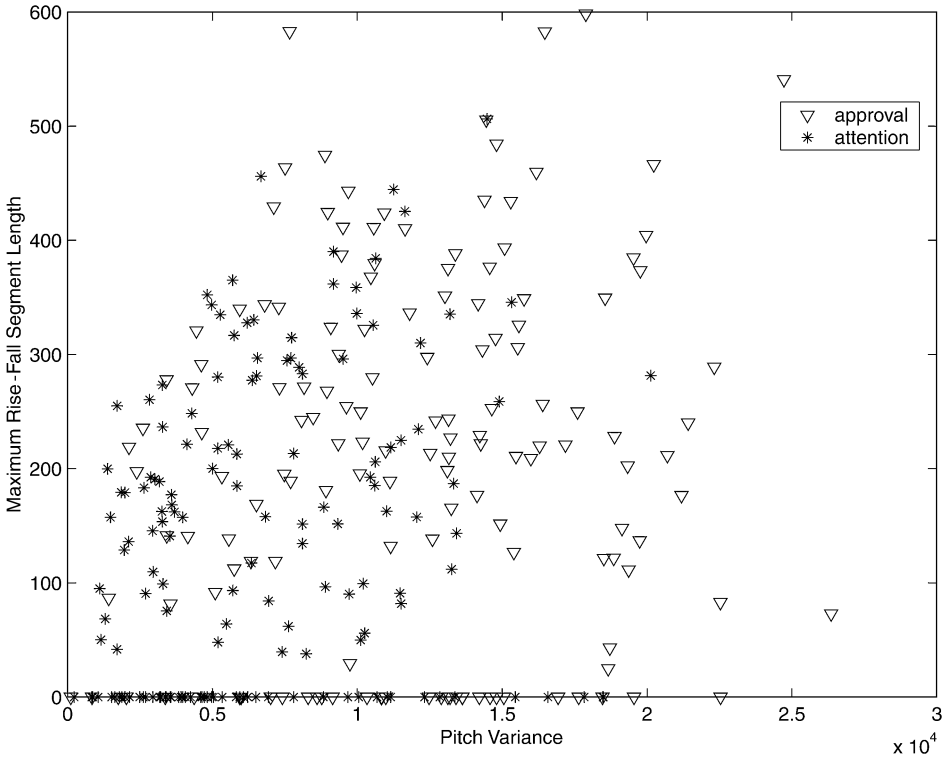

Figure 9. Feature space: Approval vs attentional bid.

a neutral as prohibition. In the second pass classifier, some neutral utterances are classified as approval, attention, and soothing. This makes sense because the neutral class covers a wide variety of utterances.

7. Integration with the Emotion System

The output of the recognizer is integrated into the rest of Kismet's synthetic nervous system as shown

Table 5. Overall classification performance.

			Classification result					
	Class	Test size	Approval	Attention	Prohibition	Soothing	Neutral	% correctly classified
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

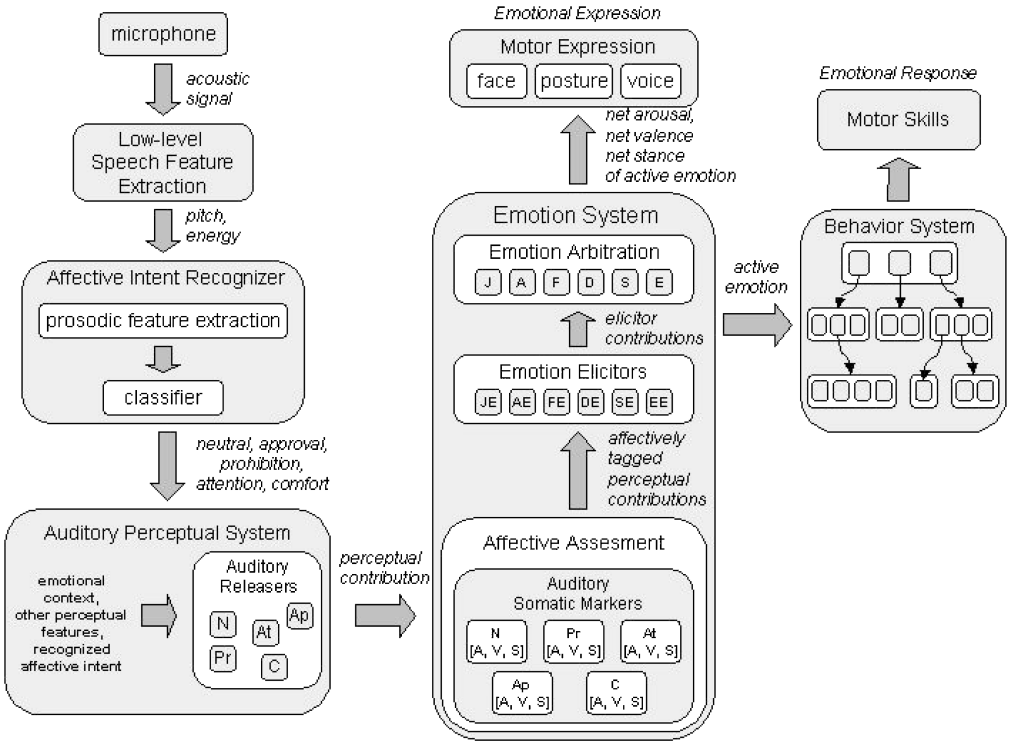

Figure 10. System architecture for Kismet.

in Fig. 10. Its entry point is at the auditory perceptual system, where 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. For the purposes here, we only discuss those releasers related to the input from the vocal affect 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 onto the affective assessment stage so that it may influence emotional behavior.

Within this assessment phase, each releaser is evaluated in affective terms by an associated *somatic marker* (SM) process. This mechanism is inspired by the *Somatic Marker Hypothesis* of Damasio (1994) where incoming perceptual information is "tagged" with affective information. Table 6 summarizes how each vocal affect releaser is somatically tagged. We have applied a slight twist to Fernald's work in using approvals and prohibitions to modulate the valence of Kismet's affective state in addition to arousal (Fernald focuses on the impact of these contours on arousal levels of infants).

Table 6. Affective tags for the output of the affective intent recognizer.

	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

There are three classes of tags the SM uses to affectively characterize its perceptual (as well as motivational and behavioral) input. Each tag has an associated intensity that scales its contribution to the overall affective state. The *arousal* tag specifies how arousing this percept is to the emotional system. Positive values correspond to a high arousal stimulus whereas negative

values correspond to a to low arousal stimulus. The *valence* tag 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 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 [*arousal, valence, stance*] 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 `sad` 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.

In the second 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). 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 becomes active. By staging the response in this manner, the caregiver gets immediate expressive feedback that the robot understood his/her intent. For Kismet, this feedback can come in a combination of facial expression, tone of voice, or posture. The facial expression also sets up the human's expectation of what robot behavior will soon follow. As a result, the human observing the robot not only can see what the robot is doing, but has an understanding of why. Readability is an important issue for social interaction with humans.

8. Use of Behavioral Context to Improve Interpretation

Most affective speech recognizers are not integrated into robots equipped with affect systems that are embedded in a social environment. As a result, they have to classify each utterance in isolation. However, for Kismet, the surrounding social context can be exploited to help reduce false categorizations; or at least to reduce the number of "bad" misclassifications (such as mixing up prohibitions for approvals).

8.1. Transition Dynamics of the Emotion System

Some of this contextual filtering is performed by the transition dynamics of the emotion processes. These processes cannot instantaneously become active or inactive. Decay rates and competition for activation with other emotion processes give the currently active process a base-level of persistence before it becomes inactive. Hence, for a sequence of approvals where the activation of the robot's `happy` process is very high, an isolated prohibition will not be sufficient to immediately switch the robot to a negatively valenced state.

However, if the caregiver in fact intended to communicate disapproval to the robot, reiteration of the prohibition will continue to increase the contribution of negative valence to the emotion system. This serves to inhibit the positively valenced processes and to excite the negatively valenced processes. Expressive feedback from the robot is sufficient for the caregiver to recognize when the intent of the vocalization has been communicated properly and has been communicated strongly enough. The smooth transition dynamics of the emotion system enhances the naturalness of the robot's behavior since a person would expect to have to "build up" to a dramatic shift in affective state from positive to negative, as opposed to being able to flip the robot's emotional state like a switch.

8.2. *Using Social Context to Disambiguate Intent*

The affective state of the robot can also be used to help disambiguate the intent behind utterances with very similar prosodic contours. A good example of this is the difference between utterances intended to soothe versus utterances intended to encourage the robot. The prosodic patterns of these vocalizations are quite similar, but the intent varies with the social context. The communicative function of soothing vocalizations are to comfort a distressed robot—there is no point in comforting the robot if it is not in a distressed state. Hence, the affective assessment phase somatically tags these types of utterances as soothing when the robot is distressed, and as encouraging otherwise.

9. Experiments

9.1. *Motivation*

We have shown that the implemented classifier performs well on the primary caregivers' utterances. Essentially, the classifier is trained to recognize the caregivers' different prosodic contours, which are shown to coincide with Fernald's prototypical patterns. In order to extend the use of the affective intent recognizer, we would like to evaluate the following issues:

- Will naïve subjects speak to the robot in an exaggerated manner (in the same way as the caregivers)? Will Kismet's infant-like appearance urge the speakers to use *motherese*?
- If so, will the classifier be able to recognize their utterances, or will it be hindered by variations in individual's style of speaking or language?
- How will the speakers react to Kismet's expressive feedback, and will the cues encourage them to adjust their speech in a way they think that Kismet will understand?

9.2. *Experimental Setup*

Five female subjects, ranging from 23 to 54 years old, were asked to interact with Kismet in different languages (English, Russian, French, German, and Indonesian). Subjects were instructed to express each communicative intent (approval, attention, prohibition, and soothing) and signal when they felt that they had communicated it to the robot. We did not include the neutral class because we expected that many neutral utterances would be spoken during the experiment. All sessions were recorded on video for further evaluations.

9.3. *Results*

A set of 266 utterances were collected from the experiment sessions. Very long and empty utterances (those containing no voiced segments) were not included. An objective observer was asked to label these utterances and to rate them based on the perceived strength of their affective message (except for neutral). As shown in the classification results (see Table 7), compared to the caregiver test set, the classifier performs almost as well on neutral, and performs decently well on all the *strong* classes, except for soothing and attentional bids. As expected, the performance reduces as the perceived strength of the utterance decreases.

A closer look at the misclassified soothing utterances showed that a high number of utterances were actually soft approvals. The pitch contours contained a rise-fall segment, but the energy level was low. A 1-degree polynomial fitting on these contours will generate a flat slope and thus classified as neutral. A few soothing utterances were confused for neutral despite having the downsweep frequency characteristic because they contained too many words and coarse pitch contours. Attentional bids generated the worst classification performance. A careful observation of the classification errors revealed that many of the misclassified attentional bids contained the word "kis-met" spoken with a bell-shaped pitch contour. This was detected by the classifier as the characteristic rise-fall pitch segment found in approvals. We also found that many other common words used in attentional bids, such as "hey" and "hello", also generated a bell-shaped pitch contour. Interestingly, these attentional bids appear to carry stronger affective message because they do not occur as much in the medium strength utterances, which are thus easier to classify. These are obviously very important issues to be resolved in future efforts to improve the system.

Based on these findings, we can draw several conclusions. Firstly, a high number of utterances are perceived to carry *strong* affective message, which implies the use of exaggerated prosody during the interaction session that we hoped for. The remaining question is whether or not the classifier will generalize to the naïve speakers' exaggerated prosodic patterns. Except for the two special cases discussed above, experimental results

Table 7. Classification performance on naïve speakers.

Test set	Strength	Class	Test size	Classification result					% correctly classified
				Approval	Attention	Prohibition	Soothing	Neutral	
Care givers		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
Naive speakers	Strong	Approval	18	14	4	0	0	0	72.2
		Attention	20	10	8	1	0	1	40
		Prohibition	23	0	1	20	0	2	86.96
		Soothing	26	0	1	0	16	10	61.54
	Medium	Approval	20	8	6	0	1	5	40
		Attention	24	10	14	0	0	0	58.33
		Prohibition	36	0	5	12	0	18	33.33
		Soothing	16	0	0	0	8	8	50
	Weak	Approval	14	1	3	0	0	10	7.14
		Attention	16	7	7	0	0	2	43.75
		Prohibition	20	0	4	6	0	10	30
		Soothing	4	0	0	0	0	4	0
		Neutral	29	0	1	0	4	24	82.76

indicate that the classifier performs very well in recognizing the naïve speakers' prosodic contours although it is trained only on the primary caregivers' utterances. Moreover, the same failure modes occur in the naïve speakers test set. No strongly valenced intents were misclassified as oppositely valenced ones. It is very encouraging to discover that the classifier not only generalizes to perform well on naïve speakers using different languages, but it also does not make any (or at least very few) unacceptable misclassifications.

10. Discussion

Results from these initial studies and other informal observations suggest that people do naturally exaggerate their prosody (characteristic of motherese) when addressing Kismet. People of different genders and ages often comment that they find the robot to be "cute", which encourages this manner of address. Naïve subjects appear to enjoy interacting with Kismet and are often impressed at how life-like it behaves. This also promotes natural interactions with the robot, making it easier for them to engage the robot as if it were a very young child or adored pet.

All of our female subjects spoke to Kismet using exaggerated prosody characteristic of infant-directed speech. It is quite different from the manner in which they spoke with the experimenters. We have informally noticed the same tendency with children (approximately twelve years of age) and adult males. It is not surprising that individual speaking styles vary. Both children and women (especially those with young children or pets) tend to be uninhibited, whereas adult males are often more reserved. For those who are relatively uninhibited, their styles for conveying affective communicative intent vary. However, Fernald's contours hold for the strongest affective statements in all of the languages that were explored in this study. This would account for the reasonable classifier performance on vocalizations belonging to the strongest affective category of each class. As argued previously, this is the desired behavior for using affective speech as an emotion-based saliency marker for training the robot.

Tables 8 and 9 illustrate sample event sequences that occured during experiment sessions of a caregiver (S1) and naïve speaker (S2) respectively. Each row represents a trial in which the subject attempts to communicate an affective intent to Kismet. For each trial, we

Table 8. Sample experiment session of a caregiver.

Speaker	Intent	Trial	No. of utterances	Robot's cues	Correct?	Subject's response	Change in prosody	Subject's comment
S1—caregiver (English)	Approval	1	2	Smile	Yes	Laugh and acknowledge		
		2	2	Lean forward	No	Giggle		
		3	1	Smile	Yes	Acknowledge		
		4	2	Smile	Yes	Acknowledge		
		5	1	Smile	Yes	Acknowledge		
		6	2	Attending	No	Ignore		
		7	1	Smile	Yes	Acknowledge		
		8	1	Smile	Yes	Acknowledge		
		9	1	Smile	Yes	Acknowledge		
		10	1	Smile	Yes	Acknowledge		"It liked that one"
		11	1	Smile	Yes	Acknowledge		
		12	1	Smile	Yes	Acknowledge		
		13	1	Smile	Yes	Acknowledge		
	Attention	14	1	Smile	No	Ignore		
		15		Attending	Yes	Acknowledge		
		16	2	Attending	Yes	Acknowledge	Louder	
		17	1	Attending	Yes	Acknowledge		
		18	1	Attending	Yes	Acknowledge		
		19	1	Smile	No	Ignore		"It thinks I'm approving it"
		20	1	Attending	Yes	Acknowledge		
		21	1	Smile	No	Acknowledge		
		22	1	Attending	Yes	Acknowledge		
		23	2	Attending	Yes	Acknowledge		
		24	1	Attending	Yes	Acknowledge		
		25	2	Smile	No	Ignore		
		26	2	Attending	Yes	Acknowledge		"There's a lag here"
		27	1	Attending	Yes	Acknowledge		
		28	1	Attending	Yes	Acknowledge		"You know when it's in between, before it gets excited, I'm all scared that it's going to get sad"
		29	1	Attending	Yes	Acknowledge		
	Prohibition	30	4	Look down	No	Ignore		
		31		Frown	Yes	Acknowledge		"Oooh..."
		32	2	Look down	No	Ignore	Lower pitch	
		33	1	Still look down	Yes	Acknowledge		"Sorry, are you okay now? It's just an experiment, Kismet. I have to do it"
		34	3	Look down	Yes	Acknowledge	Louder	
		35		Frown	Yes	Ignore		

(*continued on next page.*)

Table 8. (*continued*).

Speaker	Intent	Trial	No. of utterances	Robot's cues	Correct?	Subject's response	Change in prosody	Subject's comment
		36	4	Look down	No	Ignore		
		37	3	Frown	Yes	Acknowledge		
		38	4	Look down and frown	Yes	Acknowledge		
		39	2	Look down and frown	Yes	Acknowledge		
	Soothing	40	1	Look up and ears perk up	Yes	Acknowledge		
	Prohibition	41	4	Look down and frown	Yes	Acknowledge		
	Soothing	42	4	Look up and ears perk up	Yes	Acknowledge		
	Prohibition	43	3	Look down and frown	Yes	Acknowledge		
	Soothing	44	3	Look up and smile	Yes	Acknowledge		
	Prohibition	45	4	Look down	No	Ignore		
		46	2	Frown	Yes	Acknowledge		
(Indonesian)	Approval	47	1	Smile	Yes	Smile and acknowledge		
		48	1	Smile	Yes	Acknowledge		
		49	3	Smile	Yes	Acknowledge		
		50	2	Grin	No	Ignore		
		51	1	Smile	Yes	Acknowledge		
	Attention	52	3	Attending	Yes	Acknowledge		
		53	3	Attending	Yes	Acknowledge		
	Prohibition	54	4	Look down and frown	Yes	Acknowledge		
		55	4	Look down and frown	Yes	Acknowledge		
		56	3	Look down and frown	Yes	Acknowledge		
	Soothing	57	2	Look up	Yes	Acknowledge		
	Prohibition	58	4	Look down and frown	Yes	Acknowledge		
	Soothing	59	3	Look up	Yes	Acknowledge		

recorded the number of utterances said, Kismet's cues, subject's responses and comments, as well as changes in prosody, if any. 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 26 of subject S2's experiment session, subject giggled when Kismet smiled despite her scolding,

Table 9. Sample experiment session of a naïve speaker.

Speaker	Intent	Trial	No. of utterances	Robot's cues	Correct?	Subject's response	Change in prosody	Subject's comment
S2—naive	Approval	1	1	Ears perk up	No	Smile and acknowledge		
		2	1	Ears perk up, a little grin	No	Smile and acknowledge		
		3	2	Look down	No	Lean forward	Higher pitch	
		4	2	Look up	No	Smile and acknowledge	Higher pitch	
		5	1	Ears perk up, a little grin	No	Lean forward, smile, and acknowledge		"I had it"
		6		Lean forward and smile		Smile		
		7	2	Smile	Yes	Lean forward, smile, and acknowledge	Higher pitch	
		8	3	Smile	Yes	Lean forward, smile, and acknowledge	Higher pitch	
		9	4	Attending	No	Ignore		
		10		Smile	Yes	Lean forward, smile, and acknowledge		
	Attention	11	3	Make eye contact	No	Smile and acknowledge	Higher pitch	
		12	1	Attending	Yes	Acknowledge		
		13	1	Attending	Yes	Acknowledge		
		14	1	Attending	Yes	Acknowledge		
		15	2	Lean forward and make eye contact	No	Acknowledge		
		16	2	Lean back and make eye contact	No	Lean forward and acknowledge		
		17		Look down and frown		Ignore		
		18	4	Look up	No	Lean forward, smile, and acknowledge	Higher pitch	
	Prohibition	19	4	Look down	No	Lean forward, keep on talking		
		20	4	Frown	Yes	Acknowledge	Lower pitch	
		21	6	Look down	No	Lean forward, keep on talking		
	Soothing	23	4	Look up and make eye contact	Yes	Pauses and acknowledge		
	Prohibition	24	6	Frown	Yes	Acknowledge		
	Soothing	25	4	Look up and make eye contact	Yes	Pauses and acknowledge		

commented that volume would help, and thus spoke louder in the next trial.

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 trial 30, 32, and 36, Kismet responded to subject S1's scolding by dipping its head and subject continued prohibiting with lower voice until Kismet finally frowned.

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". Subject S1 was very apologetic throughout her prohibition session. 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.

Another interesting social dynamic we observed involved *affective mirroring* between robot and human. In this situation, the subject might first issue a medium strength prohibition to the robot, which causes it to dip its head. The subject 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 (trial 19–21 in S2's session). This technique was employed to modulate the degree to which the strength of the message was "communicated" to the robot.

11. Limitations and Extensions

The ability of naïve subjects to interact with Kismet in this affective and dynamic manner suggests that its response rate is of acceptable performance. However, the timing delays in the system can and should be improved. There is about a 500 ms delay from the time speech ends to receiving an output from the classifier. Much of this delay is due to the underlying speech recognition system, where there is a trade-off between shipping out the speech features to the NT machine immediately after a pause in speech, or waiting long enough during that pause to make sure that speech has completed. There is another delay of one to two seconds associated with interpreting the classifier in affective terms and feeding it through an emotional response. The subject will typically issue one to three short utterances during this time (of a consistent affective content). It is interesting that people seem to rarely issue just one short utterance and wait for a response. Instead, they prefer to communicate affective meanings in a sequence of a few closely related utterances ("That's right Kismet. Very good! Good robot!"). In practice, people do not seem to be bothered by or notice the delay. The majority of delays involve waiting for a sufficiently strong vocalization to be spoken, since only these are recognized by the system.

Given the motivation of being able to use natural speech as a training signal for Kismet, it remains to be seen how the existing system needs to be improved or changed to serve this purpose. Naturally occurring robot-directed speech doesn't come in nicely packaged sound bites. Often there is clipping, multiple prosodic contours of different types in long utterances, and other background noise (door's slamming, people talking, etc.). Again, targeting infant-caregiver interactions goes some ways in alleviating these issues, as infant-directed speech is slower, shorter, and more exaggerated. However, our collection of robot-directed utterances demonstrates a need to address these issues carefully.

The recognizer in its current implementation is specific to female speakers, and it is particularly tuned to women who can use motherese effectively. Granted not all people will want to use motherese to instruct their robots. However, at this early state of research we are willing to exploit *naturally occurring* simplifications of robot-directed speech to explore human-style socially situated learning scenarios. Given the classifier's strong performance for the caregivers (those who will instruct the robot intensively), and decent performance for other female speakers (especially for prohibition and approval), we are quite encouraged at these early results. Future improvements include either training a male adult model, or making the current model more gender neutral.

For instructional purposes, the question remains "how good is good enough?". Seventy to eighty percent performance of five-way classifiers for recognizing emotional speech is regarded as state of the art. In practice, within an instructional setting, this may be an unacceptable number of misclassifications. As a result, we have taken care in our approach to minimize the number of "bad" misclassifications, to exploit the social context to reduce misclassifications further (such as soothing verses neutral), and to provide expressive feedback to the caregivers so they can make sure that the robot properly "understood" their intent. By incorporating expressive feedback, we have already observed some intriguing social dynamics that arise with naïve female subjects. We intend to investigate these social

dynamics further so that we may use them to advantage in instructional scenarios.

To provide the human instructor with greater precision in issuing vocal feedback, we will need to look beyond *how* something is said to *what* is said. Since the underlying speech recognition system (running on the Linux machine) is speaker independent, this will boost recognition performance for both males and females. It is also a fascinating question of how the robot could *learn* the valence and arousal associated with particular utterances by bootstrapping from the correlation between those phonemic sequences that show particular persistence during each of the four classes of affective intents. Over time, Kismet could associate the utterance "Good robot!" with positive valence, "No, stop that!" with negative valence, "Look at this!" with increased arousal, and "Oh, it's ok." with decreased arousal by grounding it in an affective context and Kismet's emotional system. Developmental psycholinguists posit that human infants learn their first meanings through this kind of affectively-grounded social interaction with caregivers (Stern et al., 1982). Using punctuated words in this manner gives greater precision to the human caregiver's ability to issue reinforcement, thereby improving the quality of instructive feedback to the robot.

12. Conclusions

Human speech provides a natural and intuitive interface for both communicating with humanoid robots as well as for teaching them. We have implemented and demonstrated a fully integrated system whereby a humanoid robot recognizes and affectively responds to praise, prohibition, attention, and comfort in robot-directed speech. These communicative intents are well matched to human-style instruction scenarios since praise, prohibition, and directing the robot's attention to relevant aspects of a task, could be intuitively used to train a robot. Communicative efficacy has been tested and demonstrated with the robot's caregivers as well as with naïve subjects. We have argued how such an integrated approach lends robustness to the overall classification performance. Importantly, we have discovered some intriguing social dynamics that arise between robot and human when expressive feedback is introduced. This expressive feedback plays an important role in facilitating natural and intuitive human-robot communication.

Acknowledgments

This work was funded by a DARPA MARS grant BAA-9909. The authors gratefully acknowledge Jim Glass and Lee Hetherington for their assistance in porting the Spoken Language Group's speech recognizer to Kismet. We would like to thank Malcolm Slaney and Interval Research for allowing us to use their data base in earlier implementations. Paul Fitzpatrick was of tremendous assistance in helping us to integrate the many computers and processes running on Kismet. Roz Picard gave helpful insights in early discussions of this work.

Notes

1. This software was developed at MIT by the Spoken Language Systems Group.
2. The phoneme information is not currently used in the recognizer.

References

Blumberg, B. 1996. Old tricks, new dogs: Ethology and interactive creatures. Ph.D. Thesis, MIT.

Breazeal, C. 1998. A motivational system for regulation human-robot interaction. In *Proceedings of AAAI98*, pp. 54–61.

Breazeal, C. and Scassellati, B. 1999. How to build robots that make friends and influence people. In *Proceedings of IROS99*, pp. 858–863.

Breazeal, C., Edsinger, A., Fitzpatric, P., and Scassellati, B. 2000. Social constraints on animate vision. In *Proceedings of the 1st International Conference on Humanoid Robots*, Cambridge, MA.

Breazeal, C. and Foerst, A. 1999. Schmoozing with robots: Exploring the original wireless network. In *Proceedings of Cognitive Technology (CT99)*, pp. 375–390.

Breazeal, C. 1999. Robot in society: Friend or appliance? In *Proceedings of Agents99 Worshop on Emotion Based Architectures*, pp. 18–26.

Breazeal, C. and Scassellati, B. 2000. Challenges in building robots that imitate people. *Imitation in Animals and Artifacts*, MIT Press.

Bullowa, M. 1979. *Before speech: The Beginning of Interpersonal Communication*, Cambridge University Press: Cambridge, London.

Cahn, J. 1990. Generating expression in synthesized speech. Master's Thesis, MIT Media Lab.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. 1994. Animated conversation: Rule-based generation of facial expression, gesture, and spoken intonation for multiple conversational agents. In *SIGGRAPH*.

Chen, L. and Huang, T. 1998. Multimodal human emotion/expression recognition. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*.

Damasio, A.R. 1994. *Descartes's Error: Emotion, Reason, and the Human Brain*. Gosset/Putnam Press: New York, NY.

Dellaert, F., Polzin, F., and Waibel, A. 1996. Recognizing emotion in speech. In *Proceedings of the ICSLP*.

Eibl-Eibelsfeld, I. 1970. *Liebe und Hass: Zur Naturgeschichte elementarer Verhaltensweisen*, Piper: Munich, Germany.

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. *Infant Behavior and Development*, 8:181–195.

Ferrier, L.J. 1987. Intonation in discourse: Talk between 12-month-olds and their mothers. In *Children's Language*, Vol. 5, K. Nelson (Ed.), Erlbaum: Hillsdale, NJ, pp. 35–60.

Grieser, D.L. and Kuhl, P.K. 1988. Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, 24:14–20.

McRoberts, G., Fernald, A., and Moses, L. in press. An acoustic study of prosodic form-function relationships in infant-directed speech: Cross language similarities. *Development Psychology*.

Murray, I.R. and Arnott, L. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal Acoustical Society of America*, 93(2):1097–1108.

Nakatsu, R., Nicholson, J., and Tosa, N. 1999. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *ICMCS*, Vol. 2, pp. 804–808.

Papousek, M., Papousek, H., and Bornstein, M.H. 1985. The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech. In *Social Perception in Infants*, T. Field and N. Fox (Eds.), Ablex: Norwood, NJ, pp. 269–297.

Reeves, B. and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, CSLI Publications: Stanford, CA.

Roy, D. and Pentland, A. 1996. Automatic spoken affect classification and analysis. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 363–367.

Slaney, M. and McRoberts, G. 1998. Baby ears: A recognition system for affective vocalizations. In *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA.

Snow, C.E. 1972. Mother's speech to children learning language. *Child Development*, 43:549–565.

Stern, D.N., Spieker, S., and MacKain, K. 1982. Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, 18:727–735.

Velasquez, J. 1998. When robots weep: A mechanism for emotional memories. In *Proceedings of AAAI98*.

Vlassis, N. and Likas, A. 1999. A kurtosis-based dynamic approach to Gaussian mixture modeling. *IEEE Trans. on Systems, Man, and Cybernetics. Part A: Systems and Humans*, 29(4):393–399.

Yoon, S.Y., Blumberg, B., and Schneider, G. 2000. Motivation driven learning for interactive synthetic characters. In *Proceedings of Agents*.

Cynthia Breazeal directs the Robotic Presence Group at the MIT Media Lab. She has developed numerous autonomous robots, from planetary micro-rovers, to upper-torso humanoid robots, to highly expressive robotic faces. Always inspired by the behavior of living systems, scientific models and theories as well as artistic insights factor heavily into the hardware and software design of her robotic creations. Her current interests focus on social interaction and socially situated learning between people and life-like robots. She carried out her graduate work at the MIT Artificial Intelligence Lab, and received her Sc.D. and S.M. degrees from MIT in the department of Electrical Engineering and Computer Science with specialization in robotics and artificial intelligence.

Lijin Aryananda received her B.S. (1998) and M.Eng (1999) in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. She is currently pursuing a Ph.D. at the MIT Artificial Intelligence Laboratory. Her research interests include robotics, human-robot interaction, individual recognition, social intelligence, and autobiographical memory.