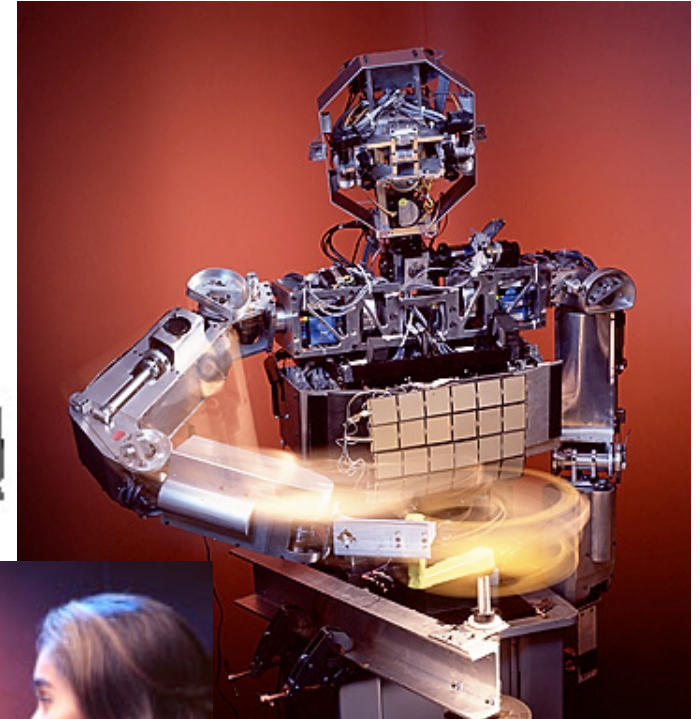
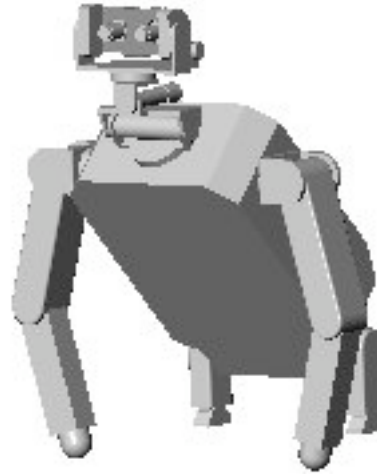




Natural Tasking of Robots Based on Human Interaction Cues--MIT AI Lab

- **Challenge:** Make commanding robots as intuitive and natural as commanding soldiers
- We will build software systems (on 4 robot platforms) that
 - interact with people using intuitive social cues
 - are self-adapting through developmental learning
 - exploit their physical embodiment to simplify computation
 - use multi-modal sensory integration to maximize effectiveness





Research Issues: Learning by Imitating Humans

- **Knowing what to imitate**
 - what sensory signals are relevant to the current task?
- **Mapping between bodies**
 - how to convert sensory view of one agent to self-mapping?
- **Correcting failures and recognizing success**
 - how can the robot automatically develop success measures?
- **Chaining actions together**
 - how can the learned pieces be chained together into complete task sequences?
- **Generalizing to more complex tasks**
 - how to use invariants and common-sense knowledge to adapt the learned sequences to new situations
- **Intuitive interactions for the human**
 - how to use natural social signals to make it intuitive for humans and robots to communicate with each other



Our Approaches Lead to Specific Progress on Each Research Issue

Research Issues

Our Approaches

	Knowing what to imitate	Mapping between bodies	Correcting failures and recognizing success	Chaining actions together	Generalizing to more complex tasks	Intuitive Interactions
Social Interaction	Uses attentional cues to recognize task relevant events		Identifies and displays emotional states for reinforcement			Allows humans and robots to share similar social cues
Development	Limits search space by incremental refinement of perception	Simplifies mapping by incremental refinement of motor skills		Provides natural decomposition of complex tasks	Exploits incremental learning to build new skills	Provides methods for building social skills
Embodiment	Constrains movement which assists directed perception	Simplifies body mapping between human and robot	Enables simple but robust low-level behaviors	Places physical limits on successive actions		Allows the human to observe natural social feedback cues
Integration	Allows robot to recognize cues in multiple modalities		Increases robustness through redundancy		Assists in transfer to new modalities	Allows human to use natural modalities (voice, gesture, facial expressions)



The Research Team



May 23, 2000

DARPA Mobile Autonomous Robot Software BAA99-09



Main Reference for Today

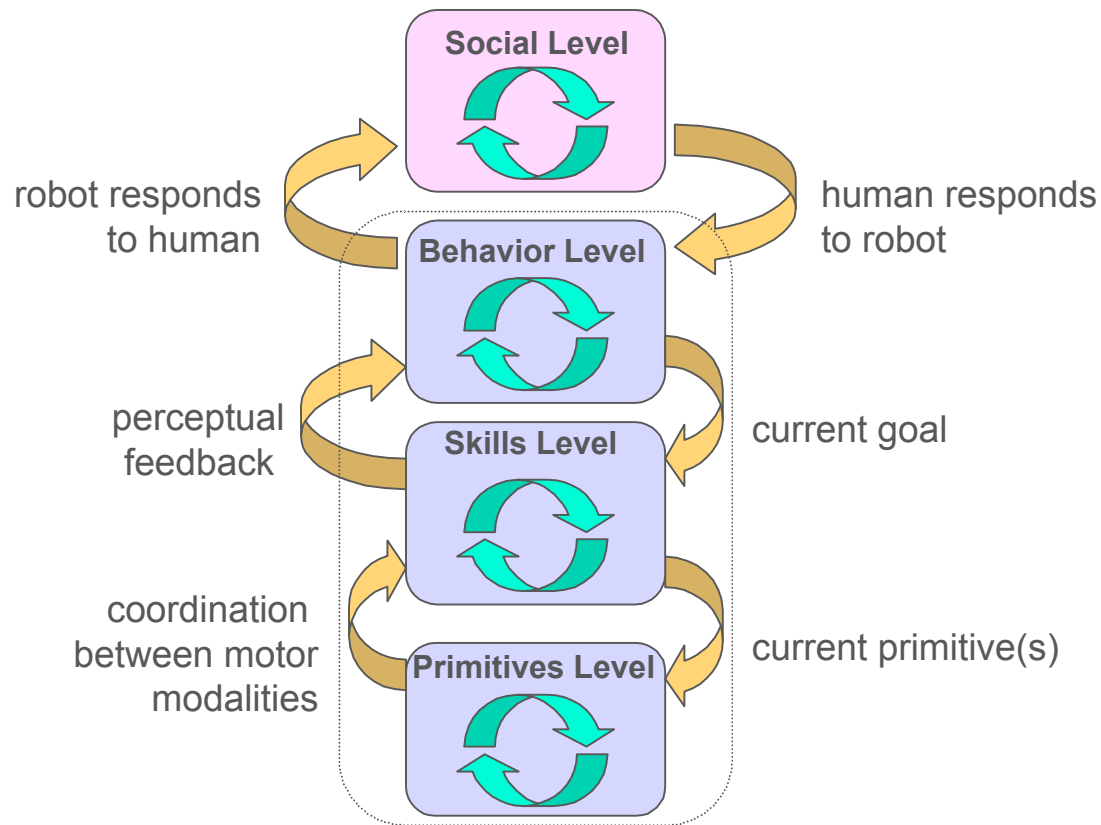
Cynthia Breazeal

**“Sociable Machines: Expressive Social
Exchange between Humans & Robots”**

D.Sc. thesis, May 19, 2000



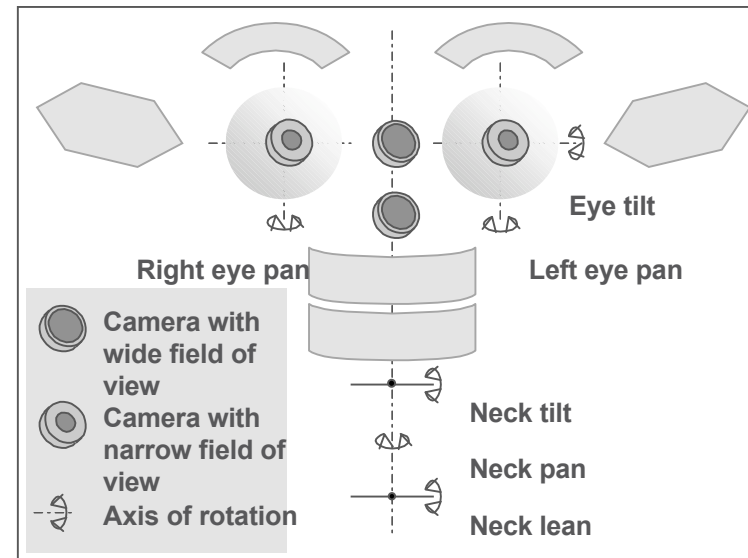
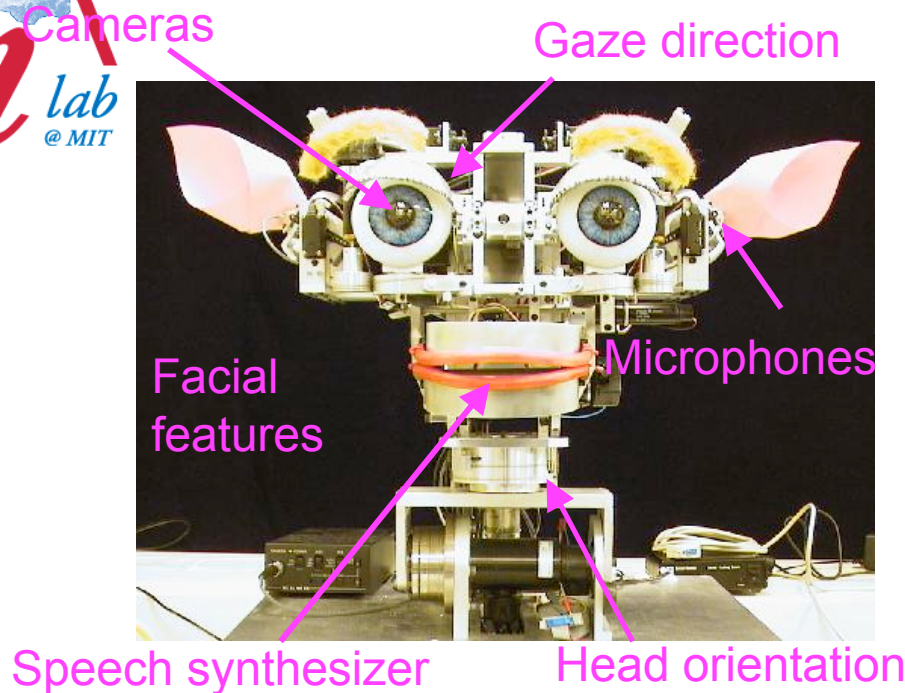
Levels of Control for Sociable Robots



- Observable behavior generated through multiple levels of interaction
- Each level has different constraints
 - temporal
 - sensory feedback
 - information flow
- Social level adds significant constraints to control
 - Robots actions have communicative value and social significance
 - Robot's actions have the ability to manipulate the environment *far beyond their immediate physical consequence*
- Design to benefit interaction for robot and human
 - matched to robot's perceptual, physical, and computational limitations
 - Interaction is natural, intuitive, and enjoyable for person



Designing for the Human In the Loop



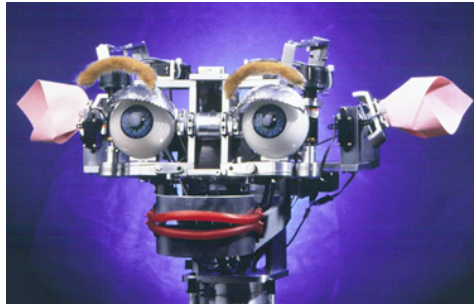
- **Social interaction with person constraints**
 - how robot appears
 - how its sensors are configured
 - its quality of movement
 - how it behaves, interacts

- **Support perception of human social cues**
 - vision, auditory
- **Display human-like social cues**
 - gaze direction, body posture
 - gesture, facial expression
 - vocal acts



perception,
attention,
motivation,
behavior,
expression

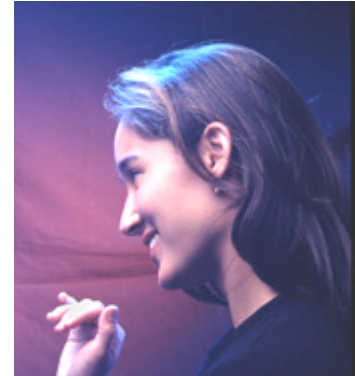
Design for Human-Robot System



tune robot
to human



tune human
to robot



evolution,
development,
culture

- **Design to support human nature**
 - look at psychology, ethology, evolution, social development of infants
 - Naturally elicit desired complexity of interaction from humans
 - Match to human perceptions, human attention, human social behavior
- **Audience perception**
 - The design should establish the right social expectations
 - People naturally engage robot at right level of interaction
 - “believability” – treat robot as socially aware
 - “readability” – what person observes and interprets accurately reflects robot’s internal state



Four Competencies to Support Social Interaction and Learning for Kismet

- **Direct Visual Attention**
 - indicate saliency
 - organize behavior about a locus of attention
 - robot should respond to natural attention directing cues of people
 - people should know what Kismet is attending to
 - useful for establishing shared reference
- **Recognize Socially Communicated Reinforcement**
 - voice as a natural instrument to convey praise, prohibition, etc.
 - robot should recognize affective feedback in robot-directed speech
 - people should know when Kismet understands their intent
 - useful as progress estimator, signal for goal attainment



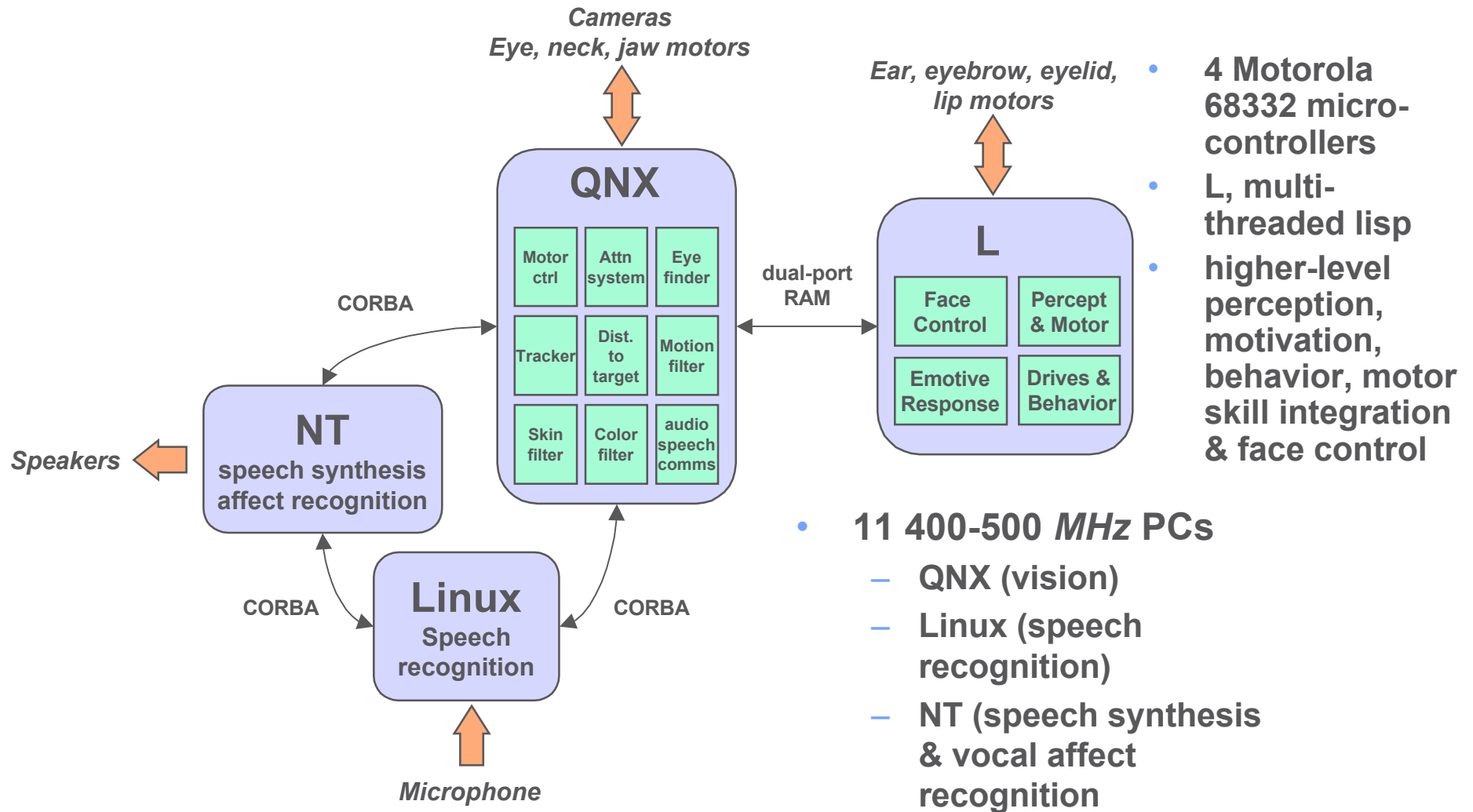
Four Competencies to Support Social Interaction and Learning for Kismet

- **Communicate Internal State to Human**
 - important cue for human to predict and understand robot's behavior
 - important cue for human to tune their own behavior to robot
 - robot should convey its motivational and behavioral state to human in an intuitive manner
 - can be used to establish better quality instruction
- **Regulation of Social Interaction**
 - turn-taking is cornerstone of social interaction and social instruction
 - keep robot from becoming overwhelmed or under-stimulated.
 - structures learning episodes, interactive games, variations upon a theme



Large Scale Integration

(15 networked computers)



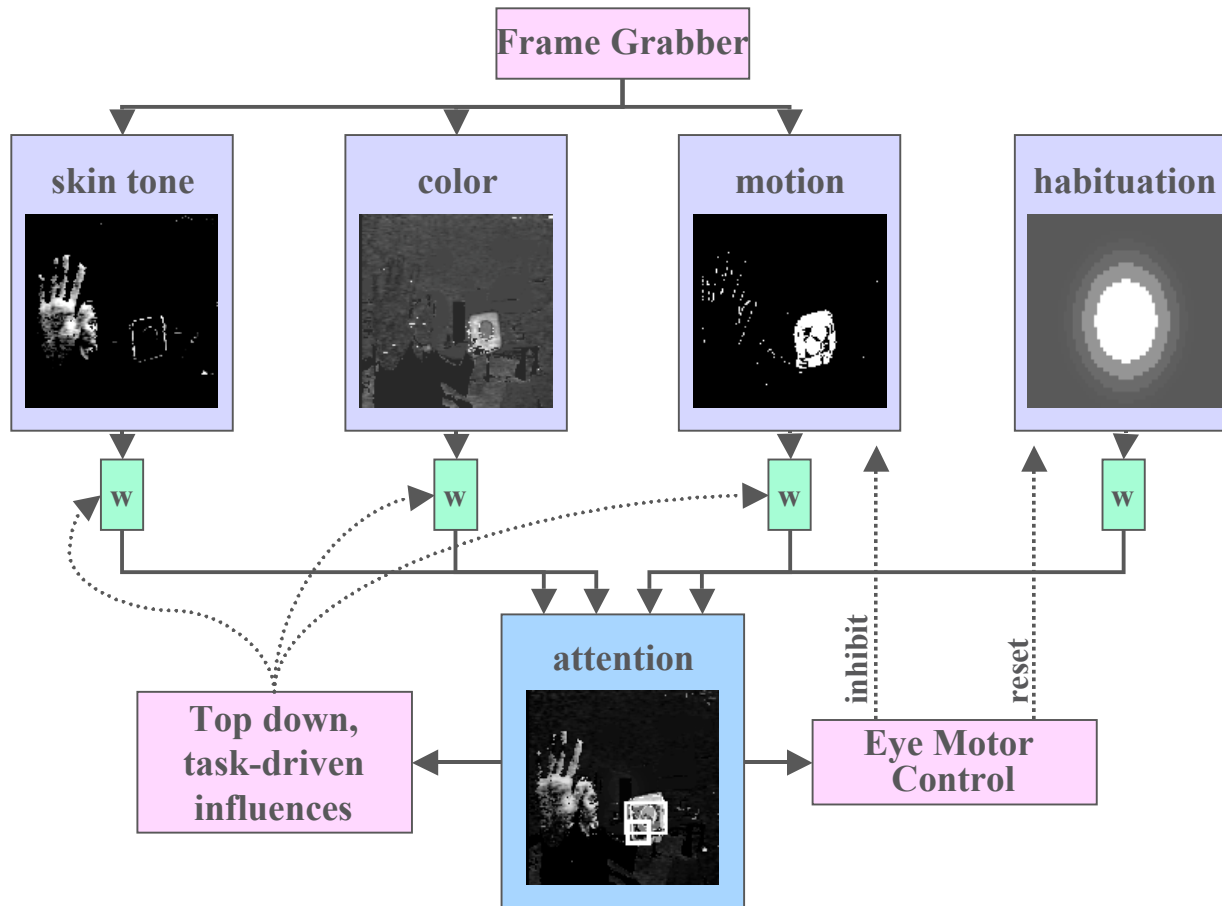


Evaluation Metrics

- **Design for natural, intuitive, social interactions**
- **Must consider human response as well as robot's behavior**
 - Subjective but quantifiable
- **Four Criteria used to evaluate Kismet:**
 - Can people intuitively read and do they naturally respond to Kismet's social cues?
 - Does Kismet readily elicit social cues (scaffolding) from the human that could be used to benefit learning?
 - Can Kismet perceive and appropriately respond to these naturally offered social cues?
 - Does the human adapt to the robot, and robot adapt to the human, in a way that benefits the interaction? (entrainment)



Visual Attention.

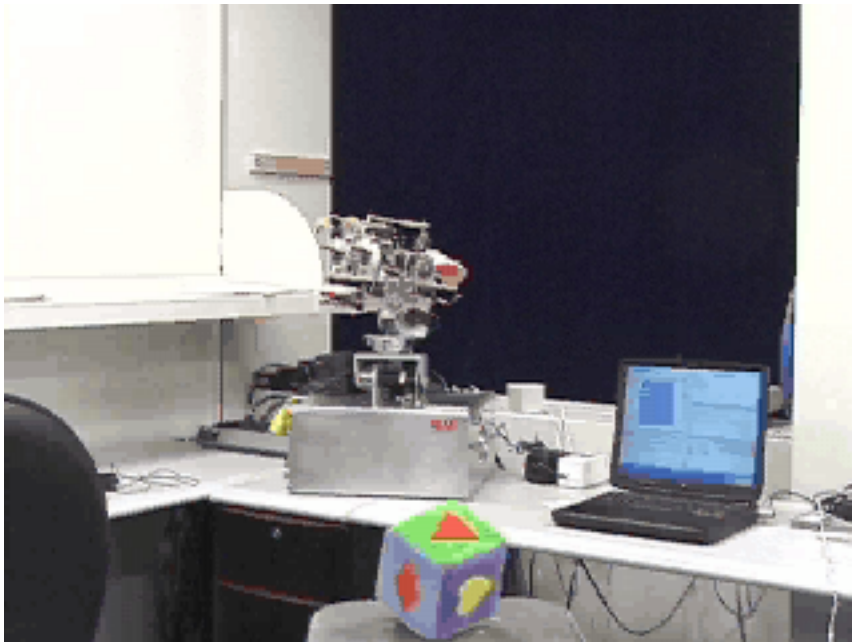


Adapted from psychological model for human attention and visual search:

- **Wolfe's VGS 2.0**
- For robot, extended model:
 - animate vision,
 - habituation,
 - time-varying goals
- Provides locus of attention within complex visual environment to organize behavior
- Human and robot should find same sorts of features salient
- Features need to match motivational requirements
- Robot's gaze direction is an important teaching cue

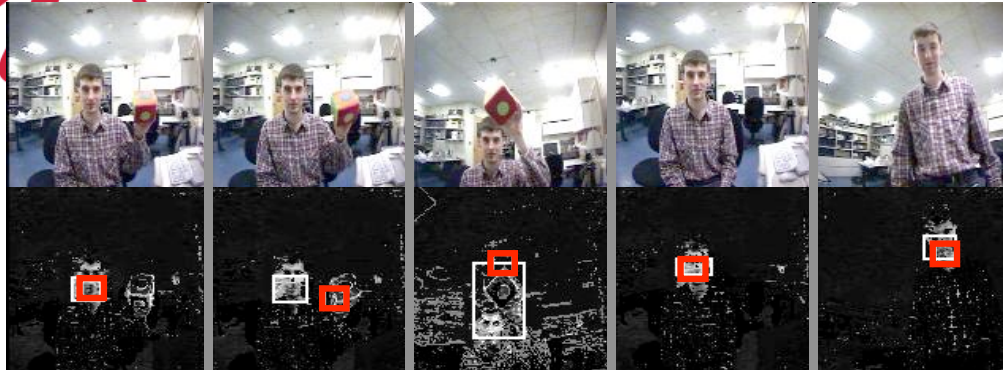


Video of Visual Search for Toy



- **Robot's search is task-specific**
 - seeks out desired stimuli
 - still opportunistic when appropriate
- **Visual behavior conveys degree of commitment**
 - fleeting glances
 - smooth pursuit
 - full body orientation
- **Gaze direction, facial expression, and body posture convey robot's interest**
- **The attention and visual systems are always running and responding to external events**
 - it's possible to divert attention by layering saliency features

Directing Kismet's Attention and Gaze

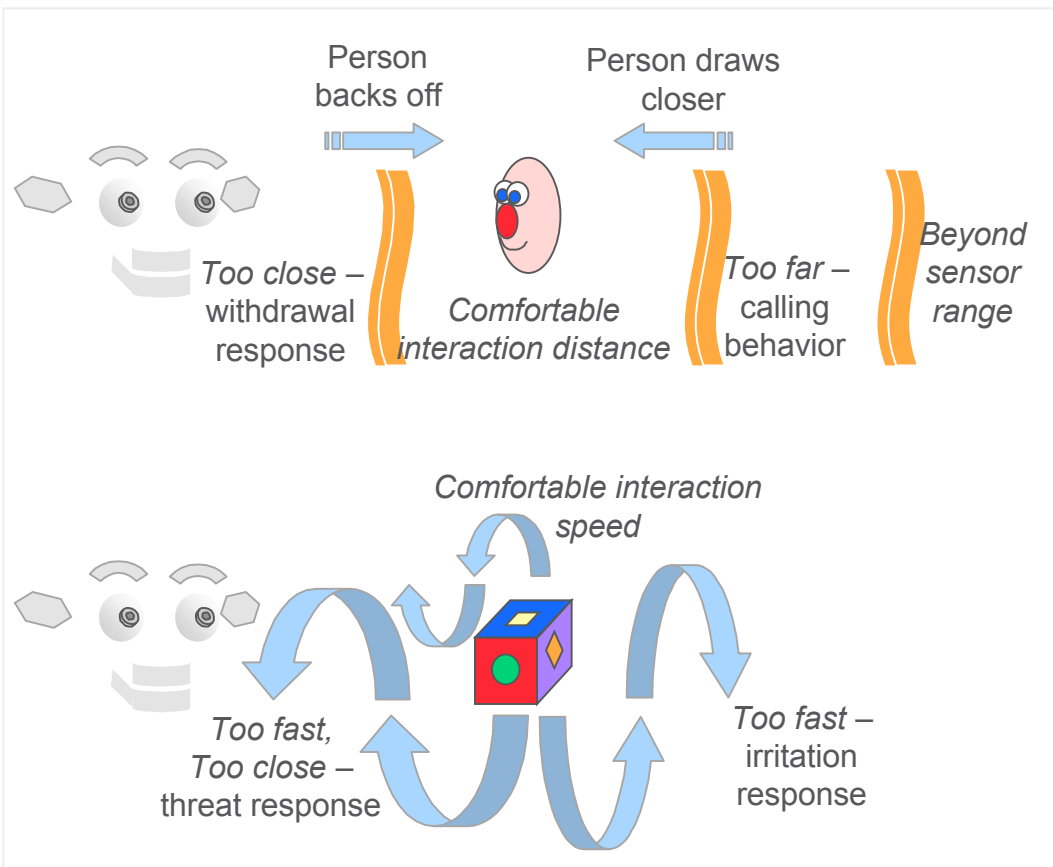


stimulus category	stimulus	presentations	average time (s)	commonly used cues	commonly read cues
color & movement	yellow dinosaur	8	8.5	motion across centerline	change in visual behavior
	multi-colored block	8	6.5		
	green cylinder	8	6.0	shaking	face reaction
motion only	b/w cow	8	5.0	bringing object close	body posture
skin-toned & movement	pink cup	8	6.5		
	hand	8	5.0		
	face	8	3.5		
Total		56	5.8		

- Wide field of view tracker follows the most salient target
- Gaze directs foveal cameras to target and eyes smoothly pursue target.
- Can overlay cues to make object more salient to robot
- These cues are naturally the ones people use to direct attention



Social Constraints on Animate Vision



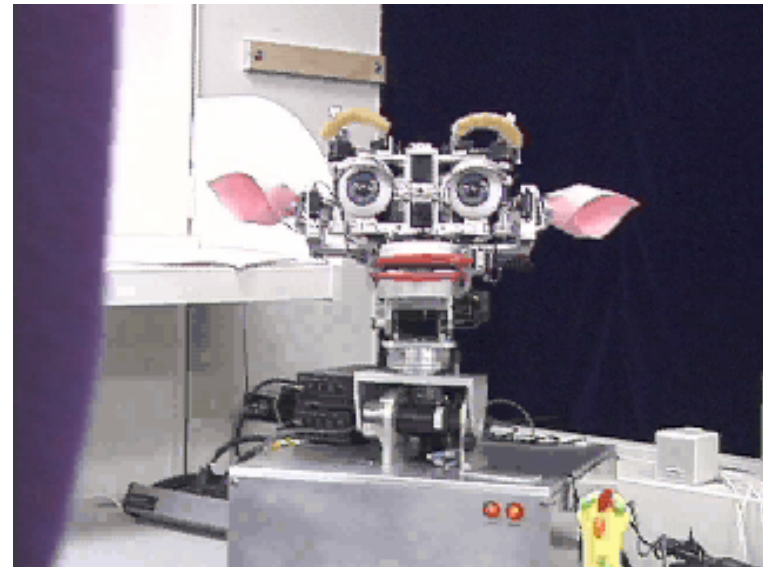
- Robot's visual behavior, facial expressions, and body posture have communicative value to the human observer
- The robot's signals can be used to **regulate** the nature of the interaction
 - To benefit robot's behavior
 - To benefit robot's perceptual processing
- These signals must be intuitive to understand for the human observer
 - correlated to emotive responses for coping with unfavorable stimuli
 - » fear, disgust, frustration, distress
 - correlated to emotive responses for bringing robot near favorable stimuli
 - » interest, joy



Video of Social Amplification



establish personal space



lure people in

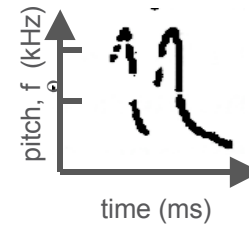
- **naive subjects**
 - robot establishes a “personal space” through expressive cues
 - regulates persons interaction to suit its visual processing limitations



Infant Recognition of Vocal Affective Intent (Anne Fernald)

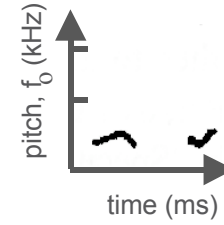
- **Four cross-cultural contours of infant-directed speech (A. Fernald)**
 - praise,
 - prohibition,
 - soothing,
 - bids for attention
- **Exaggerated prosody is well matched to infant's innate responses to auditory stimuli.**
 - modulates infant's arousal and attention
 - social referencing

That's a good bo-o-y!



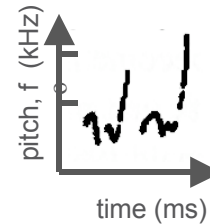
approval

No no baby.



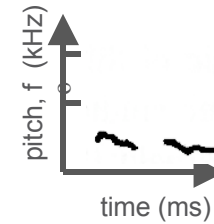
prohibition

Can you get it? Can you get it?



attention

MMMM Oh, honey.

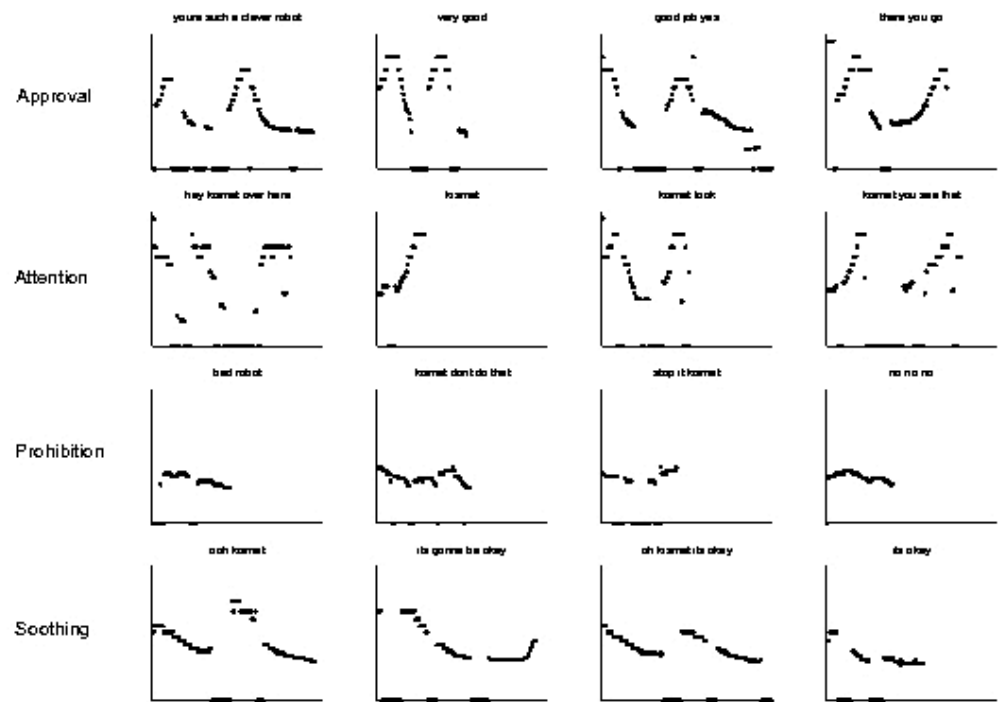


comfort



Vocal Affective Intent as Teaching Signal

- Adapt Fernald's research to Kismet
- Ultimately, use as training signal (robot does not learn yet)
 - Real-time recognition
 - Use as affective **saliency** marker (designed for *strong intent*)
- Affective classification modulates robot's affective state
 - could be used as reinforcement
- Socially situate with expressive feedback

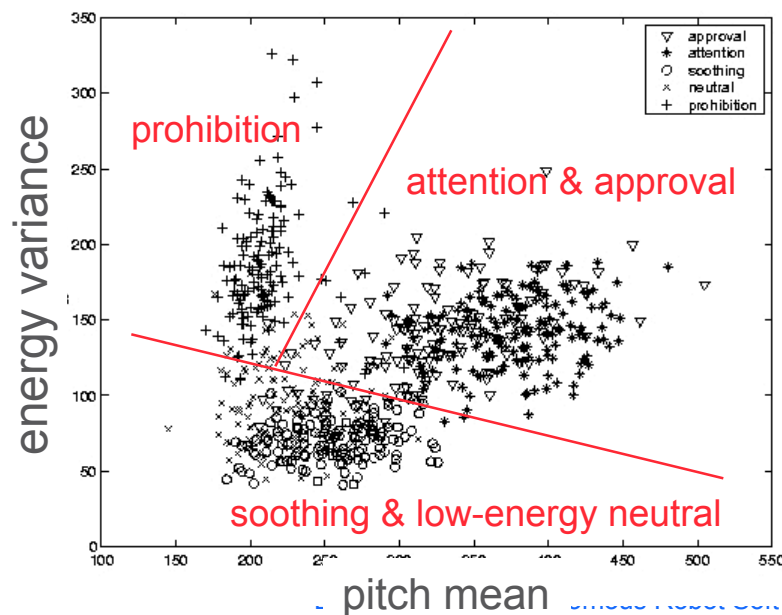
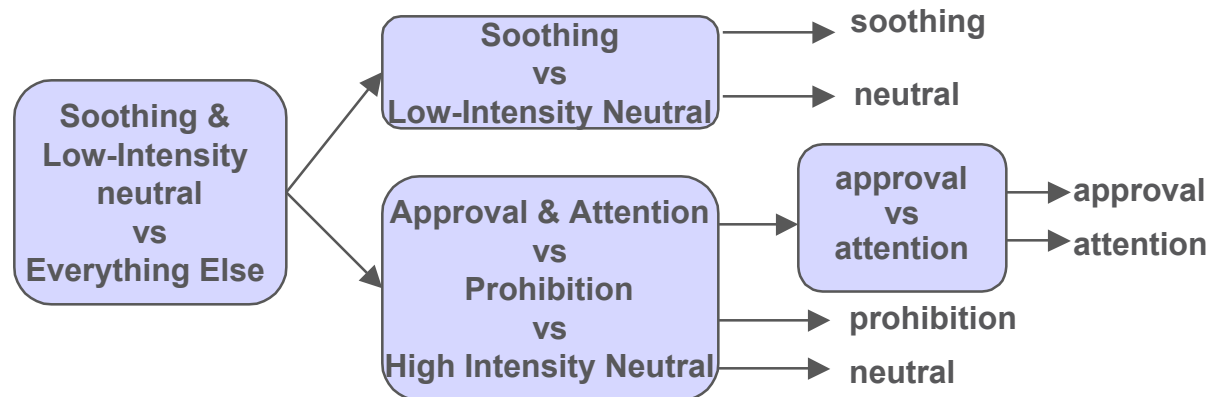


Evidence for 4 contours in
Kismet-directed speech



Recognition of Affective Intent in Robot-Directed Speech

- Database of small number of female speakers
 - Recorded in robot's environment
 - Robot-directed speech with strong affect
- Multi-stage model
 - Each stage is simple for real-time performance
 - Some classes easily separated by arousal
 - energy, pitch
 - Off-the-shelf learning mechanism for the stages (Mixture of Gaussian with EM)
- Use Fernald contour information to distinguish *difficult* classes
 - attention vs approval
 - soothing vs low intensity neutral





Recognizing Affective Intent

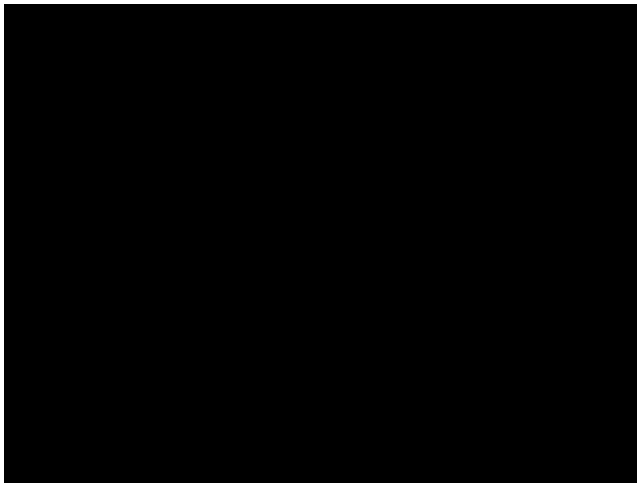
Results with Caregivers and Naive Subjects

Test set	Strength	Class	Test Size	Classification Result					% Correctly Classified
				Approval	Attention	Prohibition	Soothing	Neutral	
Caregivers		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

- **Five classes of utterances**
 - neutral speech
 - praise, prohibition, attention, soothing
- **All Female speakers**
- **Multiple languages**
 - French, German, Indonesian, English, Russian
- **Objective scorer classifies as strong, medium, weak**
- **Good overall performance for strong instances**
 - (random perf. = 20%)
 - very good for caregivers
 - good for naive subjects
- **Acceptable misclassifications**
 - minimal confusion of valence
 - some confusion of arousal



Evaluation with Naive Subjects

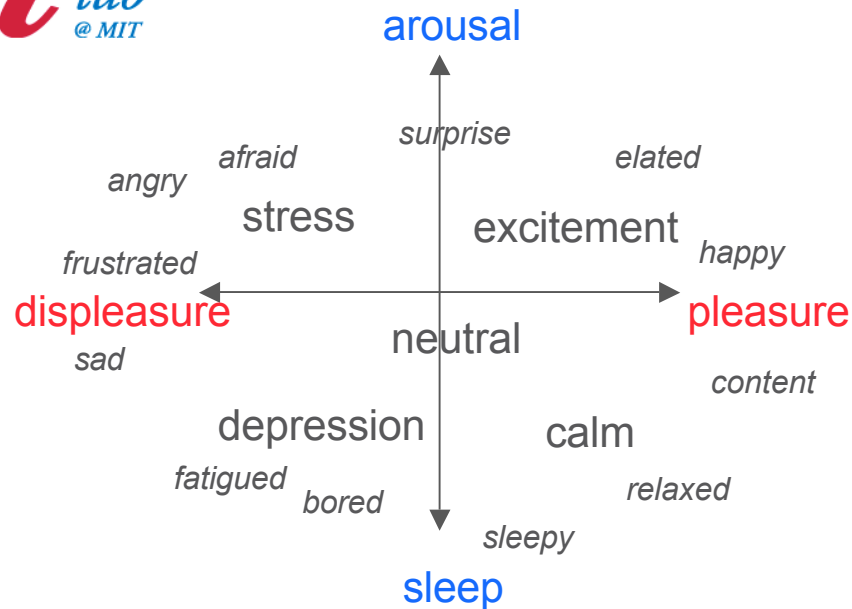


- 5 female subjects
 - 4 naive subjects
 - 1 caregiver
- Four contours and neutral speech
 - praise, prohibition, attention, soothing
- Multiple languages
 - French, German, Indonesian, English, Russian
- Driven by Human
- Socially situated with expressive feedback
 - expression easily read as signal of robot's "understanding"
 - Evidence of entrainment
 - affective mirroring
 - affective feedback



Facial Expressions

(Russell, Scott & Smith)

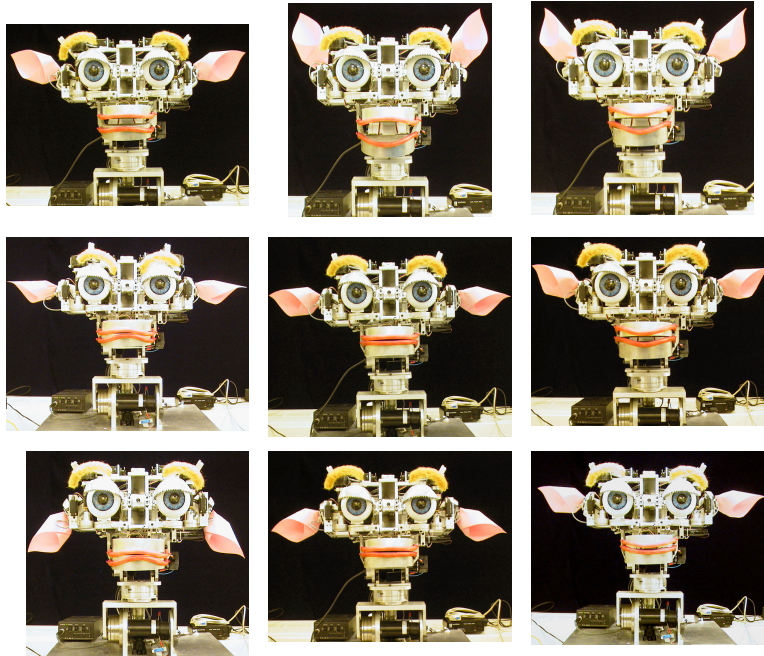


- Emotive Facial expressions are a social signal
- Is communicated message contained in facial components or entire emotive facial expression?

- Emotions can be mapped to affect dimensions (Russell)
- Facial postures are related in a systematic way to these affective dimensions (Smith & Scott)



Generating Affective Facial Expressions

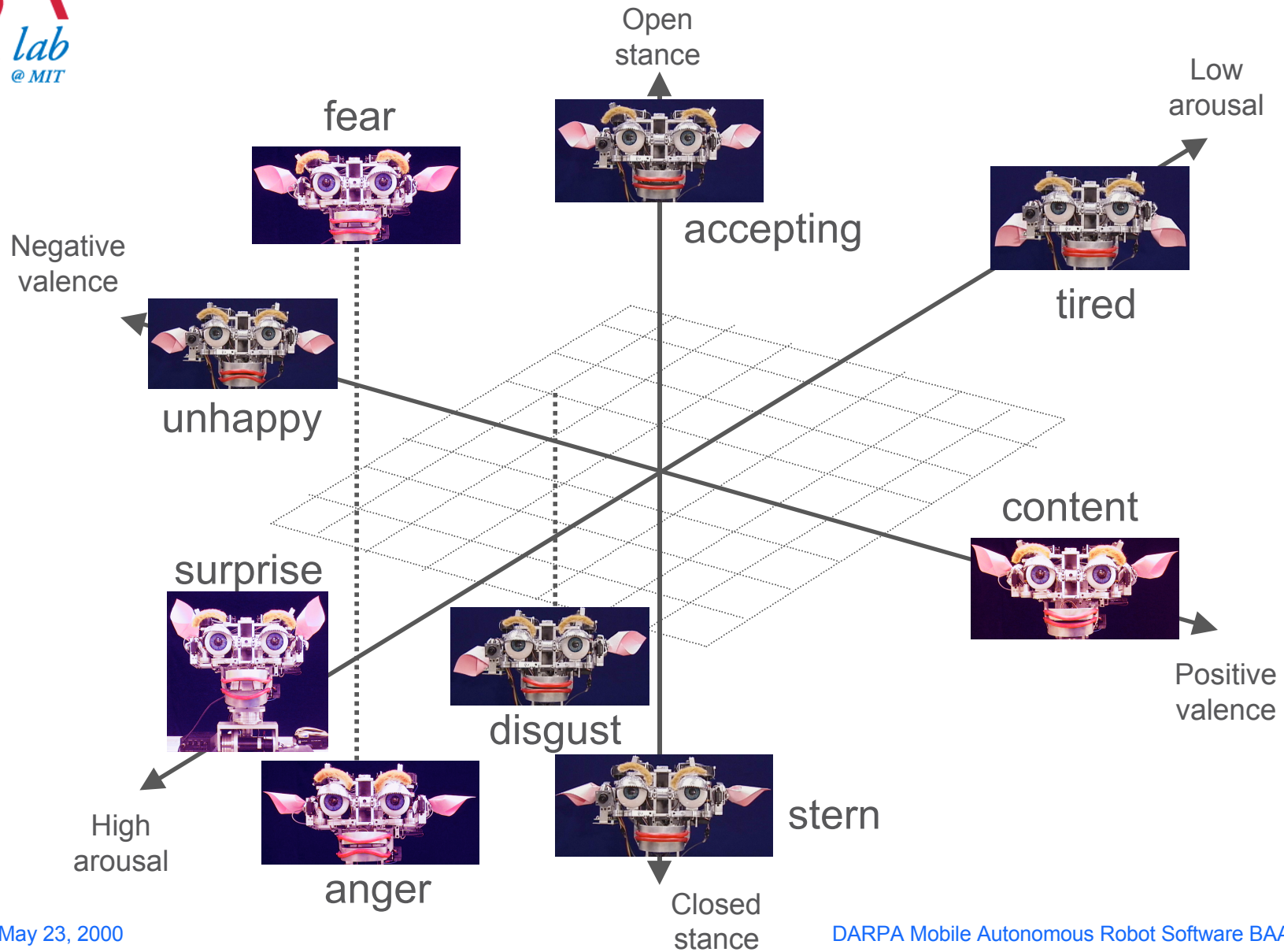


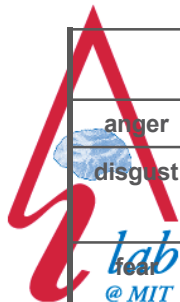
- **Emotive Facial expressions are a social signal**
- **Important Feedback to the human observer**
 - predictive of robot's behavior
 - indicative of robot's appraisal (future behavior)

- **“Affective tags” of appraisal phase map to affect Kismet's dimensions (A, V, S)**
- **Blended facial expression reflects the affective state of the robot**



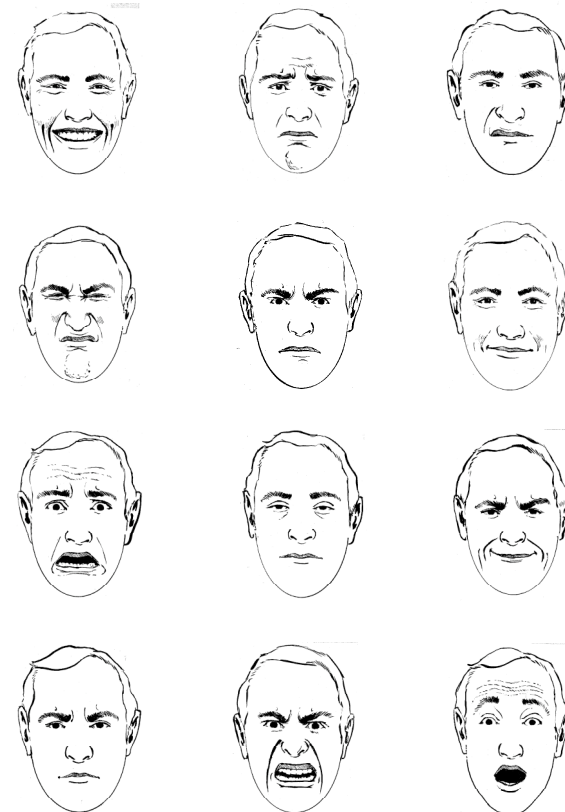
Mapping Facial Postures to Affect Space





	most similar sketch	percent age	comments
anger	anger	100%	shape of mouth and eyebrows are strongest reported cues
	disgust	80%	shape of mouth is strongest reported cue
	sly grin	20%	described as "sneering"
	fear	70%	shape of mouth and eyes are strongest reported cues. Mouth open "aghast"
	surprise	10%	subject associates look of "shock" with sketch of "surprise" over "fear"
	happy	20%	lip mechanics cause lips to turn up at ends, sometimes confused with a weak smile
joy	happy	70%	report lips and eyes are strongest cues. Ears may provide arousal cue to lend intensity.
	content	10%	report lips used as strongest cue
	repulsion	10%	lip mechanics turn lips up at end, causing shape reminiscent of lips in repulsion sketch
	surprise	10%	perked ears, wide eyes lend high arousal. sometimes associated with a pleasant surprise
sorrow	sad	90%	lips reported as strongest cue. Low ears may lend to low arousal.
	repulsion	10%	lip mechanics turn lips up and end, causing shape reminiscent of repulsion sketch
surprise	surprise	90%	reported open mouth, raised brows, wide eyes and elevated ears all lend to high arousal
	happy	10%	subject remarks on similarity of eyes, but not mouth
pleased	content	90%	reported relaxed smile, ears, and eyes lend low arousal and positive valence
	sly grin	10%	subject reports the robot exhibiting a reserved pleasure. Associated with the "sly grin" sketch
sly grin	sly grin	50%	lips and eyebrows reported as strongest cues
	content	30%	subjects use robot's grin as the primary cue
	stern	10%	subject report's the robot looking "serious", which is associated with "sly grin" sketch
	repulsion	10%	lip mechanics curve lips up at end. Subject sees similarity with lips in "repulsion" sketch
stern	stern	60%	lips and eyebrows are reported as strongest cues
	mad	10%	subject reports robot looking "slightly cross". Cue on robot's eyebrows and pressed lips.
	tired	20%	subjects may cue in on robot's pressed lips, low ears, lowered eyelids
	sly grin	10%	subject reports similarity in brows.

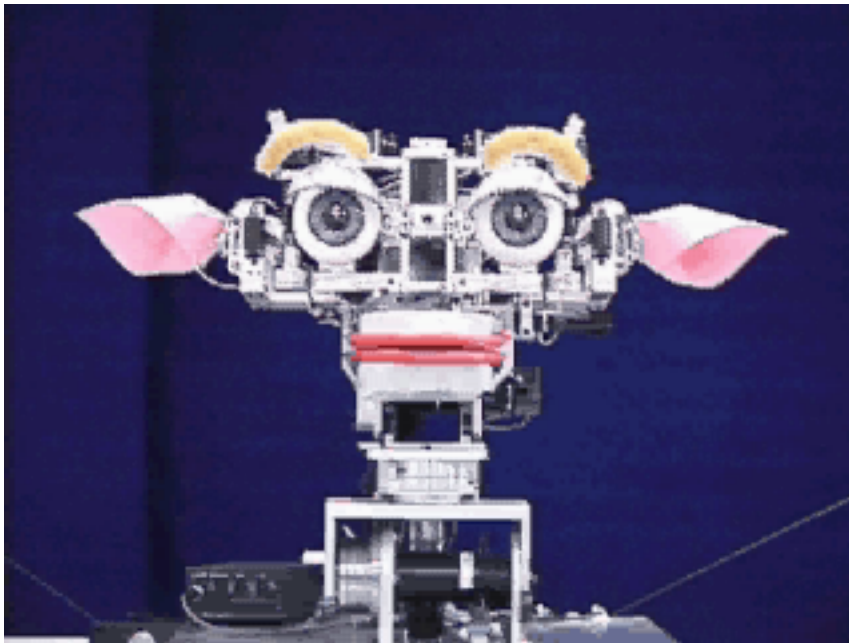
Evaluation of Emotive Expressions



random chance = 8%



Video of Emotionally Expressive Face, Voice, Posture

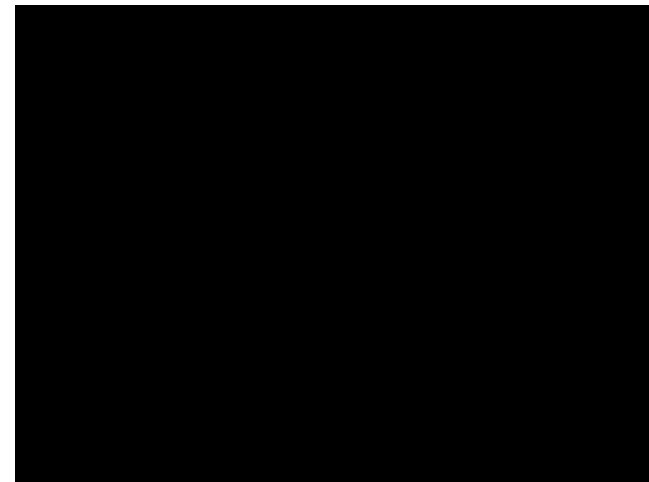


- Evaluated Kismet's expressive cues on naive subjects
 - Ranging in age 12—54
 - 9—17 subjects per study
- Human sketch comparisons
 - 70%--90% for “basic” emotion
- Still images, forced choice
 - 47%--83%
- Video of posture&expression, forced choice
 - 57%--86%
- Pre-recorded vocalizations, forced choice
 - (babbles and English phrases)
 - 50%--70% recognition
 - except “surprise” often confused for “fear”



Turn-Taking and Proto-Dialog

- Already talked about social amplification (slower time scales)
- Cornerstone of human-style communication, learning, and instruction
- Turn-taking is fine grained regulation of human's behavior
 - use of envelope displays
 - shifts of gaze
 - facial expressions
 - shifts of body posture
 - Tightly coupled dynamic of contingent responses to other
- 4 phases of turn cycle
 - relinquish floor
 - listen to speaker
 - reacquire floor
 - speak
- Integrates
 - visual behavior & attention
 - facial expression & animation
 - body posture
 - vocalization & lip synchronization



Edited down from long interactions
10-15mins



Evaluation of Performance

- **Naïve subjects**
 - ranging in age from 25 to 28
 - All young professionals.
 - No prior experience with Kismet
 - video recorded
- **Turn-taking performance**
 - 82% “clean” turn transitions
 - 11% interruptions
 - 7% delays followed by prompting
- **Significant flow disturbances**
 - tend to occur in clusters
 - 6.5% of the time, but rate diminishes
- **Evidence for entrainment**
 - shorter phrases
 - wait longer for response
 - read turn-taking cues
 - 0.5—1.5 seconds between turns

		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+

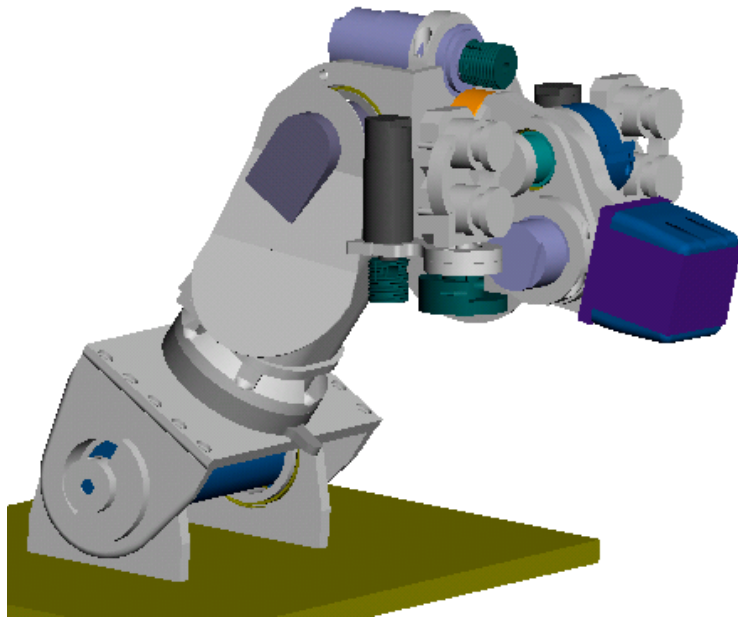


Relevance to Future Combat Systems

- We are trying to make it intuitive to interact with robots
- Currently examining social coupling in an extreme situation--parent/child
- Will take lessons learned there, and apply it to more operational scenarios
- Human/animal is one metaphor
- Human/subordinate human is another



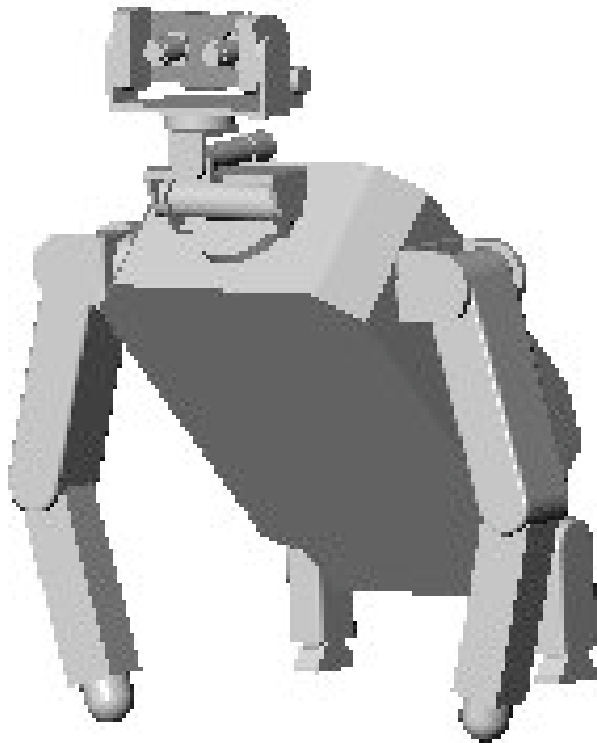
M4, A Dog-like Robot



May put head on a two legged robot



Coco, a Walking Gorilla



Expressive head
Arm gestures



May 23, 2000

DARPA Mobile Autonomous Robot Software BAA99-09