Sociable Machines: Expressive Social Exchange Between Humans and Robots

by

Cynthia L. Breazeal

B.S. Electrical and Computer Engineering
University of California at Santa Barbara, 1989
S.M. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 1993

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR of SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2000

Signature of Author ....
Department of Electrical Engineering and Computer Science
May 19, 2000

Certified by..
Rodney Brooks
Fujitsu Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Sociable Machines: Expressive Social Exchange Between Humans and Robots

by

Cynthia L. Breazeal

Submitted to the Department of Electrical Engineering and Computer Science on May 19, 2000 in Partial Fulfillment of the Requirements for the Degree of Doctor of Science in Electrical Engineering and Computer Science

ABSTRACT

Sociable humanoid robots are natural and intuitive for people to communicate with and to teach. We present recent advances in building an autonomous humanoid robot, Kismet, that can engage humans in expressive social interaction. We outline a set of design issues and a framework that we have found to be of particular importance for sociable robots. Having a human-in-the-loop places significant social constraints on how the robot aesthetically appears, how its sensors are configured, its quality of movement, and its behavior.

Inspired by infant social development, psychology, ethology, and evolutionary perspectives, this work integrates theories and concepts from these diverse viewpoints to enable Kismet to enter into natural and intuitive social interaction with a human caregiver, reminiscent of parent-infant exchanges. Kismet perceives a variety of natural social cues from visual and auditory channels, and delivers social signals to people through gaze direction, facial expression, body posture, and vocalizations.

We present the implementation of Kismet's social competencies and evaluate each with respect to: 1) the ability of naive subjects to read and interpret the robot's social cues, 2) the robot's ability to perceive and appropriately respond to naturally offered social cues, 3) the robot's ability to elicit interaction scenarios that afford rich learning potential, and 4) how this produces a rich, flexible, dynamic interaction that is physical, affective, and social. Numerous studies with naive human subjects are described that provide the data upon which we base our evaluations.

Thesis supervisor: Prof. Rodney A. Brooks
Title: Fujitsu Professor of Computer Science and Engineering

Acknowledgements

I remember seeing *Star Wars* as a little girl. I remember being absolutely captivated and fascinated by the two droids *R2D2* and *C3P0*. Their personalities and their antics made them compelling characters, far different from typical sci-fi robots. I remember the heated debates among my classmates about whether the droids were real or not. Some would argue that because you could see the wires in *C3P0's* abdomen that it must be a real robot. Alas, however, the truth was known. They weren't real at all. They existed only in the movies. I figured that I would never see anything like those two droids in my lifetime.

Many years later I would find myself at MIT in the robotics lab of Prof. Rod Brooks. He told me of autonomous robots, of their biological inspiration, all very insect-like in nature. I remember thinking to myself that this was it – these kinds of robots were the real-life pre-cursers to the droids of my childhood. I knew that this was the place for me.

I was initiated into the lab on a project that was something like baptism by fire. Serious second system syndrome for the successor to *Genghis*, the first autonomous insect-like robot from our lab. Six legs, eight microcontrollers, and sixty sensors later I was confronted with my first autonomous robot, *Hannibal*. I was responsible for "breathing the life" into it. A year and a half later, I finished my M. S. on the little critter.

It was time to start my Ph.D., and foolish me, I wanted to do something different. So, Rod threw *Cog* on my plate. *Cog* is the first general purpose autonomous humanoid robot created by our group to explore theories of human intelligence. But in many ways, building *Cog* was not unlike building *Hannibal*. Just lots more of everything. Lots more. *Cog* is also big enough to do some damage if you're not careful. Fortunately it couldn't locomote, but I could see the wisdom in building robots that you could pick up. Anyways, everyone in the group was building *Cog*, and building *Cog*, and building *Cog*, and building...

Then one day, we ran both *Cog's* arm control with its active vision system. The two didn't communicate at all. The arm would perform a reach and grasp maneuver, again, and again, and again. In the meantime the active vision system would orient *Cog's* eyes and head towards a moving stimulus. I intuitively played a simple little "grab-the-eraser" game with *Cog*. I would wiggle an eraser on the table top, the eyes and head would look at it, and the arm would perform the reach and grasp. *Cog* would of course miss because there was no visual feedback to the reaching code. But it didn't really matter. To an outside observer, it looked like *Cog* was engaged in a game of trying to pick up the eraser, and I was helping it to do so.

I realized at that point that social interaction and playing in a human-like way with these autonomous robots was "lower hanging fruit" than I had thought. But

Cog was big and intimidating. I wanted a robot that I could treat as a very young child, that I could play with, and in doing so could teach it about its world and about people. I wanted to build a robot that could learn and develop like an infant. A robot raised in human culture. A robot grounded in social interaction. Pretty radical, even for science fiction. Soon thereafter, the idea of *Kismet* was born.

Three years and fifteen computers later, *Kismet* is a dramatically different kind of autonomous robot than any other, real or fictional. It's unlike *HAL*, or *Gort*, or *Lt. Commander Data*, or even *R2D2* or *C3P0*. And that's besides the fact that *Kismet* is real. There's no teeny, weeny, tiny, little actor making *Kismet* do what it does.

As evidenced by this thesis, I've had to dive deep into the study of intelligent behavior in natural systems to build a robot like *Kismet*. I've implemented capabilities on this robot that are pretty outlandish in relation to the status quo. Who would ever give a robot "emotions" or "drives"? Why give it expressive capabilities? Why build a robot that "needs" people? That's socially engaging? Because in many ways, that's where it all begins for us.

Kismet is an adorable, amazing little robot. Not only because of what it can do, but because of how it makes you feel. *Kismet* connects to people on a physical level, on a social level, and on an emotional level. For this reason, I do not see *Kismet* as being a purely scientific or engineering endeavor. It is an artistic endeavor as well. It is my masterpiece. I do not think anyone can get a full appreciation of what *Kismet* is by reading this dissertation. Video helps. But I think you have to experience it first hand to understand the connection this robot makes with so many people.

I could have never built *Kismet* alone. There are so many people who have contributed ideas, shaped my thoughts, and hacked code for the little one. There are so many people to give my heartfelt thanks. *Kismet* would not be what it is today without you.

First, I should thank Rod Brooks. He has believed in me, and supported me, and given me the freedom to pursue not one but several ridiculously ambitious projects. I honestly cannot think of another place in the world, working for anyone else, where I would have been given the opportunity to even attempt what I have accomplished in this lab.

I should also thank those who funded Kismet. Support for Kismet was provided in part by an ONR Vision MURI Grant (No. N00014-95-1-0600), and in part by DARPA/ITO under contract DABT 63-99-1-0012. I hope they are happy with the results.

Next, there are my fellow graduate students in the Humanoid Robotics Group who have put so much of their time and effort into making *Kismet* tick. In particular, I am indebted to Brian Scassellati, Paul Fitzpatrick and Lijin Aryananda. When I went into turbo-mode this last semester, Lijin and Paul kept up with me and never complained. I owe all of them big time. Kismet would not be able to see or hear without their help.

I've had so many useful discussions with others. I've picked Juan Velasquez's brain on many occasions about theories on emotion. I've cornered Robert Irie's again and again about auditory processing. I've bugged Matto Marjanovic throughout the years for just figuring out how to build random stuff. Kerstin Dautenhahn and Brian

Scassellati are kindered spirits with the same dream of building socially intelligent robots. Our discussions have had profound impact on the ideas in this thesis.

I've formed so many close friendships with the people I've had the pleasure to work with. These people continue enrich my life. Mike and Gina Binnard have continued be close friends, and are living proof that there's life after graduate school. Anita Flynn and Dave Barrett are dreamers by nature. They have encouraged me to follow my dreams. Brian Anthony has encouraged me when I needed it most. He often reminded me that life is a process – enjoy the process. There have been so many others, past and present. I thank you all.

Kismet wouldn't hear a thing if it weren't for the help of Jim Glass and Lee Hetherington of the Spoken Language Systems Group. They were very generous with their time and support in porting the speech recognition code to *Kismet*.

Bruce Blumberg was the one who first opened my eyes to the world of animation and synthetic characters. The concepts of believability, expressiveness, and audience perception are so critical for building sociable machines. I now see many strong parallels between his field and my own. I have learned so much from him. I've had great discussions with Chris Kline and Mike Hlavac from his group.

I'd like to thank the readers on my committee (Roz Picard, Justine Cassell and Eric Grimson) for wading through numerous drafts of this document. Their comments and suggestions have been so very helpful. Their insights and experience are invaluable.

I need to thank Jim Alser, for figuring out how to make *Kismet's* captivating blue eyes. It's just not the same robot without them.

And of course, I need to thank my family. My mother Juliette, my father Norman, and my brother William have all stood by me through the best of times and the worst of times. Their unconditional love and support have helped me through some very difficult times. I do my best to give them reason to be proud.

Contents

List of Figures

Chapter 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. What might such an interface look like?

1.1 A Universal Interface?

In the field of human computer interaction (HCI), researchers are already examining how people interact with one form of interactive technology - computers. Recent research by Reeves and Nass (1996) has shown that humans generally treat computers as they might treat other people, and it does not matter whether the people are computer experts, lay-people, or computer critics. They treat computers with politeness usually reserved for humans. They are careful to not hurt the computer's "feelings"' by criticizing it. They feel good if the computer compliments them. In team play they are even are willing to side with a computer against another human if the human belongs to a different team. If asked before the respective experiment if they could imagine treating a computer like a person, they strongly deny it. Even after the experiment, they insist that they treated the computer as a machine. They do not realize that they treated it as peer.

In these experiments, why do people unconsciously treat the computers in a social manner? To explain this behavior, Reeves and Nass appeal to evolution. Their main thesis is that the "*human brain evolved in a world in which only humans exhibited rich social behaviors, and a world in which all perceived objects were real physical objects. Anything that seemed to be a real person or place was real.*" (Reeves & Nass 1996), p.12. Evolution has hardwired the human brain with innate mechanisms that enable people to interact in a social manner with others that also behave socially. In short, we have evolved to be *experts* in social interaction. Our brains have changed very little over thousands of years, yet our brains have to deal with twentieth-century technology. As a result, if a technology behaves in a socially competent manner, we evoke our evolved social machinery to interact with it. Reeves and Nass argue that it actually takes *more* effort for people to consciously inhibit their social machinery in order to *not* treat the machine in this way. From their numerous studies, they argue that a social interface may be a truly universal interface (Reeves & Nass 1996).

1.1.1 An Argument for Sociable Humanoids

From these findings, 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. Similarity of morphology and sensing modalities makes humanoid robots one form of technology particularly well suited to this.

If the findings of Reeves and Nass 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. Sociable machines offer an intriguing alternative to the way humans interact with robots today.

1.2 Our Robot, Kismet

An important and challenging aspect of building a sociable machine is to *support natural human communication*. Another critical aspect is *socially situated learning*. Any robot that co-exists with people as part of their daily lives must be able to learn and adapt to new experiences. As designers, we simply cannot predict all the possible scenarios that such a robot will encounter. The challenge is not only to build a robot that is an effective learner, but to build a robot that can learn in a way that is natural and intuitive for people to teach.

We are particularly interested in this human form of socially situated learning, and we have argued for the many advantages social cues and skills could offer robots that learn from people (Breazeal & Scassellati 2000). The human learning environment is a dramatically different learning environment from that of typical autonomous robots. It is an environment that affords a uniquely rich learning potential. However, social interaction is required to tap into that potential.

Humans are the most socially advanced of all species. As one might imagine, a humanoid robot that could interact with people in a human-like way – one that could interpret, respond, and deliver human-style social cues even at the level of a human infant – is quite a sophisticated machine. As a starting point, we are exploring the simplest kind of human-style social interaction and learning– that which occurs between a human infant with its caregiver. Our primary interest in building this kind of robot is to explore the challenge of building a socially intelligent machine that can communicate with and learn from people.

Over the past three years, we have constructed an autonomous humanoid robot, called Kismet, and have been implementing a wide variety of infant-level social competencies into it. It is a very ambitious and highly integrated system, running on fifteen networked computers. The design and implementation of Kismet has drawn significant inspiration from models, theories, and concepts from the fields of psychology, social development, ethology, and evolutionary theory. We present much of this inspiration throughout the thesis. From Kismet's inception, the design has been driven by the desire to explore the kind of socially situated learning that occurs between a (robot) infant and its (human) caregiver. Much of this thesis is concerned with supplying the infrastructure to support this style of learning. However, the learning itself is the topic of future work.

This thesis presents the design issues, the framework, and the implementation of an autonomous humanoid robot that can engage humans in natural and intuitive interaction. Following the infant-caregiver metaphor, Kismet's interaction with a human is dynamic, physical, expressive, and social. We emphasize how designing for a human-in-the-loop introduces a new level of social constraints that profoundly impact the robot control problem – far beyond those issues of traditional autonomous robot control. A number of studies with naive human subjects are presented throughout the thesis. Using the data from these studies, we evaluate the work with respect to the performance of the human-robot system as a whole, not just the performance of the robot. In the next section, we explore this issue of socially situated learning in greater detail.

1.3 Socially Situated Learning

Humans (and other animals) acquire new skills socially through direct tutelage, observational conditioning, goal emulation, imitation, and other methods (Galef 1988), (Hauser 1996). These social learning skills provide a powerful mechanism for an observer (the learner) to acquire behaviors and knowledge from a skilled individual (the instructor). In particular, imitation is an extremely powerful mechanism for social learning which has received a great deal of interest from researchers in the fields of animal behavior and child development.

Similarly, social interaction can be a powerful way for transferring important skills, tasks, and information to a robot. A socially competent robot could take advantage of the same sorts of social learning and teaching scenarios that humans readily use. From an engineering perspective, a robot that could imitate the actions of a human would provide a simple and effective means for the human to specify a task and for the robot to acquire new skills without any additional programming. From a computer science perspective, imitation and other forms of social learning provide a means for biasing interaction and constraining the search space for learning. From a developmental psychology perspective, building systems that learn from humans allows us to investigate a minimal set of competencies necessary for social learning.

By positing the presence of a human that is motivated to help the robot learn the task at hand, a powerful set of constraints can be introduced to the learning problem. A good teacher is very perceptive to the limitations of the learner and scales the instruction accordingly. As the learner's performance improves, the instructor incrementally increases the complexity of the task. In this way, the learner is competent but slightly challenged - a condition amenable to successful learning. This type of learning environment captures key aspects of the learning environment of human infants, who constantly benefit from the help and encouragement of their caregivers. An analogous approach could facilitate a robot's ability to acquire more complex tasks in more complex environments. Keeping this goal in mind, we outline three key challenges of robot learning, and how social interaction can be used to address them in interesting ways.

Knowing what matters

Faced with an incoming stream of sensory data, a robot (the learner) must figure out which of its myriad of perceptions are relevant to learning the task. As the perceptual abilities of a robot increases, the search space becomes enormous. If the robot had a way of narrowing in on those few perceptions that mattered, the learning problem can become significantly more manageable.

Knowing what matters when learning a task is fundamentally a problem of determining saliency. Objects can gain saliency (that is, they become the target of attention) through a variety of means. At times, objects are salient because of their inherent properties; objects that move quickly, objects that have bright colors, and objects that are shaped like faces are all likely to attract attention. (We call these properties *inherent* rather than *intrinsic* because they are perceptual properties, and thus are observer-dependent and not strictly a quality of an external object.)

Objects can also become salient through contextual effects. The current motivational state, emotional state, and knowledge of the learner can impact saliency. For example, when the learner is hungry, images of food will have higher saliency than they otherwise would.

Objects can also become salient if they are the focus of the instructor's attention. For example, if the human is staring intently at a specific object, that object may become a salient part of the scene even if it is otherwise uninteresting. People naturally attend to the key aspects of a task while performing that task. By directing the robot's own attention to the object of the instructor's attention, the robot would automatically attend to the critical aspects of the task.

Hence, a human instructor could play a helpful role by indicating to the robot what features it should attend to as it learns how to perform the task. The instructor can take action to bring the robot's attention to those aspects. Also, in the case of social instruction, the robot's gaze direction could also serve as an important feedback signal for the instructor.

Knowing what action to try

Once the robot has identified salient aspects of the scene, how does it determine what actions it should take? As robots become more complex, their repertoire of possible actions increases. This also contributes to a large search space. If the robot had a way of focusing on those actions that are likely to be successful, the learning problem would be simplified.

In this case, a human instructor, sharing a similar morphology with the robot, could provide considerable assistance by demonstrating the appropriate actions to try. The body mapping problem is challenging, but could provide the robot with a good first attempt. The similarity in morphology between human and humanoid robot could also make it easier and more intuitive for the instructor to correct the robot's errors.

Evaluating actions, correcting errors, and recognizing success

Once a robot can observe an action and attempt to perform it, how can the robot determine whether or not it has been successful? The robot must be able to identify the desired outcome and to judge how its performance compares to that outcome. In many of these situations this evaluation depends upon an understanding of the goals and intentions of the instructor as well as the robot's own internal motivations. Further, if the robot has been unsuccessful, how does it determine which parts of its performance were inadequate? The robot must be able to diagnose its own errors in order to incrementally improve performance.

However, the human instructor has a good understanding of the task and knows how to evaluate the robot's success and progress. If the instructor could communicate this information to the robot, in a way that the robot could use, the robot could bootstrap from the instructor's evaluation in order to shape its behavior. One way a human instructor could facilitate the robot's evaluation process is by providing the robot with expressive feedback. The robot could use this feedback to recognize success and to correct failures. In the case of social instruction, the difficulty of obtaining success criteria can be simplified by exploiting the natural structure of social interactions. As the learner acts, the facial expressions (smiles or frowns), vocalizations, gestures (nodding or shaking of the head), and other actions of the instructor all provide feedback that could allow the learner to determine whether or not it has achieved the desired goal.

In addition, as the instructor takes a turn, the instructor often looks to the learner's face to determine whether the learner appears confused or understands what is being demonstrated. The expressive displays of a robot could be used by the instructor to control the rate of information exchange – to either speed it up, to slow it down, or to elaborate as appropriate. If the learner appears confused, the instructor slows down the training scenario until the learner is ready to proceed. Facial expressions could be an important cue for the instructor as well as the robot. Monitoring the structure of the social interaction can assist the instructor in maintaining an appropriate environment for learning. This improves the quality of instruction.

Finally, the structure of instructional situations is iterative; the instructor demonstrates, the student performs, and then the instructor demonstrates again, often exaggerating or focusing on aspects of the task that were not performed successfully. The ability to take turns lends significant structure to the learning episode. The instructor continually modifies the way he/she performs the task, perhaps exaggerating those aspects that the student performed inadequately, in an effort to refine the student's subsequent performance. By repeatedly responding to the same social cues that initially allowed the learner to understand and identify the salient aspects of the scene, the learner can incrementally refine its approximation of the actions of the instructor.

In the above discussion, we introduced several challenges in robot learning, and how social interaction and social cues could be used to address these challenges in new and interesting ways. For these reasons, we have implemented a number of these abilities on Kismet. These include the ability to *direct the robot's attention* to establish shared reference, the ability for the robot to *recognize expressive feedback* such as praise and prohibition, the ability to *give expressive feedback* to the human, and the ability to *take turns* to structure the learning episodes. In chapter 2, we will see strong parallels in how human caregivers assist their infant's learning through similar social interactions.

1.4 Embodied Systems that Interact with Humans

Before we launch into the presentation of our work with Kismet, we summarize some related work. These diverse implementations overlap a variety of issues and challenges that we have had to overcome in building Kismet.

There are a number of systems from different fields of research that are designed to interact with people. Many of these systems target different application domains such as computer interfaces, web agents, synthetic characters for entertainment, or robots for physical labor. In general, these systems can be either embodied (the human interacts with a robot or an animated avatar) or disembodied (the human interacts through speech or text entered at a keyboard). The embodied systems have the advantage of sending para-linguistic communication signals to a person, such as gesture, facial expression, intonation, gaze direction, or body posture. These embodied and expressive cues can be used to complement or enhance the agent's message. At times, para-linguistic cues carry the message on their own, such as emotive facial expressions or gestures. Cassell (1999*b*) presents a good overview of how embodiment can be used by avatars to enhance conversational discourse (however, there are a number of systems that interact with people without using natural language). Further, these embodied systems must also address the issue of sensing the human, often focusing on perceiving the human's embodied social cues. Hence, the perceptual problem for these systems is more challenging than that of disembodied systems. In this section we summarize a few of the embodied efforts, as they are the most closely related to Kismet.

1.4.1 Embodied Conversation Agents

There are a number of graphics-based systems that combine natural language with an embodied avatar. The focus is on natural, conversational discourse accompanied by gesture, facial expression, and so forth. The human uses these systems to perform a task, or even to learn how to perform a task.

Fully Embodied Agents

There are several fully embodied conversation agents under development at various institutions. One of the most advanced systems is *Rea* from the Media Lab at MIT (Cassell, Bickmore, Campbell, Vilhjalmsson & Yan 2000). Rea is a synthetic real-estate agent, situated in a virtual world, that people can query about buying property. The system communicates through speech, intonation, gaze direction, gesture, and facial expression. It senses the location of people in the room and recognizes a few simple gestures. Another advanced system is *Steve*, under development at USC (Rickel & Johnson 2000). Steve is a tutoring system, where the human is immersed in virtual reality to interact with the avatar. It supports domain-independent capabilities to support task-oriented dialogs in 3D virtual worlds. For instance, Steve trains people how to operate a variety of equipment on a virtual ship, and guides them through the ship to show them where the equipment is located. Another interesting system is *Cosmo*, under development at North Carolina State University (Lester, Towns, Callaway, Voerman & FitzGerald 2000). Cosmo is an animated pedagogical agent for children that operates on the web. The character inhabits the *Internet Advisor*, a learning environment for the domain of Internet packet routing. Because the character interacts with children, particular attention is paid to the issues of life-like behavior and engaging the students at an affective level.

Agents with Faces

There are a number of graphical systems where the avatar predominantly consists of a face with minimal to no body. A good example is *Gandalf*, a precurser system of Rea. The graphical component of the agent consisted of a face and a hand. It could answer a variety of questions about the Solar system, but required the user to wear a substantial amount of equipment in order to sense the user's gestures and head orientation (Thorisson 1998). In Takeuchi & Nagao (1993), the use of an expressive graphical face to accompany dialog is explored. They found that the facial component was good for initiating new users to the system, but its benefit was not as pronounced over time.

1.4.2 Interactive Characters

There are a variety of interactive characters under development for the entertainment domain. Some systems use natural language whereas others do not. Instead, the emphasis for each system is compelling, life-like behavior and characters with personality. Expressive, readable behavior is of extreme importance for the human to understand the interactive story line. Instead of passively viewing a scripted story, the user creates the story interactively with the characters.

Sympathetic Interfaces

A number of systems have been developed by at the MIT Media Lab. One of the earliest systems was the *ALIVE* project (Maes, Darrell, Blumberg & Pentland 1996). The best known character of this project is *Silus*, an animated dog that the user could interact with using gesture within a virtual space (Blumberg 1996). Several other systems have since been developed at the Media Lab by the Synthetic Characters Group, such as *Swamped!*, *(void*)*, and *Syndy k-9.0*. In *Swamped!* and *(void*)*, the human interacts with the characters using a *sympathetic interface*. For *Swamped!*, for instance, this was a sensor laden plush chicken (Johnson, Wilson, Blumberg, Kline & Bobick 1999). By interacting with the plush toy, the user could control the behavior of an animated chicken in the virtual world, which would then interact with other characters.

Believable Agents

There are several synthetic character systems that support the use of natural language. The *Oz* project at CMU is a good example (Bates 1994). The system stressed *broad and shallow* architectures, stressing the preference for characters with a broad repertoire of behaviors over those that are narrow experts. Some of the characters were graphics oriented (such as *woggles*), whereas others were text based (such as *Leotard the cat*). Using a text based interface, Bates, Loyall & Reilly (1992) explored the development of social and emotional agents. At Microsoft Research Labs, *Peedy* was an animated parrot that users could interact with in the domain of music (Ball, Ling, Kurlander, Miller, Pugh, Skelley, Stankosky, Thiel, Dantzich & Wax 1997). In later work at Microsoft Research, Ball & Breese (2000) explore incorporating emotion and personality into conversation agents using a Baysian network technique.

1.4.3 Human Friendly Humanoids

In the robotics community, there is a growing interest in building personal robots, or in building robots that share the same workspace with humans. Some projects focus on more advanced forms of tele-operation. Since our emphasis is on autonomous robots, we will not dwell on these systems. Instead, we concentrate on those efforts in building robots that interact with people.

Robotic Faces

There are several projects that focus on the development of expressive robot faces. Researchers at the Tokyo Institute of Technology have developed the most human-like robotic faces (typically resembling a Japanese woman) that incorporate hair, teeth, silicone skin, and a large number of control points (Hara 1998). Each control point maps to a *facial action unit* of a human face. The facial action units characterize how each facial muscle (or combination of facial muscles) adjust the skin and facial features to produce human expressions and facial movements (Ekman & Friesen 1982). Using a camera mounted in the left eyeball, the robot can recognize and produce a predefined set of emotive facial expressions (corresponding to anger, fear, disgust, happiness, sorrow, and surprise). A number of simpler expressive faces have been developed at Waseda University, one of which can adjust its amount of eye-opening and neck posture in response to light intensity (Takanobu, Takanishi, Hirano, Kato, Sato & Umetsu 1998).

Full Bodied Humanoids

There are a growing number of humanoid robotic projects underway, with a particularly strong program in Japan. Some humanoid efforts focus on more traditional challenges of robot control. Honda's *P3* is a bipedal walker with an impressive human-like gait (Hirai 1998). Another full bodied (but non-locomotory) humanoid is at ATR (Schaal 1999). Here, the focus has been on arm control and in integrating arm control with vision to mimic the gestures and tasks demonstrated by a human. There are several upper torso humanoid robots. There are two relatively new efforts: one at NASA, called *robonaut* (Ambrose, Aldridge & Askew 1999), and another at Vanderbilt University (Kawamura, Wilkes, Pack, Bishay & Barile 1996). One of the most well known humanoid robots is *Cog*, under development at the MIT Artificial Intelligence Lab (Brooks, Breazeal, Marjanovic, Scassellati & Williamson 1999). Cog is a general purpose humanoid platform used to explore theories and models of intelligent behavior and learning, both physical and social.

1.4.4 Personal Robots

There are a number of robotic projects that focus on operating within human environments. Typically these robots are not humanoid in form, but are designed to support natural communication channels such as gesture or speech.

Domestic Robots

There are a few robots that are being designed for domestic use. For systems such as these, safety, and minimizing their impact on human living spaces are important issues as well as performance and ease of use. Many applications of this kind focus on providing assistance to the elderly or to the disabled. The *MOVAID* system as described in Dario & Susani (1996), and a similar project at Vanderbilt University presented in Kawamura et al. (1996) are examples. In a somewhat related effort Dautenhahn (1999) has employed autonomous robots to assist in social therapy of fairly high-functioning autistic children.

Synthetic Pets

In the entertainment market, there are a growing number of synthetic pets (both robotic and digital). Sony's robot dog *Aibo* is the most sophisticated (and expensive). It can perceive a few simple visual and auditory features that allow it to interact with a pink ball and objects that appear skin-toned. It is mechanically quite sophisticated, able to locomote, to get up if it falls down, and performs an assortment of tricks. There are simpler, less expensive robotic toys such as Tiger Electronic's *Furby*. Successful digital pets include *Tomogotchis* which the child can carry with them, or animated pets that live on the computer screen such as PF Magic's *Petz*. The toys are designed to encourage people to establish a long term relationship with their toys.

1.5 Summary

In this chapter, we have motivated the construction of sociable machines from the viewpoint of building robots that are natural and intuitive to communicate with and to teach. We summarized a variety of related efforts in building embodied technologies that interact with people. We introduced Kismet, the subject of this thesis. Our work with Kismet is concerned both with supporting human-style communication as well as providing the infrastructure to support socially situated learning. We discussed how social interaction and social cues can address some of the key challenges in robot learning in new and interesting ways. These are the capabilities we have taken particular interest in building into Kismet.

Below, we outline the remainder of the thesis. We take care in each chapter to emphasize the constraints that interacting with a human imposes on the design of each system. We tie these issues back to supporting socially situated learning. Evaluation studies with naive subjects are presented at the end of many of the chapters to tie Kismet's behavior back to interacting with people. We have found that designing for a human-in-the-loop has placed profound constraints on how we think about the physical design of autonomous robots as well as their socially situated behavior. The outline of the remaining chapters is as follows:

- *Chapter 2*: We highlight some key insights from developmental psychology. These concepts have had a profound impact on the types of capabilities and interactions we have tried to achieve with Kismet.

- *Chapter 3*: We present an overview of the key design issues for sociable machines, an overview of Kismet's system architecture, and a set of the evaluation criteria.

- *Chapter 4*: We present the system hardware including the physical robot, its sensory configuration, and the computational platform.

- *Chapter 5*: We present an overview of Kismet's low level visual and auditory perceptions. A detailed presentation of the visual and auditory systems follows in later chapters.

- *Chapter 6*: We offer a detailed presentation of Kismet's visual attention system.

- *Chapter 7*: We present an in-depth description of Kismet's ability to recognize affective intent from the human caregiver's voice.

- *Chapter 8*: We give a detailed presentation of Kismet's motivation system, consisting of both homeostatic regulatory mechanisms as well as models of emotions. This system serves to motivate Kismet's behavior to maintain Kismet's internal state of "well being".

- *Chapter 9*: Kismet has several time-varying motivations and a broad repertoire of behavioral strategies to satiate them. This chapter presents Kismet's behavior system that arbitrates among these competing behaviors to establish the current goal of the robot.

- *Chapter 10*: Given the goal of the robot, the motor systems are responsible for controlling Kismet's output modalities (body, face, and voice) to carry out the task. This chapter presents an overview of Kismet's diverse motor systems and the different levels of control that produce Kismet's observable behavior.

- *Chapter 11*: We present an in-depth look at the motor system that controls Kismet's face. It must accommodate various functions such as emotive facial expression, communicative facial displays, and facial animation to accommodate speech.

- *Chapter 12*: We present Kismet's expressive vocalization system and lip synchronization abilities.

- *Chapter 13*: We present a multi-level view of Kismet's visual behavior, from low level oculo-motor control to using gaze direction as a powerful social cue.

- *Chapter 14*: We summarize our results, present future work for Kismet, and offer a set of grand challenges for building sociable machines.

Chapter 2

Insights from Developmental Psychology

> *Human babies become human beings because they are treated as if they already were human beings.* (Newson 1979).

In this chapter, we discuss the role social interaction plays in learning during infant-caregiver exchanges. First, we illustrate how the human newborn is primed for social interaction immediately after birth. This fact alone suggests how critically important it is for the infant to establish a social bond with his caregiver, both for survival purposes as well as to ensure normal development. Next, we focus on the caregiver and discuss how she employs various social acts to foster her infant's development. We discuss how infants acquire meaningful communication acts through ongoing interaction with adults. We conclude this chapter by relating these lessons to Kismet's design.

We have taken strong inspiration from developmental psychology in the design of Kismet's synthetic nervous system. In this chapter we see strong parallels to the previous chapter in how social interaction with a benevolent caregiver can foster robot learning. By implementing similar capabilities as the initial perceptual and behavioral repertoire of human infants, we hope to prime Kismet for natural social exchanges with humans and for socially situated learning.

2.1 Development of Communication and Meaning

Most of what a human infant learns is acquired within an ongoing, dynamic, and social process. This process begins immediately after birth with his caregiver, whom the infant depends upon for survival. Hence the social experience to which all infants are naturally exposed is one in which one member of the interaction pair is highly sophisticated and culturally competent, whereas the other is culturally naive.

2.1.1 Infant Preference for Social Stimuli

From birth, human infants are primed for social interaction with their caregivers. In general, infants exhibit a strong preference for humans over other forms of stimuli. Certain types of spontaneously occurring events may momentarily dominate their attention, or cause them to react in a quasi-reflex manner. However, the classes of events which dominate and hold their sustained attention leads one to conclude that they are biologically tuned to react to person-mediated events. They show a particular responsiveness to human caregivers, who very often react specifically to their immediately preceding actions. Hence, a caregiver's behavior is by no means random with respect to her infant's actions. This simple contingent reactivity makes her an object of absolute, compelling interest to her baby.

2.1.2 Innate Social Responses

Soon after birth, babies respond to their caregivers in a well coordinated manner. They seem to be born with a set of "pre-programmed" *proto-social responses*, which are specific to human infants. Their adaptive advantage seems to be their power to attract the attention of adults and to engage them in social interaction, the richness of which appears to be unique to the human species.

For instance, Bateson (1979) argues that the infant's inability to distinguish separate words in his caregiver's vocalizations may allow him to treat her clauses as unitary utterances analogous to his own coos and murmurs. This allows the infant to participate in "dialogues" with her. From these early dialogues, he can learn the cadence, rhythm, intonation, and emotional content of language long before speaking and understanding his first words (Fernald 1984). As another example, Johnson (1993) argues that the combination of having a limited depth of field[1] with early fixation patterns forces the infant to look predominantly at his caregiver's face. This brings the infant into face-to-face contact with his caregiver, which encourages her to try to engage him socially.

Kaye (1979) discusses a scenario where the burst-pause-burst pattern in suckling behavior, coupled with the caregiver's tendency to jiggle the infant during the pauses, lays the foundation of the earliest forms of turn-taking. Over time, the baby's ability to take turns becomes more flexible and regular; it is a critical skill for social learning. Turn-taking leads to dynamic exchanges between caregiver and infant.

Trevarthen (1979) discusses how the wide variety of facial expressions displayed by infants are interpreted by the caregiver as indications of the infant's motivational state. The caregiver views these as responses to her efforts to engage him, and they encourage her to treat him as an intentional being. These expressive responses provide the caregiver with feedback, which she uses to carry the dialog along.

2.1.3 Regulating Social Interaction

Given that the caregiver and infant engage in social interactions, there are a number of ways in which an infant limits the complexity of his interactions with the world. This is a critical skill for social learning because it allows the infant to keep himself from being overwhelmed or under stimulated for prolonged periods of time. For instance,

[1] A newborn's resolution is restricted to objects about 20 cm away, about the distance to his caregiver's face when she holds him.

the infant's own physically immature state serves to limit his perceptual and motor abilities, which simplifies his interaction with the world. In addition, the infant is born with a number of innate behavioral responses which constrain the sorts of stimulation that can impinge upon him. Various reflexes such as quickly withdrawing his hand from a painful stimulus, evoking the looming reflex in response to a quickly approaching object, closing his eyelids in response to a bright light, etc. all serve to protect the infant from stimuli that are potentially dangerous or too intense. In addition, whenever the infant is in a situation where his environment contains too much commotion and confusing stimuli, he either cries or tightly shuts his eyes. By doing so, he shuts out the disturbing stimulation.

To assist the caregiver in regulating the intensity of interaction, the infant provides her with cues as to whether he is being under stimulated or overwhelmed. For instance, when the infant feels comfortable in his surroundings, he generally appears content and alert. Too much commotion results in an appearance of anxiety, or crying, if the caregiver does not act to "correct" the environment. On the other hand, many experiments with infants exploit their tendency to show habituation or boredom (looking away from the stimulus) when a stimulus scenario is repeated often enough.

For the caregiver, her ability to present an appropriately complex view of the world to her infant strongly depends on how good she is at reading her infant's expressive and behavioral cues. It is interesting how adults naturally engage infants in appropriate interactions without realizing it, and caregivers seem to be instinctually biased to do so. For instance, *motherese* is a well known example of how adults simplify and exaggerate important aspects of language (Bateson 1979). By doing so, adults may draw the infant's attention to salient features of the adult's vocalizations (Fernald 1984). Exaggerated facial expressions to show extreme happiness or surprise during face-to-face exchanges with infants is another example.

2.1.4 Attributing Precocious Social Abilities to Infants

The early proto-social responses exhibited by infants are a close enough approximation to the adult forms that the caregiver immediately interprets her infant's reactions by a process of adultomorphism. Simply stated, she assumes her infant is fully socially responsive; with wishes, intentions, and feelings which can be communicated to others and which must be respected within certain limits. Events which may at first be the result of automatic action patterns, or may even be spontaneous or accidental, are endowed with social significance by the caregiver. By assuming that her infant is attempting some form of meaningful dialog, and by crediting him with having thoughts, feelings, and intentions like all other members of society, she imputes meaning to the exchange in a consistent and reliable manner. By doing so, she establishes a dialog with her infant, from which the communication of shared meanings gradually begins to take place.

By six weeks, human infants and their caregivers are communicating extensively face-to-face. The baby's expressions have become much more varied – they include coos, murmurs, smiles, frowns, waving and kicking. The caregiver interprets these activities as indications of the infant's emotional state, of his "beliefs" and "desires", and of his responses to her own acts of mothering. At such an early age, Kaye (1979) and Newson (1979) point out that it is the caregiver who supplies the meaning to the exchange, and it is the mechanism of flexible turn-taking that allows her to maintain the illusion that a meaningful exchange is taking place. For instance, whenever her infant does anything that can be interpreted as a turn in the "conversation", she will treat it as such. She fills in the gaps, and pauses to allow her infant to respond. She allows herself to be paced by him, but also subtly leads him on. She could not do this without the conviction that an actual dialog is taking place.

Although the caregiver-infant dialog still has no specific content, the pragmatics of conversation are being established. This is an important element for how meaning emerges for the infant. Schaffer (1977) writes that turn-taking of the non-specific, flexible, human variety is eminently suited to a number of important developments that occur over the next few months. It allows the infant to discover what sorts of activity on his part will get responses from his caregiver. It allows routine sequences of a predictable nature to be built up. And it provides a context of mutual expectations. It is the predictable and consistent behavior of the caregiver when interacting with her infant that makes this possible. She behaves in this consistent manner because she assumes the infant shares the same meanings that she applies to the interaction. Eventually, the infant picks up on these consistencies to the point where he also shares the same meanings. That is, he learns the significance that his actions and expressions have for other people.

In a similar way, caregivers bootstrap their infants to performing intentional acts (i.e., acts *about* something) arguably long before the infant is capable of intentional thought (Siegel 1999). Around the age of four months (after the caregiver has enjoyed extensive face-to-face interactions with her infant), the infant displays a new species typical activity pattern. Now the infant is able to break his caregiver's gaze to look at other things in the world. The caregiver interprets this break of gaze as an intentional act where the infant is now directing his gaze at some other object. In fact Collis (1979) points out that the infant's gaze does not seem to be directed at anything in particular. Furthermore, the infant does not seem to be trying to inform his caregiver of a newly found interest in objects. However, it is the caregiver who then converts a particular object into the object of attention. For instance, if an infant makes a reach and grasping motion in the direction of a given object, the she will assume that the infant is interested in that object and is trying to hold it. She inevitably intervenes by giving the object to the infant, thereby "completing" the infant's action. In this way, she has converted an arbitrary activity pattern into an action *about* something. The caregiver provides the supporting action in which the activity pattern acquires intentional significance. With this assistance supplied by the caregiver, the infant is performing intentional acts long before he is capable of intentional thought.

Hence, it is essential for the infant's psychological development that adults treat their infants as intentional beings. Both the infants' responses and their own maternal responses have been selected for because they foster this kind of interaction. It is by treating infants as intentional beings that the caregivers can bootstrap them into a cultural world. The infant's conception of himself and his actions, his beliefs, desires, and goals take shape from the situated interactive processes that his proto-social response patterns enable him to engage in with his caregiver.

Learning to Mean

Halliday (1975) explores the acquisition of meaningful communication acts from the viewpoint of how children *use* language to serve themselves in the course of daily life. From this perspective, a very young child may already have a linguistic system *long before* he has any words or grammar. Prior to uttering his first words, a baby is capable of expressing a considerable range of meanings which bear little resemblance to adult language, but which can be readily interpreted from a functional perspective, i.e. "what has the baby learned to do by means of language?". At a very young age, he is able to use his voice for doing something; it is a form of action that influences the behavior of the external world (such as the caregiver), and these meaningful vocal acts soon develop their own patterns and are used in their own significant contexts. To paraphrase Halliday: *He uses his voice to order people about, to get them to do things for him; he uses it to demand certain objects or services; he uses it to make contact with people, to feel close to them; and so on. All these things are meaningful actions.*

Halliday, refers to the child's first language as the "child's tongue" or *proto-language*. It comes into being around the middle of the first year of life. Hence, the child has already been meaning long before he ever utters his first words (which typically doesn't occur until about a year later). The infant arrives at meanings, i.e. a proto-language, through constant interaction with his caregivers. They unconsciously track his language, understanding what he meant, and respond with meanings of their own. They talk to him in a way that he can interpret with his own functional resources of meaning, while stretching his understanding without going beyond it. By doing so, they share in the child's language and its development at every stage.

2.2 Scaffolding for Social Learning

It is commonplace to say that caregiver-infant interaction is bi-directional, where each partner adapts to the other over time. However, each has a distinctive role in the dyad – they are not equal partners. The kinds of effects that infants have upon their caregivers are very different from those which go the other way. This is not surprising given that the caregiver is socially sophisticated, but the infant is not. Indeed, the caregiver's role is targeted towards developing the social sophistication of her infant. She does this by providing her infant with various forms of *scaffolding*.

Traditional Scaffolding

As viewed by the field of developmental psychology, *scaffolding* is traditionally conceptualized as a supportive structure provided by an adult (Wood, Bruner & Ross 1976). It is thought of in social terms where a more able adult manipulates the infant's interactions with the environment to foster novel abilities. Commonly it involves reducing distractions, marking the task's critical attributes, giving the infant affective forms of feedback, reducing the number of degrees of freedom in the target task, enabling the infant to experience the desired outcome before he is cognitively or physically able of seeking and attaining it for himself, and so forth. This view of scaffolding emphasizes the intentional contribution of the adult in providing conscious and deliberate support and guidance to enable the infant to learn new skills. It is used as a pedagogical device where the adult pushes the infant a little beyond his current abilities, and in the direction the adult wishes him to go. For instance, by exploiting the infant's instinct to perform a walking motion when supported upright, parents encourage their infant to learn how to walk before he is physically able.

Emergent Scaffolding

Another notion of scaffolding stresses the importance of early infant action patterns and their ability to attract the attention of adults and engage them in social interaction. This form of scaffolding is referred to as *emergent scaffolding* by (Hendriks-Jansen 1996). It relies on the caregiver-infant dyad being seen as two tightly coupled dynamic systems. In contrast to the previous case where the adult guides the infant's behavior to a desired outcome, here the response patterns arise from the continuous mutual adjustments between the two participants. For instance, the interaction between a suckling infant and the caregiver who jiggles him whenever he pauses in feeding creates a recognizable interactive pattern that emerges from low-level actions. This pattern of behavior encourages the habit of turn-taking upon which face-to-face exchanges will later be built. Many of these early action patterns that newborns exhibit have no place in adult behavior. They simply serve a bootstrapping role to launch the infant into an environment of adults who think in intentional terms, communicate through language, and manipulate objects. Within this socio-cultural context, these same skills are transferred from adult to child.

Internal Scaffolding

Looking within the infant, there is a third form of scaffolding. We call it *internal scaffolding*. This internal aspect refers to the incremental construction of the cognitive structures themselves that underlie observable behavior. Here, the form of the more mature cognitive structures are bootstrapped from earlier forms. Because these earlier forms provide the infant with some level of competence in the world, they are a good jumping off point for the later competencies to improve upon. In this way, the earlier structures foster and facilitate the learning of more sophisticated capabilities.

2.3 Specific Forms of Scaffolding for Social Learning

Above we presented three forms of scaffolding. The last (internal scaffolding) has to do with learning mechanisms. For the remainder of this section, we are concerned with the other two types of scaffolding, the specific forms they take during social exchange, and how this promotes the infant's continued learning and development. The way the caregiver provides this scaffolding reflects her superior level of sophistication over her infant, and the way she uses her expertise to coax and guide her infant down a viable developmental path.

Tronick, Als & Adamson (1979) likens the interaction between caregiver and infant to a duet played by a maestro and inept pupil. The maestro continually makes adjustments to add variety and richness to the interplay, while allowing the pupil to participate in, experience, and learn from a higher level of performance than the pupil could accomplish on his own. Similarly, within each session with her infant, the caregiver makes constant micro-adjustments to changes in her infant's behavior. To make these adjustments, she takes into account her infant's current abilities, his attention span, and his level of arousal. Based on these considerations, she adjusts the timing of her responses, introduces variations about a common theme to the interaction, and tries to balance his agenda with her own agenda for him (Kaye 1979).

Allowing Infants to Lead

During social interactions, the caregiver actually plays a subservient role to her infant. For instance, when talking with him she fills his pauses with her own utterances or gestures, and immediately pauses in anticipation when he is about to respond. However, she is the one actually in charge. She will purposely leave spaces between her own repetitious utterances and gestures for the infant to fill. In the meantime, she is constantly watching and listening for new initiatives from him. She imitates vocalizations, smiles, funny faces, tongue protrusions, and flurries of limb movement. If she can produce or prolong a run of alternations between herself and her infant, she will do so. All the while, she tries to prolong the duration of her infant's attention and activity cycles, and specifically tries to get him to respond to her. When he stops performing his part of the dialog, she may continue hers for a while to re-establish the dialog. Sometimes she will try to initiate a game. All the while, she tries to pull the infant along an intuitive curriculum of socialization.

Adjusting Behavior to Suit the Infant Limitations

The caregiver's performance exhibits tremendous implicit knowledge of her infant's physiological and psychological capabilities and limitations. Aware of her infant's limited attention span, her responses are aimed toward establishing and maintaining his interest. Often she tries to re-orient his eyes and face towards her so that they hold each other in mutual gaze. Once in mutual regard, she exaggerates, slows down, and simplifies her behavioral displays to fit within her infant's information processing abilities, which are slower and more limited than her own.

Directing the Infant's Attention

The ability of infants to direct their attention to salient stimuli is present at the earliest stages of development. It plays a critical role in social interactions with adults as well as learning during these exchanges. The caregiver initiates social exchange with her infant by first getting his attention so that they can establish mutual regard. During the exchange she may direct his attention to other objects and events, such as directing the interaction to be about a particular toy. If his attention wanes, she will try to re-engage him by making either herself or the toy more salient. She may shake the toy, she may assume a staccato manner of speech, etc. By directing the infant's attention to the most salient aspects of a task she would like him to learn, she facilitates the learning problem presented to him. By directing the infant's attention to a desired stimulus, the caregiver can establish *shared reference* which is a key component of social modeling theory (Pepperberg 1988). It is argued by Bateson (1979) that the infant's learning rate is accelerated when in social exchange because the caregiver focuses his attention on what is important.

Adjusting Timing of Responses

In general, the caregiver exhibits superior flexibility with respect to her own timing and anticipation of her infant's fairly regular cycling of his needs and level of arousal. She is aware that her windows for interaction are limited, and carefully times her responses to fit within them. For instance, she quickly learns to read his signals for sleep, food, emotional discharge, and arousal, and she detects the periodicity of these events so that she can fit face-to-face communication in at the appropriate time.

Entraining to the Infant

On a smaller time scale, during each session with her infant, she continually looks for pauses in the interaction and fills them with her responses. Because his attention span is short and intermittent, she times her responses so that they occur immediately after his gaze shifts back to her. She observes her infant's behavioral and affective cues and adapts her behavior in response. By doing so, his responses appear to be contingent upon hers. The interaction becomes smoother, more synchronized over time.

Regulating Infant Arousal to Promote Learning

The caregiver is also careful to maintain her infant's arousal at an appropriate level. Her maternal responses can be classified along a continuum from "turning on" to "turning off" her infant. She serves as a buffer to keep him at a moderate level of arousal, neither too high or too low. Of course, she partly does this for her own convenience and pleasure. However, according to (Kaye 1979), she also does this for the same reason an animal trainer maintains the animal at a moderate level of hunger. Performance and learning depend upon the infant's state, and caregivers devote a great deal of energy and vigilance to the maintenance of an optimal state.

Providing Affective Assessments

Human infants engage in a process of *social referencing* with their caregivers. In social referencing, the infant uses the caregiver's affective assessment of a novel situation to organize his own behavior. This assessment can occur via visual channels whereby the infant looks to the caregiver's face to see her own affective reaction to an unfamiliar situation (Siegel 1999). The assessment can also be communicated via auditory channels. Developmental psycholinguists have found that the prosodic exaggerations typical of infant-directed speech are particularly well matched to the innate affective responses of human infants. This allows caregivers to readily use their voice to directly influence the infant's emotive state, causing the infant to relax or become more vigilant in certain situations, and to either avoid or approach objects that may be unfamiliar (Fernald 1993). The caregiver's affective responses serve as socially communicated reinforcers for the infant. Given the number of important and novel situations that the human infant encounters (which do not result in immediate pain or some other innate reinforcer) social referencing plays an important role in the infant's social and behavioral development.

Using Repetition for Teaching

When interacting with her infant, the caregiver's movements and vocalizations are repetitive in nature, but she demonstrates ample creativity in introducing variations in her own repetitions. This sort of variation on a theme for stimulating the infant is optimal for holding the infant's attention and establishes a good learning environment for the infant (Stern 1975). According to Stern, these repetitive presentations dominate the kinds of stimulation the infant receives from his caregiver. She presents her responses in the form of content runs where an act or utterance re-occurs in nearly identical form multiple times, separated by short pauses. She may also present her responses in the form of temporal runs in which different acts or utterances occur, occupying nearly identical slots of time.

Shaping Infant's Agenda

During instructional interaction, the caregiver allows her infant to take the lead, but shapes his agenda to meet her own. She tries to meet him where he is, and accommodates quickly to his behavior changes. However, her behavior has a direction with respect to his. For instance, caregivers will tend to look and point in the direction the infant is already looking. At an early age (before 6 months), it is not the case that infants look where their caregivers tell them to look; yet caregivers behave as if that is the case. They fit their own behavior into that of the infant's, so that the infant's subsequent behavior will seem to be a contingent response. Gradually the infant does seem to fit his behavior into his caregiver's dialogue.

Imitating the Infant

This agenda-shaping process can also be seen when a caregiver imitates her infant. This is much more than a simple mirroring of her baby. Specifically, she pulls him from where he is into the direction she wants him to go. To do so, she uses several imitative strategies. For instance, she may employ *maximizing imitation* – if the baby opens his mouth, she will open her mouth in an exaggerated manner. Alternatively, she may employ *minimizing imitation*. For example, if the baby begins to make a cry face, she responds with a quick cry face that immediately flows back into a bright expression. Here, the caregiver flashes to where her infant is, and attempts to draw him back to where she wants him to be. She may also employ *modulating imitation*. For instance, when a baby whines "waaah", the caregiver responds with the same pitch intonation and duration, but mellows it to a sympathetic "awwwwww". There is an important characteristic here to imitation, it is *not* a perfect match. There is variation, in the direction of an individual's personal style, a learner's incompetence, or an instructor's agenda.

Playing Games with Infants

Another important observation is that each caregiver and infant develop a set of games of their own. These conventional games are the foundation of later communication and language-learning skills. What seems to be important is the process of conventionalization, the mutual topic-comment, the modularization of dyadic routines of some kind, and learning to anticipate when and how a partner's behavior will change (Kaye 1979).

Summary

The social programming an infant is subjected to is continuous and cumulative. The infant begins life with the capacity to elicit certain instructive kinds of behavior from adults. The caregiver constantly engages her infant using attention-creating and interest holding strategies. She acts to alleviate the baby's frustrations and discomforts, and tries to entertain and stimulate him. The infant will respond initially with various pre-programmed proto-social gestures like smiling, intent and interested looking, crying, or satisfied sucking or snuggling.

Soon, the infant will take more of the initiative, demanding and using attention-seeking patterns in attempts to attract or solicit caregiver attention. These initiatives rapidly become unmistakenly deliberate and intentional. Somehow he gradually takes upon himself some of the aspects of the adult's role in interaction: imitation, adjustment of timing, etc. This in turn gives him even finer control over the adult's behavior, so that he gains further information and more and more models of motor skills, of communication, and eventually of language. Indeed, he very soon learns to operate as powerful social manipulator of those who care about and care for him. By the time his representational and phonemic systems are ready to begin learning language, he is already able to make his intentions understood most of the time, to orient himself in order to read and interpret other's responses, to elicit repetitions and variations.

2.4 Lessons from Infants

> *Human caregivers program social shared meanings and intentions into babies.*(Newson 1979).

There are several key insights we have gleaned from the discussion in this chapter. The first is that human infants are born ready for social interaction with the caregiver. The initial perceptual and behavioral responses bias the infant to interact with adults, and encourage adults to interact with and care for him. Specifically, many of these responses enable the caregiver to carry on a "dialog" with him. Second, the caregiver uses scaffolding to establish a consistent and appropriately complicated social environment for the infant that he can predict, steer, and learn from. She allows him to act as if he is in charge of leading the dialog, but she is actually the one in charge. By doing so, she allows the infant to experiment and learn how his responses influence her. Third, the development of the baby's acts of meaning is inherently a social process, and it is grounded in having the infant learn *how* he can use his voice to serve himself. It is important to consider the infant's motivations — why he is motivated to use language and for what reasons. These motivations drive what he learns and why.

2.5 Proto-social Responses for Kismet

For people to treat Kismet as a socially aware being, it needs to convey subjective internal states: intents, beliefs, desires, and feelings. The robot can be designed to exploit our natural human tendencies to respond socially to certain behaviors. To accomplish this, we have implemented several infant-like social cues and responses that human infants exhibit.

Acts that make subjective processes overt include focusing attention on objects, orienting to external events, handling or exploring objects with interest, and so forth. Summarizing the discussions of this chapter, we divide these responses into four categories. These are listed below. By implementing these four classes of responses (*affective*, *exploratory*, *protective*, and *regulatory*) we aim to encourage the human to treat Kismet as an social creature and to establish meaningful communication with it.

- *Affective responses* allow the human to attribute feelings to the robot.
- *Exploratory responses* allow the human to attribute curiosity, interest, and desires to the robot, and can be used to direct the interaction toward objects and events in the world.
- *Protective responses* keep the robot away from damaging stimuli and elicit concerned and caring responses from the human.
- *Regulatory responses* maintain a suitable environment that is neither too overwhelming nor under-stimulating, and tunes the human's behavior in a natural and intuitive way to the competency of the robot.

Of course, once Kismet can partake in social interactions with people, it is also important that the *dynamics* of the interaction be natural and intuitive. For this, we take the work of Tronick et al. (1979) as a guide. They identify five phases that characterize social exchanges between three-month old infants and their caregivers: *initiation, mutual-orientation, greeting, play-dialog,* and *disengagement*. Each phase represents a collection of behaviors which mark the state of the communication. Not every phase is present in every interaction. For example, a greeting does not ensue if mutual orientation is not established. Furthermore, a sequence of phases may appear multiple times within a given exchange, such as repeated greetings before the play-dialog phase begins. This is discussed in depth in chapter 9.

Acquiring a genuine proto-language is beyond the scope of this dissertation, but learning how to mean and how to communicate those meanings to another (through voice face, body, etc.) is a fundamental capacity of a socially intelligent being. These capacities have profoundly motivated the creation of Kismet. Hence what is conceptualized and implemented in this dissertation is heavily inspired and motivated by the processes highlighted in this chapter. We endeavor to develop a framework that could ultimately be extended to support the acquisition of a proto-language and these characteristically human social learning process. This is the topic of the next chapter.

Chapter 3

Designing Sociable Machines

3.1 Design Issues for Sociable Machines

Our challenge is to build a robot that is capable of engaging humans in natural social exchanges that adhere to the infant-caregiver metaphor. Our motivation for this kind of interaction highlights our interest in social development and in socially situated learning for humanoid robots. Consequently, this thesis focuses on the problem of building the physical and computational infrastructure needed to support these sorts of interactions and learning scenarios. The social learning, however, is beyond the scope of this thesis.

Inspired by infant social development, psychology, ethology, and evolutionary perspectives, this work integrates theories and concepts from these diverse viewpoints to enable Kismet to enter into natural and intuitive social interaction with a human caregiver. For lack of a better metaphor, we refer to this infrastructure as the robot's *synthetic nervous system* (SNS). Kismet is designed to perceive a variety of natural social cues from visual and auditory channels, and to deliver social signals to the human caregiver through gaze direction, facial expression, body posture, and vocalizations. Every aspect of its design is directed toward making the robot proficient at interpreting and sending readable social cues to the human caregiver, as well as employing a variety of social skills, to foster its behavioral and communication performance (and ultimately its learning performance). This requires that the robot have a rich enough perceptual repertoire to interpret these interactions, and a rich enough behavioral repertoire to act upon them. As such, the design must address the following issues:

Situated in a Social Environment

Kismet must be situated in a social and benevolent learning environment that provides scaffolding interactions. For our purposes, this means that the environment contains a benevolent human caregiver.

Real-Time Performance

Fundamentally, Kismet's world is a social world containing a keenly interesting stimulus: an interested human (sometimes more than one) who is actively trying to engage the robot in a dynamic social manner, to play with it, and to teach it about its world.

It is difficult to imagine a more dynamic and complex environment. We have found that it demands a relatively broad and well integrated perceptual system that must run at natural interactive rates. The same holds true for the robot's behavioral repertoire and expressive abilities. Rich perceptual, behavioral, and expressive repertoires and real-time performance are a must for the nature and quality of interaction we are trying to achieve.

Establish Appropriate Social Expectations

Kismet should have an appealing appearance and a natural interface that encourages humans to interact with Kismet as if it were a young, socially aware creature. If successful, humans will naturally provide scaffolding interactions without consciously thinking about it. Furthermore, they will expect the robot to behave at a competency-level of an infant-like creature. In particular, at a level that is achievable given the robot's perceptual, mechanical, and computational limitations.

Self-Motivated Interaction

Kismet's synthetic nervous system must motivate the robot to pro-actively engage in social exchanges with the caregiver and to take an interest in things in the environment. Each social exchange can be viewed as an episode where the robot tries to manipulate the caregiver into addressing its "needs" and "wants". This serves as the basic impetus for social interaction, upon which richer forms of communication could be built. This internal motivation frees the robot from being a slave to its environment, responding only in a reflexive manner to incoming stimuli. Given its own motivations, the robot can internally influence the kinds of interactions it pursues.

Regulate Interactions

Kismet must be capable of regulating the complexity of its interactions with the world and its caregiver. To do this, Kismet should provide the caregiver with social cues (through facial expressions, body posture, or voice) as to whether the interaction is appropriate for it or not – i.e., the robot should communicate whether the interaction is overwhelming or under stimulating. For instance, it should signal to the caregiver when the interaction is overtaxing its perceptual or motor abilities. Further, it should provide readable cues as to what the appropriate level of interaction is. Kismet should exhibit interest in its surroundings, interest in the humans that engage it, and behave in a way to bring itself closer to desirable aspects and to shield itself from undesirable aspects. By doing so, the robot behaves to promote an environment for which its capabilities are well matched. Ideally, an environment where it is slightly challenged but largely competent, in order to foster its social development.

Readable Social Cues

Kismet should send social signals to the human caregiver that provide the human with feedback of its internal state. If designed properly, humans should intuitively and naturally use this feedback to tune their performance in the exchange. Through a process of entraining to the robot, both the human and robot benefit. The resulting interaction should be natural, intuitive, and enjoyable for the person. It should allow the robot to perform effectively and be commensurate with its perceptual, computational, and behavioral limits. Ultimately, these cues will allow humans to improve the quality of their instruction.

Read the Human's Social Cues

During social exchanges, the person sends social cues to Kismet to shape its behavior. Hence, Kismet must be able to perceive and respond to these cues appropriately. By doing so, the quality of the interaction improves. Furthermore, many of these social cues will eventually be offered in the context of teaching the robot. To be able to take advantage of this scaffolding, the robot must be able to correctly interpret and react to these social cues.

Competent Behavior in a Complex World

Any convincing robotic creature must address similar behavioral issues as living, breathing creatures. The robot must exhibit robust, flexible, and appropriate behavior in a complex dynamic environment to maintain its "well being". This often entails having the robot apply its limited resources (finite number of sensors, actuators and limbs, energy, etc.) to perform various tasks. Given a specific task, the robot should exhibit a reasonable amount of persistence. It should work to accomplish a goal, but not at the risk of ignoring other important tasks if the current task is taking too long. Frequently the robot must address multiple goals at the same time. Sometimes these goals are not at cross-purposes and can be satisfied concurrently. Sometimes these goals conflict and the robot must figure out how to allocate its resources to address both adequately. Which goals the robot pursues, and how it does so, depends both on external influences coming from the environment as well as internal influences from the creature's motivations, perceptions, and so forth.

Believable Behavior

The above issue targets the challenges that an artificial creature must solve to operate well in a complex dynamic environment. However, they do not address the issue of portraying convincing, life-like behavior. For Kismet, it is critical that the caregiver perceive the robot as an intentional creature that responds in meaningful ways to his/her attempts at communication. As previously discussed in section 2, the scaffolding the human provides through these interactions is based upon this assumption. Hence, the synthetic nervous system must address a variety of issues to promote the illusion of a socially aware robotic creature. Blumberg (1996) provides such a list, slightly modified as shown here: *convey intentionality*, *promote empathy*, *expressiveness*, and *variability*.

These are the high-level design issues of the overall human-robot system. The system encompasses the robot, its environment, the human, and the nature of interactions between them. The human's behavior is governed by many internal factors that arise from evolution, physiological and psychological processes, development, learning, and cultural norms (and more). Hence the human brings a complex set of well-established social machinery to the interaction. Hence, our aim is not a matter of re-engineering the human side of the equation. Instead we need to engineer *for* the human side of the equation. We need to design Kismet's synthetic nervous system so that it supports what comes naturally to people. Humans are already experts at social communication and of social forms of learning and instruction.

If we are clever, we can design Kismet so that people intuitively engage in appropriate interactions with the robot. This can be accomplished in a variety of ways, such as physically designing the robot to establish the correct set of social expectations for humans, or having Kismet send social cues to humans that they intuitively use to fine tune their performance.

The following sections present a high level overview of the synthetic nervous system. It encompasses the robot's perceptual, motor, attention, motivation, and behavior systems. Eventually, it should include learning mechanisms so that robot becomes better adapted to its environment over time.

3.2 Design Hints from Animals, Humans, and Infants

In this section, we briefly present ideas for how natural systems address similar issues as those outlined above. Many of these ideas have shaped the design of Kismet's synthetic nervous system. Accordingly, we motivate the high level design of each component system, how each interfaces with the other, and the responsibility each carries out for the overall synthetic nervous system. The following chapters of this thesis present each component system in more detail.

The design of the underlying architecture of the SNS is heavily inspired by models, mechanisms, and theories from the scientific study of intelligent behavior in living creatures. For many years, these fields have sought explanatory models for how natural systems address the aforementioned issues. However, it is important to distinguish the psychological theory/hypothesis from its underlying implementation in Kismet.

The particular models used to design Kismet's synthetic nervous system are not necessarily the most recent nor popular in their respective fields. They were chosen based on how easily they could be applied to this application, how compatible they are with other aspects of the system, and how well they could address the aforementioned issues within synthetic creatures. Our focus has been to engineer a system that exhibits the desired behavior, and we have found scientific findings from the study of natural systems to be useful in this endeavor. Our aim has not been to explicitly test or verify the validity of these models or theories. Limitations of Kismet's performance could be ascribed to limitations in the mechanics of the implementation (dynamic response of the actuators, processing power, latencies in communication), as well as to the limitations of the models used.

Hence, we do not claim explanatory power for understanding human behavior with our implementation. We do not claim equivalence with psychological aspects of human behavior such as emotions, attention, affect, motivation, etc.. However, we have implemented synthetic analogs of proposed models, we have integrated them within the same robot, and we have situated Kismet in a social environment. The emergent behavior between Kismet's synthetic nervous system and its social environment is quite compelling. When we evaluate Kismet, we do so with an engineer's eye. We are testing the adequacy of Kismet's performance, not that of the underlying psychological models.

Below, we highlight special considerations from natural systems that have inspired the design of the robot's synthetic nervous system. Infants do not come into this world as mindless, flailing skin bags. Instead, they are born as a coherent system, albeit immature, with the ability to respond to and act within their environment in a manner that promotes their survival and continued growth. It is the designer's challenge to endow the robot with the "innate" endowments (i.e., the initial set of software and hardware) that implements similar abilities to that of a newborn. This forms the foundation upon which learning can take place.

Ethology

Models from ethology have a strong influence in addressing the behavioral issues of the system (i.e. relevance, coherence, concurrency, persistence, and opportunism). As such, they have shaped the manner in which behaviors are organized, expressed, and arbitrated among. Ethology also provides important insights as to how other systems influence behavior (i.e. motivation, perception, attention, and motor expression).

Social Development and Evolutionary Perspectives

These ethology-based models of behavior are supplemented with models, theories, and behavioral observations from developmental psychology and evolutionary perspectives. In particular, these ideas have had a strong influence in the specification of the "innate endowments" of the synthetic nervous system, such as early perceptual skills (visual and auditory) and proto-social responses. The field has also provided many insights into the nature of social interaction and learning with a caregiver, and the importance of motivations and emotional responses for this process.

Psychology

Models from psychology have influenced the design details of several systems. In particular, psychological models of the attention system, facial expressions, the emotion system, and various perceptual abilities have been adapted for the Kismet's synthetic nervous system.

3.3 A Framework for the Synthetic Nervous System

The design details of each system and how they have incorporated concepts from these scientific perspectives are presented in depth in later chapters. Here, we simply present a bird's eye view of the overall synthetic nervous system to give the reader a sense of how the global system fits together. The overall architecture is shown in figure 3-1.

The system architecture consists of six subsystems: the *low-level feature extraction system*, the *high-level perception system*, the *attention system*, the *motivation system*, the *behavior system*, and the *motor system*. The low-level feature extraction system extracts sensor-based features from the world, and the high-level perceptual system encapsulates these features into percepts that can influence behavior, motivation, and motor processes. The attention system determines what the most salient and relevant stimulus of the environment is at any time so that the robot can organize its behavior about it. The motivation system regulates and maintains the robot's state of "well being" in the form of homeostatic regulation processes and emotive responses. The behavior system implements and arbitrates between competing behaviors. The winning behavior defines the current task (i.e., the goal) of the robot. The robot has many behaviors in its repertoire, and several motivations to satiate, so its goals vary over time. The motor system carries out these goals by orchestrating the output modalities (actuator or vocal) to achieve them. For Kismet, these actions are realized as motor skills that accomplish the task physically, or expressive motor acts that accomplish the task via social signals.

Learning mechanisms will eventually be incorporated into this framework. Most likely, they will be distributed through out the synthetic nervous system to foster change within various subsystems as well as between them. It is known that natural systems possess many different kinds of interacting learning mechanisms (Gallistel 1990). Such will be the case with the synthetic nervous system described here. However, this is the topic of future work. Below, we summarize the systems that comprise the synthetic nervous system.

The Low-Level Feature Extraction System

The low-level feature extraction system is responsible for processing the raw sensory information into quantities that have behavioral significance for the robot. The routines are designed to be cheap, fast, and just adequate. Of particular interest are those perceptual cues that infants seem to rely on. For instance, visual and auditory cues such as detecting eyes and the recognition of vocal affect are important for infants. The low level perceptual features incorporated into this system are presented in chapter 5 and 6. More specific auditory percepts are presented in chapter 7.

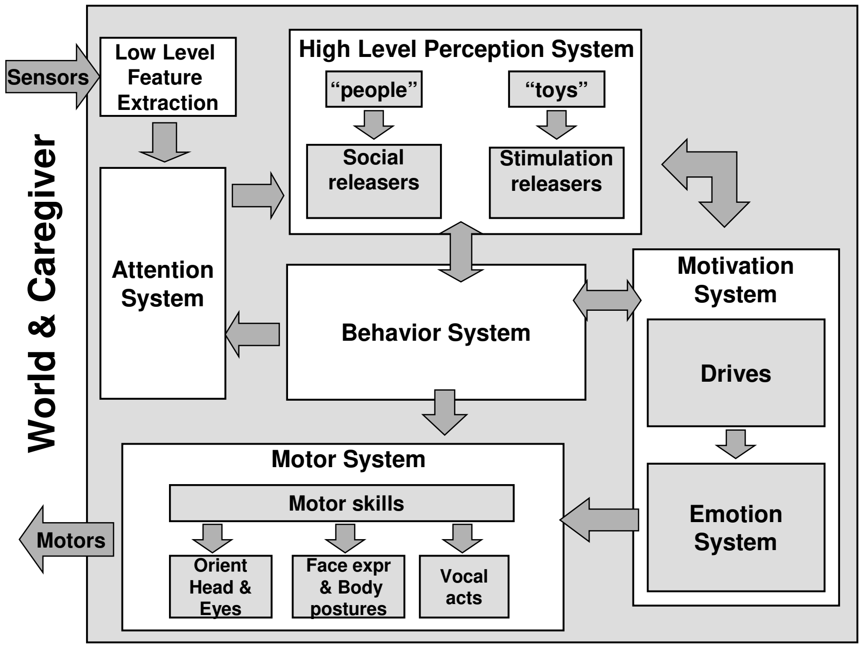


Figure 3-1: A framework for designing synthetic nervous systems. Six sub-systems interact to enable the robot to behave coherently and effectively. See text.

The Attention System

The low-level visual percepts are sent to the attention system. The purpose of the attention system is to pick out low-level perceptual stimuli that are particularly salient or relevant at that time, and to direct the robot's attention and gaze toward them. This provides the robot with a locus of attention that it can use to organize its behavior. A perceptual stimulus may be salient for several reasons. It may capture the robot's attention because of its sudden appearance, or perhaps due to its sudden change. It may stand out because of its inherent saliency such as a red ball may stand out from the background. Or perhaps its quality has special behavioral significance for the robot such as being a typical indication of danger. See chapter 6 for more details.

The Perceptual System

The low-level features corresponding to the target stimuli of the attention system are fed into the perceptual system. Here they are encapsulated into behaviorally relevant percepts. To environmentally elicit processes in these systems, each behavior and emotive response has an associated *releaser*. As conceptualized by Tinbergen (1951) and Lorenz (1973), a releaser can be viewed as a collection of feature detectors that are minimally necessary to identify a particular object or event of behavioral significance. Their function of the releasers is to ascertain if all environmental (perceptual) conditions are right for the response to become active. High level perceptions that influence emotive responses are presented in chapter 8, and those that influence task-based behavior are presented in chapter 9.

The Motivation System

The motivation system consists of the robot's basic "drives" and "emotions" (see chapter 8). The `drives` represent the basic "needs" of the robot and are modeled as simple homeostatic regulation mechanisms (Carver & Scheier 1998). When the needs of the robot are being adequately met, the intensity level of each "drive" is within a desired regime. However, as the intensity level moves farther away from the homeostatic regime, the robot becomes more strongly motivated to engage in behaviors that restore that "drive". Hence the "drives" largely establish the robot's own agenda, and play a significant role in determining which behavior(s) the robot activates at any one time.

The "emotions" are modeled from a functional perspective. Based on simple appraisals of the benefit or detriment of a given stimulus, the robot evokes positive emotive responses that serve to bring itself closer to it, or negative emotive responses in order to withdraw from it. There is a distinct emotive response for each class of eliciting conditions. Currently, six *basic* emotions are modeled that give the robot synthetic analogs of anger, disgust, fear, joy, sorrow, and surprise (Ekman 1992). There are also arousal-based responses that correspond to interest, calm, and boredom that are modeled in a similar way. The expression of emotive responses promotes empathy from the caregiver and plays an important role in regulating social interaction with the human.

The Behavior System

The behavior system organizes the robot's task-based behaviors into a coherent structure. Each behavior is viewed as a self-interested, goal-directed entity that competes with other behaviors to establish the current task. An arbitration mechanism is required to determine which behavior(s) to activate and for how long, given that the robot has several motivations that it must tend to and different behaviors that it can use to achieve them. The main responsibility of the behavior system is to carry out this arbitration. In particular, it addresses the issues of relevancy, coherency, persistence, and opportunism. By doing so, the robot is able to behave in a sensible manner in a complex and dynamic environment. The behavior system is described in depth in section 9.

The Motor System

The motor system arbitrates the robot's motor skills and expressions. It consists of four subsystems: the *motor skills system*, the *facial animation system*, the *expressive vocalization system* , and the *oculo-motor system*. Given that a particular goal and behavioral strategy have been selected, the motor system determines how to move the robot so as to carry out that course of action. Overall, the motor skills system coordinates body posture, gaze direction, vocalizations, and facial expressions to address issues of blending and sequencing the action primitives from the specialized motor systems.

3.4 Mechanics of the Synthetic Nervous System

The overall architecture is agent-based as conceptualized by Minsky (1988), Maes (1990), Brooks (1986), and bears strongest resemblance to that of Blumberg (1996). As such, the synthetic nervous system is implemented as a highly distributed network of interacting elements. Each computational element (or node) receives messages from those elements connected to its inputs, performs some sort of specific computation based on these messages, and then sends the results to those connected to its outputs. The elements connect to form networks, and networks are connected to form the component systems of the SNS.

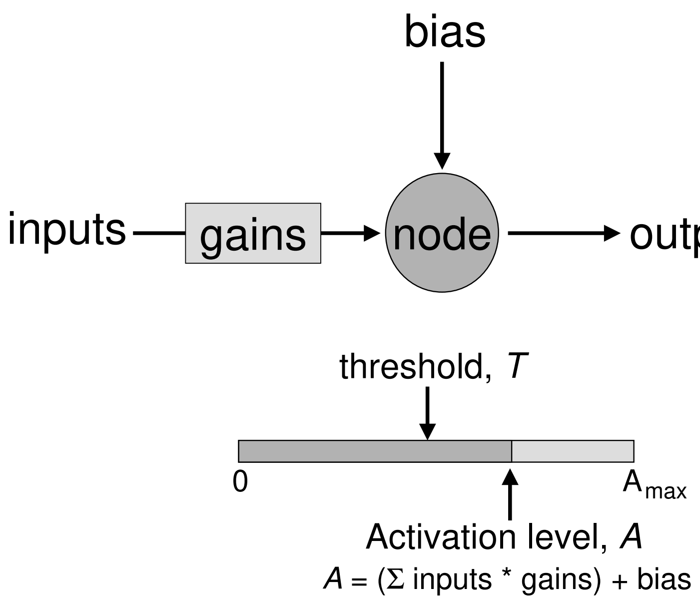


Figure 3-2: A schematic of a basic computational process. The process is active when the activation level A exceeds threshold T.

The Basic Computational Unit

For this implementation, the basic computational process is modeled as shown in figure 3-2. Its *activation level* A is computed by the equation: $A = (\sum_{n}^{j=1} w_j \cdot i_j) + b$ for integer values of inputs i_j, weights w_j, and bias b over the number of inputs, n.

The weights can be either positive or negative; a positive weight corresponds to an excitatory connection and a negative weight corresponds to an inhibitory connection. Each process is responsible for computing its own activation level to determine when it should become active. The process is active when its activation level exceeds an *activation threshold*. When active, the process can send activation energy to other nodes to favor their activation. It may also perform some special computation, send output messages to connected processes, and/or express itself through motor acts by sending outputs to actuators. Each drive, emotion, behavior, perceptual releaser, and motor process is modeled as a different type that is specifically tailored for its role in the overall system architecture. Hence, although they differ in function, they all follow the basic activation scheme.

Networks of Units

Units are connected together to form networks of interacting processes that allows for more complex computation. This involves connecting the output(s) of one unit to the input(s) of other unit(s). When a unit is active, besides passing messages to the units connected to it, it can also pass some of its activation energy. This is called *spreading activation* and is a mechanism by which units can influence the activation or suppression of other units (Maes 1990). This mechanism was originally conceptualized by Lorenz (1973) in his *Hydraulic Model*. Minsky (1988) uses a similar scheme in his ideas of memory formation using *K-lines*.

Subsystems of Networks

Groups of connected networks form subsystems. Within each subsystem the active nodes perform special computations to carry out tasks for that subsystem. To do this, the messages that are passed among and within these networks must *share* a common currency so that the information contained in the messages can be processed and combined in a principled manner (McFarland & Bosser 1993). Furthermore, as the subsystem becomes more complex, it is possible that some agents may conflict with others (such as when competing for shared resources). In this case, the agents must have some means for competing for expression. If each agent computes its relevance in terms of a shared currency, conflicting agents can compete based on this value.

Common Currency

This raises an important issue with respect to communication within and between different subsystems. Observable behavior is a product of many interacting processes. For instance, ethology, comparative psychology, and neuroscience have shown that observable behavior is influenced by internal factors (motivations, past experience, etc.) as well as by external factors (perception). This demands that the subsystems be able to communicate and influence each other despite their different functions and modes of computation. This has led ethologists such as McFarland & Bosser (1993) and Lorenz (1973) to propose that there must be a *common currency* that is shared between perceptual, motivational, and behavioral subsystems. In this scheme, the perceptual subsystem generates values based on environmental stimuli, and the motivational subsystem generates values based on internal factors. Both sets of values are passed to the behavioral subsystem, where competing behaviors use them to compute their relevance and then compete for expression based on this value. Within different subsystems, each can operate on their own currencies. This is the case of Kismet's behavior system (chapter 9) and emotion system (chapter 8). However, the currency that is passed between different systems must be shared.

Value Based

Based upon this idea, the robot's synthetic nervous system is implemented as a *value based system*. This simply means that each process computes numeric values (in a common currency) from its inputs. These values are passed as messages (or activation energy) throughout the network, either within a subsystem or between subsystems. Conceptually, the magnitude of the value represents the strength of the contribution in influencing other processes. Using a value-based approach has the nice effect of allowing influences to be graded in intensity, instead of simply being "on" or "off". Other processes compute their relevance based on the incoming activation energies or messages, and use their computed activation level to compete with others for exerting influence upon the synthetic nervous system.

3.5 Criteria for Evaluation

Thus far in this chapter, we have presented the key design issues for Kismet. To address them, we have outlined the framework for the synthetic nervous system. We now turn to the question of evaluation criteria.

Kismet is neither designed to be a tool nor an interface. One does not use Kismet to perform a task. Kismet is designed to be a robotic creature that can interact socially with humans and ultimately learn from them. As a result, it is difficult or inappropriate to apply standard Human Computer Interface (HCI) evaluation criteria to Kismet. Many of these relate to the ability for the system to use natural language, which Kismet is not designed to handle. Some evaluation criteria for embodied conversation agents are somewhat related, such as the use of embodied social cues to regulate turn taking during dialogs. However, many of these are also closely related to conversational discourse (Sanders & Scholtz 2000). Currently, Kismet only babbles, it does not speak any natural language.

Instead, Kismet's interactions with humans are fundamentally physical, affective, and social. The robot is designed to elicit interactions with the caregiver that afford rich learning potential. We have endowed the robot with a substantial amount of infrastructure that we believe will enable the robot to leverage from these interactions to foster its social development. As a result, we evaluate Kismet with respect to *interact-ability* criteria. These are inherently subjective, yet quantifiable, measures that evaluate the quality and ease of interaction between robot and human. They address the behavior of both partners, not just the performance of the robot. The evaluation criteria for interact-ability are as follows:

- Can people intuitively read and do they naturally respond to Kismet's social cues?
- Can Kismet perceive and appropriately respond to these naturally offered cues?
- Does the human adapt to the robot, and the robot adapt to the human, in a way that benefits the interaction? Specifically, we want to determine whether the resulting interaction is natural, intuitive, and enjoyable for the human, and if Kismet can perform well despite its perceptual, mechanical, behavioral, and computational limitations.
- Does Kismet readily elicit scaffolding interactions from the human that could be used to benefit learning?

3.6 Summary

In this chapter, we have outlined our approach for the design of a robot that can engage humans in a natural, intuitive, social manner. We have carefully considered a set of design issues that are of particular importance when interacting with people. Humans will perceive and interpret the robot's actions as socially significant and possessing communicative value. They will respond to them accordingly. This defines a very different set of constraints and challenges for autonomous robot control that lie along a social dimension.

We are interested in giving Kismet the ability to enter into social interactions reminiscent of those that occur between infant and caregiver. These include interactive games, having the human treat Kismet's babbles and expressions as though they are meaningful, and to treat Kismet as a socially aware creature whose behavior is governed by perceived mental states such as intents, beliefs, desires, and feelings. As discussed in chapter 2, these interactions are critical for the social development of infants. Continuing with the infant-caregiver metaphor for Kismet, these interactions could also prove important for Kismet's social development. In chapter 1 we outlined several interesting ways in which various forms of scaffolding address several key challenges of robot learning.

As such, this dissertation is concerned with providing the infrastructure to elicit and support these future learning scenarios. In this chapter, we have outlined a framework for this infrastructure that adapts theories, concepts, and models from psychology, social development, ethology, and evolutionary perspectives. The result is a synthetic nervous system that is responsible for generating the observable behavior of the robot and for regulating the robot's internal state of "well being". To evaluate the performance of both the robot and the human, we have introduced a set of evaluation criteria for interact-ability. Throughout the thesis, we will present a set of studies with naive human subjects that provide the data for our evaluations. In the following chapter, we begin our in-depth presentation of Kismet starting with a description of the physical robot and its computational platform.

Chapter 4

The Physical Robot

Our design task is to build a physical robot that encourages humans to treat it as if it were a young socially aware creature. This entails that the robot have an appealing infant-like appearance so that humans naturally fall into this mode of interaction. The robot must have a natural and intuitive interface (with respect to its inputs and outputs) so that a human can interact with it using natural communication channels. This enables the robot to both read and send human-like social cues. Finally, the robot must have sufficient sensory, motor, and computational resources for real-time performance during dynamic social interactions with people.

4.1 Design Issues and Robot Aesthetics

When designing robots that interact socially with people, the aesthetics of the robot should be carefully considered. The robot's physical appearance, its manner of movement, and its manner of expression convey personality traits to the person who interacts with it. This fundamentally influences the manner in which people engage the robot.

Youthful and Appealing

It will be quite a while before we are able to build autonomous humanoids that can rival the social competence of human adults. For this reason, Kismet is designed to have an infant-like appearance of a fanciful robotic creature. Note that the human is a critical part of the environment, so evoking appropriate behaviors from the human is essential for this project. The key set of features that evoke nurturing responses of human adults (see figure 4-1) has been studied across many different cultures (Eibl-Eibesfeldt 1972), and these features have been explicitly incorporated into Kismet's design (Breazeal & Foerst 1999). Other issues such as physical size and stature also matter. For instance, when people are standing they look down to Kismet and when they are seated they can engage the robot at eye level. As a result, people tend to intuitively treat Kismet as a very young creature and modify their behavior in characteristic baby-directed ways. As argued in chapter 2, these same could be used to benefit the robot by simplifying the perceptual challenges it faces when behaving in the physical world. It also allows the robot to participate in interesting social interactions that are well matched to the robot's level of competence.

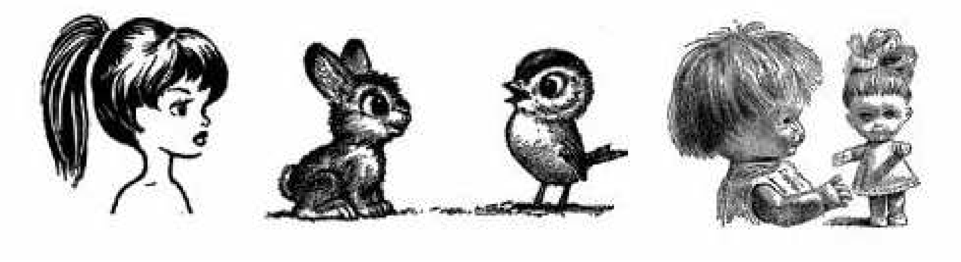


Figure 4-1: Examples of the baby scheme of Eibl-Eibelsfeldt (Eibl-Eibelsfeldt 1970). He posits that there is a set of facial characteristics that cross-culturally trigger nurturing responses from adults. These include a large head with respect to the body, large eyes with respect to the face, a high forehead, and lips that suggest the ability to suck. These features are commonly incorporated into dolls and cartoons, as shown here.

Believability verses Realism

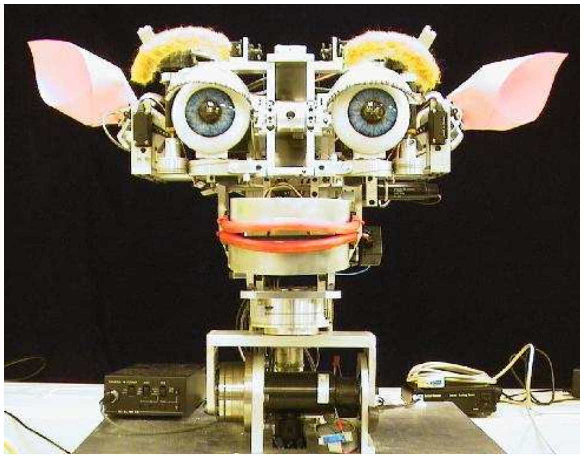
Along a similar vein, the design should minimize factors that could detract from a natural infant-caretaker interaction. Ironically, humans are particularly sensitive (in a negative way) to systems that try to imitate humans but inevitably fall short. Humans have strong implicit assumptions regarding the nature of human-like interactions, and they are disturbed when interacting with a system that violates these assumptions (Cole 1998). For this reason, we have made a conscious decision to *not* make the robot look human. Instead the robot resembles a young fanciful creature with anthropomorphic expressions that are easily recognizable to a human.

As long argued by animators, a character does not have to be realistic to be *believable*, i.e. to convey the illusion of life and to portray a thinking and feeling being (Thomas & Johnston 1981). We want people to treat Kismet as if it were a socially aware creature with thoughts, intents, desires, and feelings. Hence, believability is our goal. Realism is not necessary.

Audience Perception

A deep appreciation of audience perception is a fundamental issue for classical animation (Thomas & Johnston 1981) and has more recently been argued for by Bates (1994) in his work on believable agents. For sociable robots, this issue holds as well (albeit for different reasons), and we have experienced this first hand with Kismet. How the human perceives the robot establishes a set of expectations that fundamentally shape how the human interacts with it. This is not surprising as Reeves and Nass (1996) have demonstrated this phenomena for media characters, cartoon characters, as well as embodied conversation agents.

Being aware of these social factors can be played to advantage by establishing an



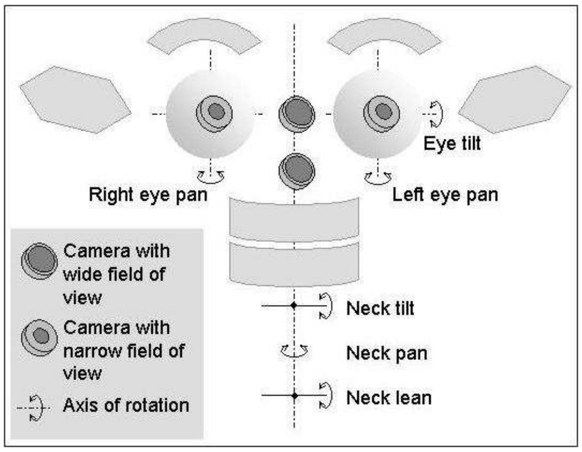


Figure 4-2: Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom (DoF) relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together along the vertical (tilt). The neck can turn the whole head horizontally and vertically, and can also crane forward. Two cameras with narrow "foveal" fields of view rotate with the eyes. Two central cameras with wide fields of view rotate with the neck. These cameras are unaffected by the orientation of the eyes. A human wears a microphone to talk to the robot.

appropriate set of expectations through robotic design. If done properly, people tend to naturally tune their behavior to the robot's current level of competence. This leads to a better quality of interaction for both robot and human.

4.2 The Hardware Design

Kismet is an expressive robotic creature with perceptual and motor modalities tailored to natural human communication channels. To facilitate a natural infant-caretaker interaction, the robot is equipped with input and output modalities roughly analogous to those of an infant (of course, missing many that infants have). For Kismet, the inputs include visual, auditory, and proprioceptive sensory inputs.

The motor outputs include vocalizations, facial expressions, and motor capabilities to adjust the gaze direction of the eyes and the orientation of the head. Note that these motor systems serve to steer the visual and auditory sensors to the source of the stimulus and can also be used to display communicative cues. The choice of these input and output modalities is geared to enable the system to participate in social interactions with a human, as opposed to traditional robot tasks such as manipulating physical objects or navigating through a cluttered space.

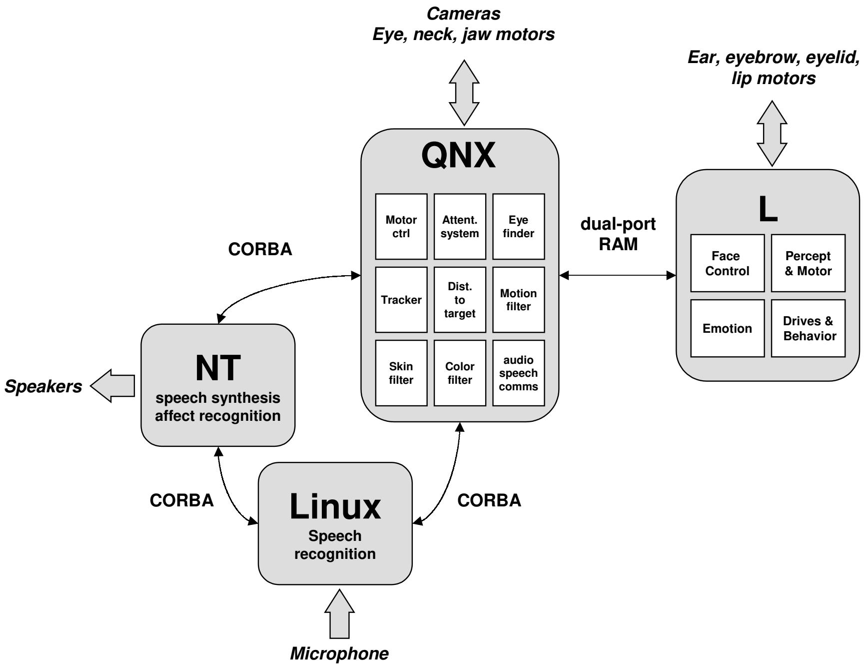


Figure 4-3: Our hardware and software control architectures have been designed to meet the challenge of real-time processing of visual signals (approaching 30 Hz) and auditory signals (8 kHz sample rate and frame windows of 10 ms) with minimal latencies (less than 500 ms). The high-level perception system, the motivation system, the behavior system, the motor skill system, and the face motor system execute on four Motorola 68332 microprocessors running L, a multi-threaded Lisp developed in our lab. Vision processing, visual attention and eye/neck control is performed by nine networked 400 MHz PCs running QNX (a real-time Unix operating system). Expressive speech synthesis and vocal affective intent recognition runs on a dual 450 MHz PC running NT, and the speech recognition system runs on a 500 MHz PC running Linux.

Vision System

The robot's vision system consists of four color CCD cameras mounted on a stereo active vision head. Two wide field of view (fov) cameras are mounted centrally and move with respect to the head. These are 0.25 inch CCD lipstick cameras with 2.2 mm lenses manufactured by Elmo Corporation. 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 are 0.5 inch CCD foveal cameras with an 8 mm focal length lenses, and are used for higher resolution post-attentional processing, such as eye detection.

Kismet has three degrees of freedom to control gaze direction and three degrees of freedom to control its neck (see figure 4-2). Each eye has an independent pan DoF, and both eyes share a common tilt DoF. The degrees of freedom are driven by Maxon

DC servo motors with high resolution optical encoders for accurate position control. This gives the robot the ability to move and orient its eyes like a human, engaging in a variety of human visual behaviors. This is not only advantageous from a visual processing perspective (as advocated by the active vision community such as Ballard (1989)), but humans attribute a communicative value to these eye movements as well. For instance, humans use gaze direction to infer whether a person is attending to them, to an object of shared interest, or not. This is important information when trying to carry out face-to-face interaction.

Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system. The PCs are connected together via 100MB Ethernet. There are frame grabbers and video distribution amplifiers to distribute multiple copies of a given image with minimal latencies. The cameras that are used to compute stereo measures are externally synchronized.

Auditory System

The caregiver can influence the robot's behavior through speech by wearing a small unobtrusive wireless microphone. This auditory signal is fed into a 500 MHz PC running Linux. The real-time, low-level speech processing and recognition software was developed at MIT by the Spoken Language Systems Group. These auditory features are sent to a dual 450 mHz PC running NT. The NT machine processes these features in real-time to recognize the spoken affective intent of the caregiver. The Linux and NT machines are connected via 100MB Ethernet to a shared hub and use CORBA for communication.

Expressive Motor System

Kismet is able to display a wide assortment of facial expressions that mirror its affective state, as well as produce numerous facial displays for other communicative purposes (Breazeal & Scassellati 1999*b*). Figure 4-4 illustrates a few examples. Fourteen of the face actuators are Futaba micro servos, which come in a light weight and compact package. Each ear has two degrees of freedom that enable each to elevate and rotate. This allows the robot to perk its ears in an interested fashion, or fold them back in a manner reminiscent of an angry animal. Each eyebrow has two degrees of freedom that enable each to elevate and to arc towards and away from the centerline. This allows the brows to lower and furrow in frustration, or to elevate upwards for surprise. Each eyelid can open and close independently, allowing the robot to wink an eye or blink both. The robot has four lip actuators, two for the upper lip corners and two for the lower lip corners. Each actuator moves a lip corner either up (to form a smile), or down (to form a frown). There is also a single degree of freedom jaw that is driven by a high performance DC servo motor from the MEI card. This level of performance is important for real-time lip synchronization with speech.

The face control software runs on a Motorola 68332 node running *L*. This processor is responsible for arbitrating between facial expression, real-time lip synchronization, communicative social displays, as well as behavioral responses. It communicates

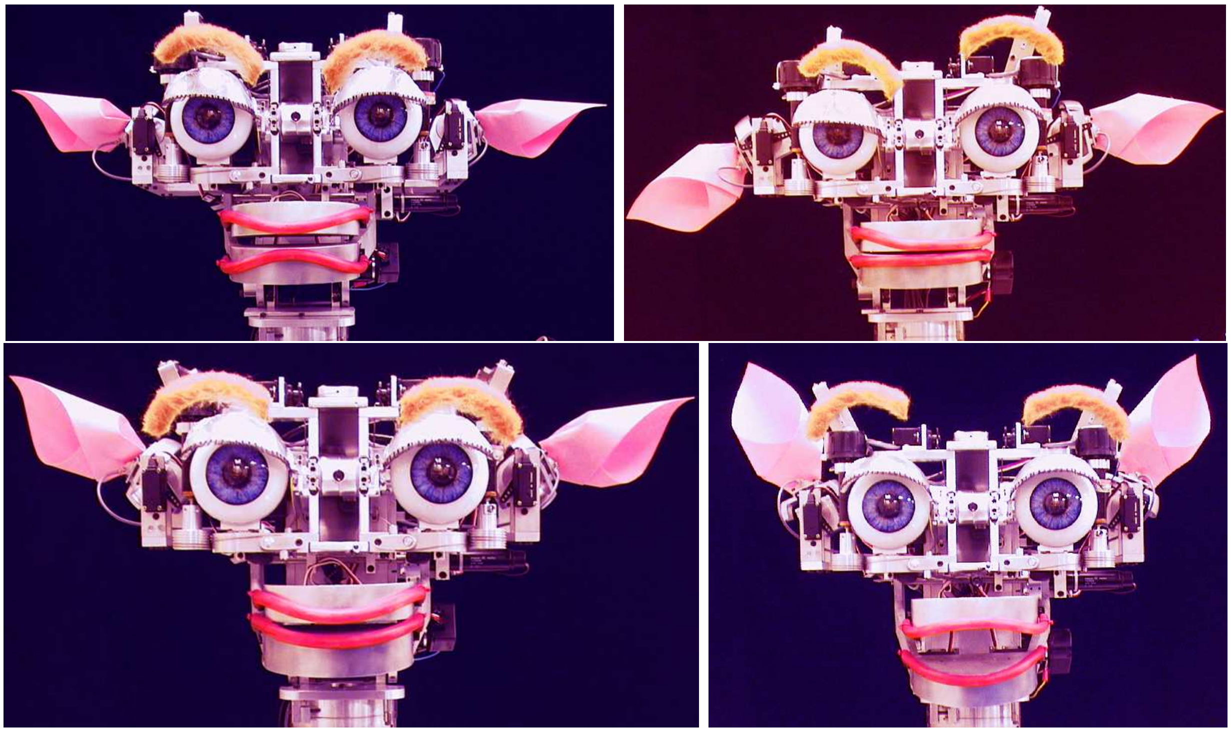


Figure 4-4: Some example facial expressions that illustrate the movement of Kismet's facial features. Top left is an expression of anger, top right is an expression of disapproval, lower left is an expression of happiness, and lower right is an expression of surprise.

to other 68332 nodes through a 16 KByte dual-ported RAM (DPRAM).

High Level Perception, Behavior, and Motivation, and Motor Skills

The high-level perception system, the behavior system, the motivation system, and the motor skills system run on the network of Motorola 68332 micro-controllers. Each of these systems communicate with the others by using threads if they are implemented on the same processor, or via DPRAM communication if implemented on different processors. Currently, each 68332 node can hook up to at most 8 DPRAMs. Another single DPRAM tethers the 68332 network to the network of PC machines via a QNX node.

Vocalization System

The robot's vocalization capabilities are generated through an articulatory synthesizer. The software, *DECtalk v4.5* sold by Digital Equipment Corporation, is based on the Klatt articulation synthesizer and it runs on a PC under Windows NT with a Creative Labs sound card. The parameters of the model are based on the physiological characteristics of the human articulatory tract. Although typically used as a text-to-speech system, it was chosen over other systems because it gives the user low level control over the vocalizations through physiologically based parameter settings. These parameters make it possible to convey affective information through vocalizations (Cahn 1990), and to convey personality by designing a custom voice for the robot. As such, Kismet's voice is that of a young child. The system also has the ability to play back files in a `.wav` format, so the robot could in principle produce infant-like vocalizations (laughter, coos, gurgles, etc.) that the synthesizer itself cannot generate.

Instead of relying on written text as an interface to the synthesizer, the software can accept strings of phonemes along with commands to specify the pitch and timing of the utterance. Hence, Kismet's vocalization system generates both phoneme strings and command settings, and says them in near real-time. The synthesizer also extracts phoneme and pitch information which are used to coordinate real-time lip synchronization. Ultimately, this capability would permit the robot to play and experiment with its own vocal tract, and to learn the effect these vocalizations have on human behavior. Kismet's voice is one of the most versatile instruments it has to interact with the caregiver.

4.3 Summary

Kismet is an expressive robotic creature with perceptual and motor modalities tailored to natural human communication channels. To facilitate a natural infant-caretaker interaction, the robot is equipped with visual, auditory, and proprioceptive sensory inputs. Its motor modalities consist of a high performance six degree of freedom active vision head supplemented with expressive facial features. 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 windows of 10 ms) with minimal latencies ($<$ 500 ms). These fifteen networked computers run the robot's synthetic nervous system that integrates perception, attention, motivations, behaviors, and motor acts.

Chapter 5

Overview of the Perceptual System

Human infants discriminate readily between social stimuli (faces, voices, etc.) and salient non-social stimuli (brightly colored objects, loud noises, large motion, etc.). For Kismet, the perceptual system is designed to discriminate a subset of both social and non-social stimuli from visual images as well as auditory streams. The specific percepts within each category (social verses non-social) are targeted for social exchanges. Specifically, the social stimuli are geared toward detecting the affective state of the caregiver, whether or not the caregiver is paying attention to the robot, and other people related percepts that are important during face-to-face exchanges such as the prosody of the caregiver's vocalizations. The non-social percepts are selected for their ability to command the attention of the robot. These are useful during social exchanges when the caregiver wants to direct the robot's attention to events outside pure face-to-face exchange. In this way, the caregiver can maneuver the interaction to be about things and events in the world, such as centering an interaction around playing with a specific toy.

5.1 Perceptual Abilities of Infants

From the earliest stages in development, infants treat people differently from other sources of stimulation in their environment. In their second month, reactions to things and people are so different that Trevarthen (1979) concludes that these two classes of objects must be distinct in the infant's awareness. They see physical objects as interesting sources of perceptual information and interact with them through grasping, chewing, kicking, et cetera. However, people are interacted with by facial expressions, vocalizations, and gestures. In fact, examinations for assessing normal infant development specifically characterize how infants interact with social stimuli and respond to non-social stimulation (Brazelton 1979). Such examinations attest to the infant's ability to distinguish people from other sources of stimulation.

5.1.1 Social Stimuli

Infants show a preference for social stimuli over non-social stimuli. They prefer even simple face-like stimuli over other pleasing stimuli such as a red shiny ball (Brazelton 1979). When encountering a human face, their face often softens, their eyes grow wide and eager, and they may crane their neck forward or make soft cooing sounds

(Trevarthen 1979). While gazing upon a face, they seem to explore its configuration while paying particular interest to the eyes and mouth (Trevarthen 1979). They seem to recognize when an adult is paying attention to them, and fuss when the adult fails to respond to their own attempts at engagement (Trevarthen 1979), (Tronick et al. 1979). During face-to-face exchanges with an adult, infants around five months of age show imitative accommodation to the pitch and duration of sounds, to facial expressions, and to various gestures such as tongue protrusion, mouth opening, and hand opening (Trevarthen 1979), (Meltzoff & Moore 1977). The perception of human sounds is acute in very young infants, and speech is reacted to with particular interest. In particular, the pitch characteristics of human voices are preferred to non-voice sounds (Trevarthen 1979). Even the mother's individual voice or manner of speaking is preferred early on (Trevarthen 1979), (Hauser 1996). Infants also seem capable of perceiving the affective state of the caregiver. The infant's mood can be affected by the mother's as conveyed both by facial expression or her speech (Trevarthen 1979).

5.1.2 Non-Social Stimuli

Much research in infant perception has been directed towards discovering what features of an object will make it naturally interesting for an infant. According to Newson (1979), "the most obviously attention-commanding objects are mobile, self-deforming, brightly colored, and noise emitting devices". Infants can discriminate color, and there seems to be a built in categorization for primary colors (red, green, blue, and yellow) (Trevarthen 1979). They have a preference for red which may assist them in finding a face or hand as light reflected onto the skin of all humans is reddish (Trevarthen 1979). Infants are particularly attentive to motion. It has been observed that, infants younger than six months will not attend to a brightly colored object in their visual field unless put into motion so as to appear "lively" (Newson 1979). They have coarse depth perception which starts developing after the first month (Tronick et al. 1979). They also demonstrate a strong response to periodic motion of an object in an otherwise inactive field (Trevarthen 1979). This may contribute to the infant's perception of people and their communication signals. In general there is close integration of rhythm between mother and infant during social exchanges, their coordinated action being synchronized about a common beat. This forms a turn-taking framework upon which the reciprocal exchange of complementary messages is based. Stern (1975) argues that repetitive acts of the caregiver, or stimulation that can be characterized as variations on a theme, is optimal for holding the infant's attention.

It should be noted that infants are also sensitive to the intensity of the impinging stimulation, and have a variety of mechanisms they employ to regulate their intensity (Brazelton 1979). The caregiver can use this to advantage, to either arouse or quiet the infant. For instance, speaking to an upset infant in a soft soothing tone will tend to quiet him. However, speaking to the infant in a high pitched staccato may build him up to crying. Sudden, intense stimuli may also cause the infant to shut out the stimulation, either by crying or clenching the eyes shut (Brazelton 1979).

5.2 Perceptual Limitations of Infants and their Consequences

The phenomenal world of the infant is quite limited compared to that of an adult. They have a slower rate of processing information. For instance, an infant may perceive a sequence of two visual events as only a single event (Tronick et al. 1979). They also have a narrower and shallower field of view as compared to adults. (Tronick et al. 1979). Hence, only objects within the infant's immediate vicinity serve to capture the attention of the infant. They have low visual acuity, and cannot perceive the same amount of detail in a visual scene as adults do (Tronick et al. 1979). In the auditory realm, infants cannot perceive many of the subtle variations of the adult tongue. Instead, they may very well perceive their mother's vocalizations as a single utterance, where prosody is the most salient feature (Fernald 1989), (Trehub & Trainor 1990). In section 2 we discussed how these limited capacities early in development actually facilitates the infant's learning and continued growth. Adult caregivers are aware of the infant's limitations, and alter their behavior to cater to the infant's current abilities.

During social exchanges with the infant, adults modify their actions to be more appropriate for the infant. Almost everything they do is exaggerated and slowed down. They vary the rate, intensity, amplitude, and quality of the action to benefit the infant (Tronick et al. 1979), (Trevarthen 1979). For instance, facial expressions become "baby faces" which are far more exaggerated than those used between adults. Their voice assumes "baby talk" characteristics where prosody and pronunciation are magnified (Hirsh-Pasek, Jusczyk, Cassidy, Druss & Kennedy 1987). They perform "baby movements" such as coming very close to the infant, orienting to face the baby, and moving their body both perpendicularly and parallel to the infant. These exaggerations seem to increase the information content of the caregiver's activities while facilitating the coordination of the infant's activities with those of the caregiver (Tronick et al. 1979).

By having the caregiver appropriately match her actions and displays to the infant's current abilities, the infant is able to function within his limitations. In this way, the infant can organize his actions based upon what he perceives and can practice his current capabilities in this context. However, as Tronick et al. (1979) points out, there is actually no way the caregiver can perfectly match her actions with the intentions or actions of the infant. Mis-matching during face-to-face interaction is bound to occur. It is doubtful that the infant can process all of the information the caregiver presents or is able to react to all of it. Nonetheless, this is a critical aspect of the environment to assure continued development. Mis-matching provides more complicated events to learn about. Hence, as the infant's capabilities develop at one level, there is an environment to develop into that slightly challenges him. An environment that is always perfectly matched to the infant's abilities would not allow for continued growth. Communication might be better at that moment, but there would be no impetus for it to improve and to become more elaborated. Hence, the normal social environment is the proper environment for both the maintenance and growth of the infant's skills.

This has important implications for how to design Kismet's perceptual system. Clearly we do not need to implement the ultimate, most versatile and complete perceptual system. Clearly, we do not need to develop a perceptual system that rivals the performance and sophistication of the adult. As argued above, this is not appropriate and would actually hinder development by overwhelming the robot with more perceptual information than the robot's synthetic nervous system could possibly handle or learn from. It is also inappropriate to place the robot in an overly simplified environment where it would ultimately learn and predict everything about that environment. There would be no impetus for continued growth. Instead, the perceptual system should start out as simple as possible, but rich enough to distinguish important social cues and interaction scenarios that are typical of caregiver-infant interactions. In the meantime, the caregiver must do her part to simplify the robot's perceptual task by slowing down and exaggerating her behavior in appropriate ways. She should repeat her behavior until she feels it has been adequately perceived by the robot. Hence the robot does not need to get the perception exactly right upon its first appearance. The challenge is to specify a perceptual system that can detect the right kinds of information at the right resolution.

5.3 Overview of the Perceptual System

As argued above, the robot does not necessarily need a perceptual system that rivals that of human adults. It can be simpler, more like that of an infant. Furthermore, at any one time there are often several interesting stimuli that the robot could respond to. We have found that a relatively broad and well integrated real-time perceptual system is critical.

The real-time constraint imposes some fairly stringent restrictions in the algorithms we use. As a result, they tend to be simple and of low resolution so that they can run quickly. One might characterize Kismet's perceptual system as being broad and simple where the perceptual abilities are robust enough and detailed enough for these early human-robot interactions. Deep and complicated perceptual algorithms certainly exist. However, as we have learned from human infants, there are developmental advantages to starting out broad and simple and allowing the perceptual, behavioral, and motor systems to develop in step. Kismet's initial perceptual system specification is designed to be roughly analogous to a human infant by implementing many of the perceptual abilities outlined above (and human infants certainly perceive more things than Kismet). Nonetheless, for an autonomous robot, it is quite a sophisticated perceptual system.

The perceptual system is decomposed into six subsystems (see figure 5-1). The development of Kismet's overall perceptual system is a large scale engineering endeavor which includes the efforts of many collaborators. We include citations wherever possible, although some of the work has yet to be published. Please see the acknowledgement section where we gratefully recognize the efforts of these researchers. We describe the visual attention system in chapter 6. We cover the affective speech

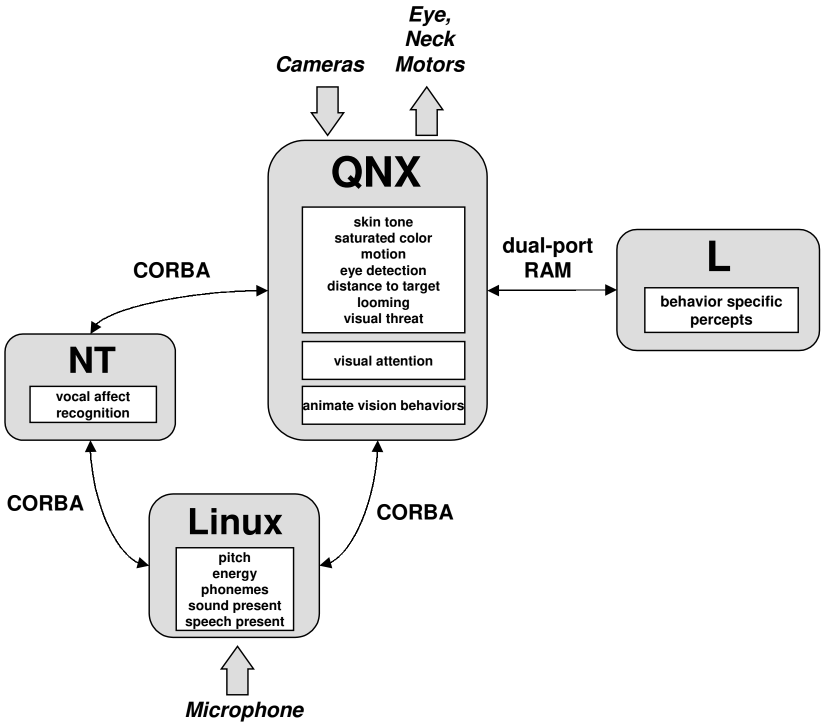


Figure 5-1: Schematic of Kismet's perceptual systems. See text.

recognition system in chapter 7. The behavior-specific and emotion-specific perceptions (organized about the social/non-social perceptual categories) are discussed in chapters 8 and 9. For the remainder of this chapter, we briefly outline the low-level perceptual abilities for visual and auditory channels.

5.4 Low-Level Visual Perception

Kismet's low-level visual perception system extracts a number of features that human infants seem to be particularly responsive toward. These low-level features were selected for their ability to help Kismet distinguish social stimuli (i.e. people, that is based on skin tone, eye detection, and motion) from non-social stimuli (i.e. toys, that is based on saturated color and motion), and to interact with each in interesting ways (often modulated by the distance of the target stimulus to the robot). There are a few perceptual abilities that serve self-protection responses. These include detecting looming stimuli as well as potentially dangerous stimuli (characterized by excessive motion close to the robot). We have previously reported an overview of Kismet's visual abilities in (Breazeal, Fitzpatrick, Edsinger & Scassellati 2000), (Breazeal & Scassellati 1999*a*), (Breazeal & Scassellati 1999*b*). Kismet's low-level visual features are as follows:

- Highly saturated color: red, blue, green, yellow. (Brian Scassellati)
- Colors representative of skin tone. (Paul Fitzpatrick)
- Motion detection. (Brian Scasselatti)
- Eye detection. (Aaron Edsinger)
- Distance to target. (Paul Fitzpatrick)
- Looming. (Paul Fitzpatrick)
- Threatening, very close, excessive motion. (Paul Fitzpatrick)

5.5 Low-Level Auditory Perception

Kismet's low-level auditory perception system extracts a number of features that are also useful for distinguishing people from other sound emitting objects such as rattles, bells, and so forth. The software runs in real-time and was developed at MIT by the Spoken Language Systems Group (*www.sls.lcs.mit.edu/sls*). Jim Glass and Lee Hetherington were tremendously helpful in tailoring the code for our specific needs and in assisting us to port this sophisticated speech recognition system to Kismet. The software delivers a variety of information that is used to distinguish speech-like sounds from non-speech sounds, to recognize vocal affect, and to regulate vocal turn-taking behavior. The phonemic information may ultimately be used to shape the robot's own vocalizations during imitative vocal games, and to enable the robot to acquire a proto-language from long term interactions with human caregivers. Kismet's low level auditory features are as follows:

- sound present
- speech present
- time stamped pitch tracking
- time stamped energy tracking
- time stamped phonemes

5.6 Summary

Kismet's perceptual system is designed to support a variety of important functions. Many aspects address behavioral and protective responses that evolution has endowed to living creatures so that they may behave and survive in the physical world. Given the perceptual richness and complexity of the physical world, we have implemented specific systems to explicitly organize this flood of information. By doing so, the robot can organize its behavior about a locus of attention.

The robot's perceptual abilities have been explicitly tailored to support social interaction with people and to support social learning/instruction processes. The robot must share enough of a perceptual world with humans so that communication can take place. The robot must be able to perceive the social cues that people naturally and intuitively use to communicate with it. The robot and a human should share enough commonality in those features of the perceptual world that are of particular interest, so that both are drawn to attend to similar events and stimuli. Meeting these criteria enables a human to naturally and intuitively direct the robot's attention to interesting things in order to establish shared reference. It also allows a human to communicate affective assessments to the robot which could make social referencing possible. Ultimately these abilities will play an important role in the robot's social development, as they do for the social development of human infants.

Chapter 6

The Vision System: Attention and Low Level Perception

> *Certain types of spontaneously occurring events may momentarily dominate his attention or cause him to react in a quasi-reflex manner, but a mere description of the classes of events which dominate and hold the infants' sustained attention quickly leads one to the conclusion that the infant is biologically tuned to react to person-mediated events. These being the only events he is likely to encounter which will be phased, in their timing, to coordinate in a non-predictable or non-redundant way with his own activities and spontaneous reactions.* (Newson 1979)

6.1 Human Infant Attention

The ability for infants to direct their attention to salient stimuli is present at the earliest stages of development. It plays a critical role in social interactions with adults as well as learning during these exchanges. The caregiver initiates social exchange with her infant by first getting his attention so that they can establish mutual regard. During the exchange she may direct his attention to other objects and events, such as directing the interaction to be about a particular toy. If his attention wanes, she will try to re-engage him by making either herself or the toy more salient. She may shake the toy, she may assume a staccato manner of speech, etc. By directing the infant's attention to the most salient aspects of a task she would like him to learn, she facilitates the learning problem presented to him. This is one important form of scaffolding. By directing the infant's attention to a desired stimulus, the caregiver can establish joint reference.

6.2 Design Issues of Attention Systems for Robots that Interact with People

Above, we discussed a number of stimuli that infants have a bias to attend to. They can be categorized according to visual versus auditory sensory channels (among others), and whether they correspond to social or non-social forms of stimulation. From these, we can outline those specific percepts that have been implemented on Kismet because we deem them important for social interaction (of course, there are other important features that have yet to be implemented). The attention system is designed to direct the robot's attention to those visual sensory stimuli that can be characterized by these selected perceptions. Later extensions to the mechanism could include other perceptual features.

To benefit communication and social learning, it is important that both robot and human find the same sorts of perceptual features interesting. Otherwise there will be a mismatch between the sorts of stimuli and cues that humans use to direct the robot's attention versus those that attract the robot's attention. For instance, if designed improperly it could prove to be very difficult to achieve joint reference with the robot. Even if the human could learn what attracts the robot's attention, this defeats the goal of allowing the person to use natural and intuitive cues. Designing for the set of perceptual cues that human infants find salient allows us to implement an initial set that are evolutionary significant for humans.

Kismet's attention system acts to direct computational and behavioral resources toward salient stimuli and to organize subsequent behavior around them. In an environment suitably complex for interesting learning, perceptual processing will invariably result in many potential target stimuli. Its critical that this be accomplished in real-time. In order to determine where to assign resources, the attention system must incorporate raw sensory saliency with task-driven influences.

6.3 Specification of the Attention system

The attention system is shown in figure 6-1 and is heavily inspired by the *Guided Search v2.0* system of Wolfe (1994). Wolfe proposed this work as a model for human visual search behavior. We have extended it to account for moving cameras, dynamically changing task-driven influences, and habituation effects (Breazeal & Scassellati 1999*a*).

The attention system is a two stage system. The first stage is a *pre-attentive*, massively parallel stage that processes information about basic visual features (i.e., color, motion, depth cues, etc.) across the entire visual field (Triesman 1986). For Kismet, these bottom-up features include highly saturated color, motion, and colors representative of skin tone. The second stage is a *limited capacity* stage which performs other more complex operations, such as facial expression recognition, eye detection, or object identification, over a localized region of the visual field. These limited capacity processes are deployed serially from location to location under attentional control. This is guided by the properties of the visual stimuli processed by the first stage (an exogenous contribution), by task-driven influences, and by habituation effects (both are endogenous contributions). The habituation influence provides Kismet with a primitive attention span. For Kismet, the second stage includes an eye-detector that operates over the foveal image, and a target proximity estimator that operates on the stereo images of the two central wide field of view (FoV) cameras. Figure 6-1 shows an overview of the attention system which we describe below.

All four factors influence the direction of Kismet's gaze. This in turn determines

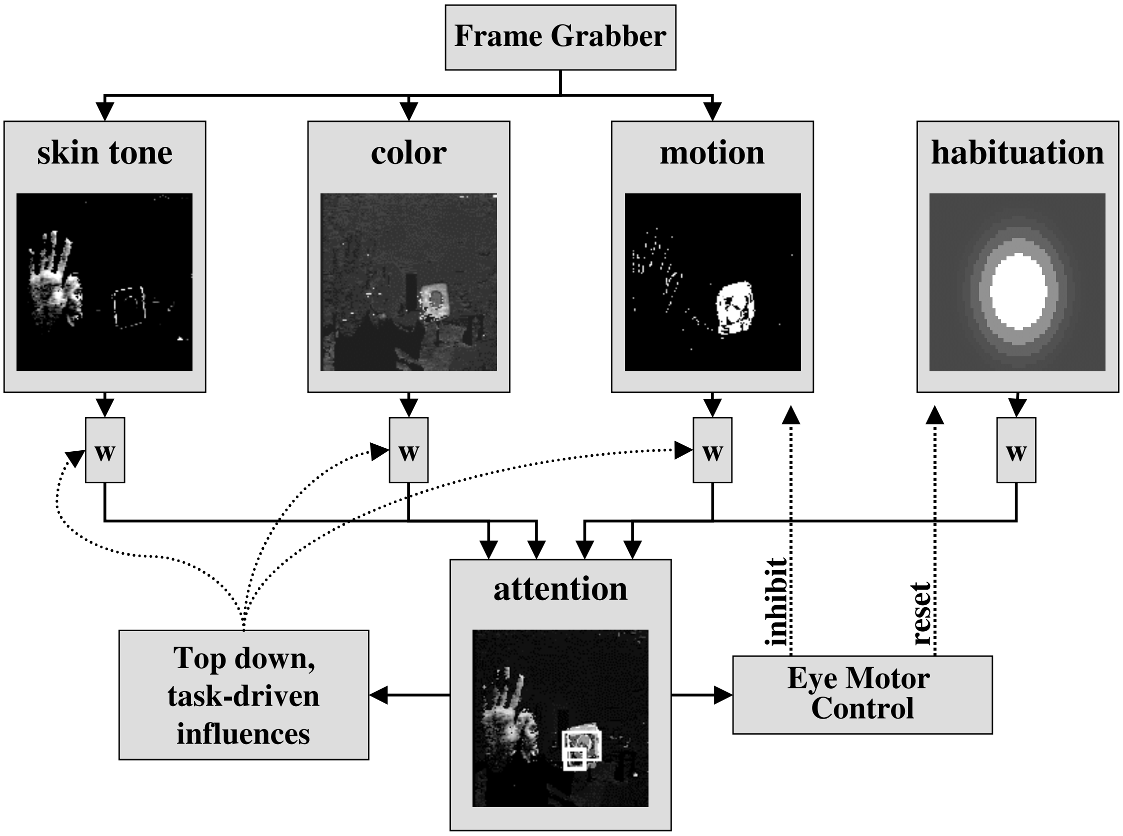


Figure 6-1: The robot's attention is determined by a combination of low-level perceptual stimuli. The relative weightings of the stimuli are modulated by high-level behavior and motivational influences. A sufficiently salient stimulus in any modality can pre-empt attention, similar to the human response to sudden motion. All else being equal, larger objects are considered more salient than smaller ones. The design is intended to keep the robot responsive to unexpected events, while avoiding making it a slave to every whim of its environment. With this model, people intuitively provide the right cues to direct the robot's attention (shake object, move closer, wave hand, etc.). Displayed images were captured during a behavioral trial session.

the robot's subsequent perception, which ultimately feeds back to behavior. Hence the robot is in a continuous cycle of behavior influencing what is perceived and perception influencing subsequent behavior.

6.4 Bottom-up contributions: Computing Feature Maps

The purpose of the first massively parallel stage is to identity locations that are worthy of further attention. This is considered to be a *bottom-up* or stimulus-driven contribution. Raw sensory saliency cues are equivalent to those "pop-out" effects studied by Triesman (1986), such as color intensity, motion, and orientation for visual stimuli. As such, it serves to bias attention toward distinctive items in the visual field, and will not guide attention if the feature properties of that item are not inherently salient.

This contribution is computed from a series of *feature maps* which are updated in parallel over the entire visual field (of the wide fov camera) for a limited set of basic visual features. There is a separate feature map for each basic feature (for Kismet these correspond to color, motion, and skin tone) and each map is topographically organized and in retinotopic coordinates. The computation of these maps is described below. The value of each location is called the *activation level* and represents the saliency of that location in the visual field with respect to the other locations. In our implementation, the overall bottom-up contribution comes from combining the results of these feature maps in a weighted sum.

The video signal from each of Kismet's cameras is digitized by one of the 400MHz nodes with frame grabbing hardware. The image is then subsampled and averaged to an appropriate size. Currently, we use an image size of 128×128, which allows us to complete all of the processing in near real-time. To minimize latency, each feature map is computed by a separate 400MHz processor (each of which also has additional computational task load). All of the feature detectors discussed here can operate at multiple scales.

Color saliency feature map

One of the most basic and widely recognized visual features is color. Our models of color saliency are drawn from the complementary work on visual search and attention from Itti, Koch & Niebur (1998). The incoming video stream contains three 8-bit color channels (r, g, and b) which are transformed into four color-opponency channels (r', g', b', and y'). Each input color channel is first normalized by the luminance l (a weighted average of the three input color channels):

$$r_n = \frac{255}{3} \cdot \frac{r}{l} \qquad g_n = \frac{255}{3} \cdot \frac{g}{l} \qquad b_n = \frac{255}{3} \cdot \frac{b}{l} \tag{6.1}$$

These normalized color channels are then used to produce four opponent-color channels:

$$r' = r_n - (g_n + b_n)/2 \tag{6.2}$$
$$g' = g_n - (r_n + b_n)/2 \tag{6.3}$$
$$b' = b_n - (r_n + g_n)/2 \tag{6.4}$$
$$y' = \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\| \tag{6.5}$$

The four opponent-color channels are clamped to 8-bit values by thresholding. While some research seems to indicate that each color channel should be considered individually (Nothdurft 1993), we choose to maintain *all* of the color information in a *single* feature map to simplify the processing requirements (as does Wolfe (1994) for more theoretical reasons). The result is a 2-D map where pixels containing a bright, saturated color component (red, green, blue, and yellow) increases the intensity value of that pixel. We have found the robot to be particularly sensitive to bright red, green, yellow, blue, and even orange. Figure 6-1 gives an example of the color feature map when the robot looks at a brightly colored block.

Motion Saliency Feature Maps

In parallel with the color saliency computations, a second processor receives input images from the frame grabber and computes temporal differences to detect motion. Motion detection is performed on the wide field of view, which is often at rest since it does not move with the eyes. The incoming image is converted to grayscale and placed into a ring of frame buffers. A raw motion map is computed by passing the absolute difference between consecutive images through a threshold function $\mathcal{T}$:

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|) \tag{6.6}$$

This raw motion map is then smoothed with a uniform 7×8 field. The result is a binary 2-D map where regions corresponding to motion have a high intensity value. The motion saliency feature map is computed at 25-30 Hz by a single 400MHz processor node. Figure 6-1 gives an example of the motion feature map when the robot looks at a toy block that is being shaken.

Skin tone feature map

Colors consistent with skin are also filtered for (see figure 6-1). This is a computationally inexpensive means to rule out regions which are unlikely to contain faces or hands. A large fraction of pixels on faces will pass these tests over a wide range of lighting conditions and skin color. Pixels that pass these tests are weighted according to a function learned from instances of skin tone from images taken by Kismet's cameras. See figure 6-2. In our implementation, a pixel is *not* skin-toned if:

- $r < 1.1 \cdot g$, the red component fails to dominate green sufficiently
- $r < 0.9 \cdot b$, the red component is excessively dominated by blue

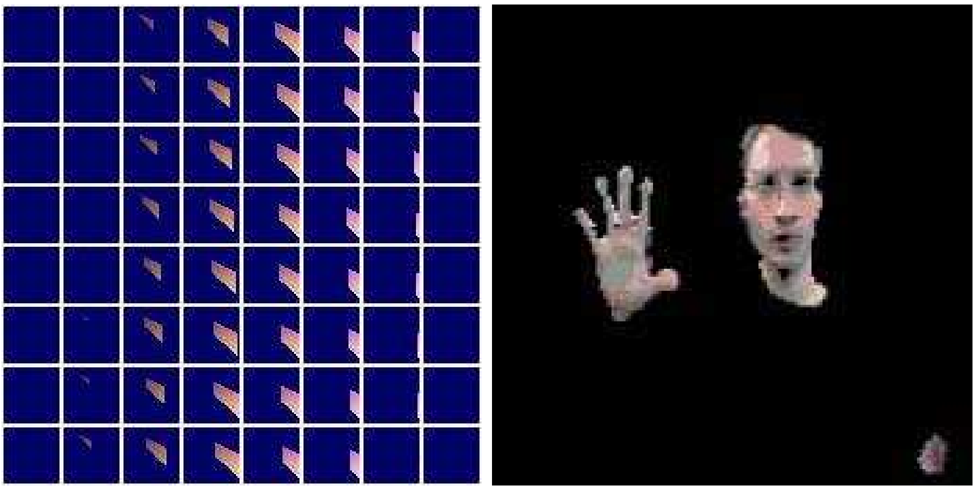


Figure 6-2: The skin tone filter responds to 4.7% of possible (R, G, B) values. Each grid element in the figure to the left shows the response of the filter to all values of red and green for a fixed value of blue. Within a cell, the x-axis corresponds to red and the y-axis corresponds to green. The image to the right shows the filter in operation. Typical indoor objects that may also be consistent with skin tone include wooden doors, cream walls, etc.

- $r > 2.0 \cdot max(g, b)$, the red component completely dominates both blue and green
- $r < 20$, the red component is too low to give good estimates of ratios
- $r > 250$, the red component is too saturated to give a good estimate of ratios

6.5 Top-down contributions: task-based influences

For a goal achieving creature, the behavioral state should also bias what the creature attends to next. For instance, when performing visual search, humans seem to be able to preferentially select the output of one broadly tuned channel per feature (e.g. "red" for color and "shallow" for orientation if searching for red horizontal lines) (Kandel, Schwartz & Jessell 2000).

In our system these top-down, behavior-driven factors modulate the output of the individual feature maps before they are summed to produce the bottom-up contribution. This process selectively enhances or suppresses the contribution of certain features, but does not alter the underlying raw saliency of a stimulus (Niedenthal & Kityama 1994). To implement this, the bottom-up results of each feature map are each passed through a filter (effectively a gain). The value of each gain is determined by the active behavior. These modulated feature maps are then summed to compute the overall attention activation map.

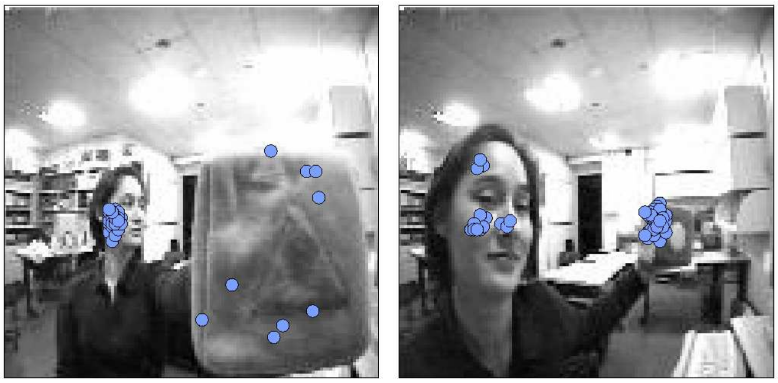


Figure 6-3: Effect of gain adjustment on looking preference. Circles correspond to fixation points, sampled at one second intervals. On the left, the gain of the skin tone filter is higher. The robot spends more time looking at the face in the scene (86% face, 14% block). This bias occurs despite the fact that the face is dwarfed by the block in the visual scene. On the right, the gain of the color saliency filter is higher. The robot now spends more time looking at the brightly colored block (28% face, 72% block).

This serves to bias attention in a way that facilitates achieving the goal of the active behavior. For example, if the robot is searching for social stimuli, it becomes sensitive to skin tone and less sensitive to color. Behaviorally, the robot may encounter toys in its search, but will continue until a skin toned stimulus is found (often a person's face). Figure 6-3 illustrates how gain adjustment biases what the robot finds to be more salient.

As shown in Figure 6-4, the skin tone gain is enhanced when the `seek-people` behavior is active and is suppressed when the `avoid-people` behavior is active. Similarly, the color gain is enhanced when the `seek-toys` behavior is active, and suppressed when the `avoid-toys` behavior is active. Whenever the `engage-people` or `engage-toys` behaviors are active, the face and color gains are restored to slightly favor the desired stimulus. Weight adjustments are constrained such that the total sum of the weights remains constant at all times.

6.6 Computing the Attention Activation Map

The attention activation map can be thought of as an activation "landscape" with higher hills marking locations receiving substantial bottom-up or top-down activation. The purpose of the attention activation map (using the terminology of Wolfe) is to direct attention, where attention is attracted to the highest hill. Hence, the greater the activation at a location, the more likely it is that the attention will be directed to that location. Note that by using this approach, the locus of activation contains no information as to its source (e.g. a high activation for color looks the same as high activation for motion information). Hence, the activation map makes it possible to guide attention based on information from more than one feature (such

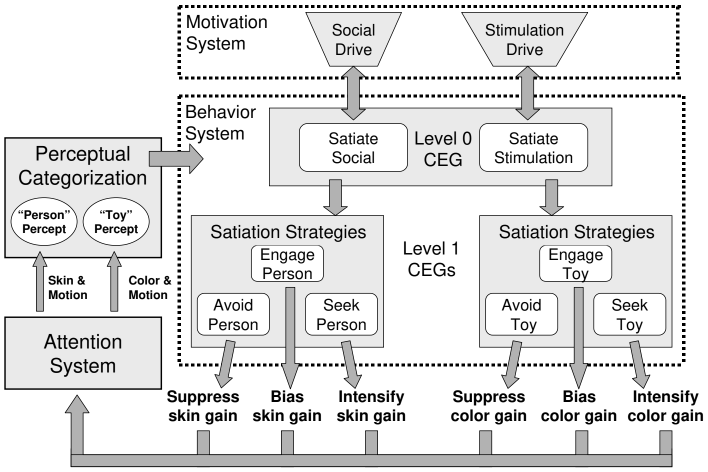


Figure 6-4: Schematic of behaviors relevant to attention. The activation of a particular behavior depends on both perceptual factors and motivation factors. The perceptual factors come from post attentive processing of the target stimulus into behaviorally relevant percepts (called *releasers* as described in chapter 9). The drives within the motivation system have an indirect influence on attention by influencing the behavioral context. The behaviors at Level 1 of the behavior system directly manipulate the gains of the attention system to benefit their goals. Through behavior arbitration, only one of these behaviors is active at any time. These behaviors are further elaborated in deeper levels of the behavior system.

as a conjunction of features).

To prevent drawing attention to non-salient regions, the attention activation map is thresholded to remove noise values, and normalized by the sum of the gains. Connected object regions are extracted using a grow-and-merge procedure with 4-connectivity (Horn 1986). To further combine related regions, any regions whose bounding boxes have a significant overlap are also merged. The attention process runs at 20 Hz on a single 400 MHz processor.

Statistics on each region are then collected, including the centroid, bounding box, area, average attention activation score, and average score for each of the feature maps in that region. The tagged regions that are large enough (having an area of at least thirty pixels) are sorted based upon their average attention activation score. The attention process provides the top three regions to both the eye motor control system and the behavior and motivational systems.

The most salient region is the new visual target. The individual feature map scores of the target are passed onto higher level perceptual stages where these features are combined to form behaviorally meaningful percepts. Hence the robot's subsequent behavior is organized about this locus of attention.

6.7 Attention Drives Eye Movement

Gaze direction is a powerful social cue that people use to determine what others are interested in. By directing the robot's gaze to the visual target, the person interacting with the robot can accurately use the robot's gaze as an indicator of what the robot is indeed attending to. This greatly facilitates the interpretation and readability of the robot's behavior, since the robot reacts specifically to the thing that it is looking at.

The eye motor control system uses the centroid of the most salient region as the target of interest. The eye motor control process acts on the data from the attention process to center the eyes on an object within the visual field. Using a data-driven mapping between image position and eye position, the retinotopic coordinates of the target's centroid are used to compute where to look next (Scassellati 1998). Each time that the neck moves, the eye/neck motor process sends two signals. The first signal inhibits the motion detection system for approximately 600 msec, which prevents self-motion from appearing in the motion feature map. The second signal resets the habituation state, described in the next section. We save a detailed discussion of how the motor component from the attention system is integrated into the rest of Kismet's visual behavior (such as smooth pursuit, looming, etc.) for chapter 13.

6.8 Habituation Effects

To build a believable creature, the attention system must also implement habituation effects. Infants respond strongly to novel stimuli, but soon habituate and respond less as familiarity increases (Carey & Gelman 1991). This acts both to keep the infant from being continually fascinated with any single object and to force the caregiver to continually engage the infant with slightly new and interesting interactions. For a robot, a habituation mechanism removes the effects of highly salient background objects that are not currently involved in direct interactions as well as placing requirements on the caregiver to maintain interaction with different kinds of stimulation.

To implement habituation effects, a *habituation filter* is applied to the activation map over the location currently being attended to. The habituation filter effectively decays the activation level of the location currently being attended to, making other locations of lesser activation bias attention more strongly.

The habituation function can be viewed as a feature map that initially maintains eye fixation by increasing the saliency of the center of the field of view and slowly decays the saliency values of central objects until a salient off-center object causes the neck to move. The habituation function is a Gaussian field $G(x, y)$ centered in the field of view with peak amplitude of 255 (to remain consistent with the other 8-bit values) and $\theta = 50$ pixels. It is combined linearly with the other feature maps using the weight

$$w = W \cdot max(-1, 1 - \Delta t/\tau) \tag{6.7}$$

where w is the weight, Δt is the time since the last habituation reset, τ is a time constant, and W is the maximum habituation gain. Whenever the neck moves, the habituation function is reset, forcing w to W and amplifying the saliency of central objects until a time τ when $w = 0$ and there is no influence from the habituation map. As time progresses, w decays to a minimum value of $-W$ which suppresses the saliency of central objects. In the current implementation, we use a value of $W = 10$ and a time constant $\tau = 5$ seconds. When the robot's neck shifts, the habituation map is reset, allowing that region to be re-visited after some period of time.

6.9 Second Stage Processing

Once the attention system has selected regions of the visual field that are potentially behaviorally relevant, more intensive computation can be applied to these regions than could be applied across the whole field. Searching for eyes is one such task. Locating eyes is important to us for engaging in eye contact. We search for eyes after the robot directs its gaze to a locus of attention so that a relatively high resolution image of the area being searched is available from the narrow field of view cameras (Figure 6-5).

Once the target of interest has been selected, we also estimate its proximity to the robot using a stereo match between the two central wide fov cameras. Proximity is an important factor for interaction. Things closer to the robot should be of greater interest. It is also useful for interaction at a distance. For instance, a person standing too far from Kismet for face-to-face interaction may be close enough to be beckoned closer. Clearly the relevant behavior (beckoning or playing) is dependent on the proximity of the human to the robot.

6.9.1 Eye Detection

Eye-detection in a real-time robotic domain is computationally expensive and prone to error due to the large variance in head posture, lighting conditions and feature scales. Aaron Edsinger developed an approach based on successive feature extraction, combined with some inherent domain constraints, to achieve a robust and fast eye-detection system for Kismet (Breazeal et al. 2000). First, a set of feature filters are applied successively to the image in increasing feature granularity. This serves to reduce the computational overhead while maintaining a robust system. The successive filter stages are:

- Detect skin colored patches in the image (abort if this does not pass above a threshold).
- Scan the image for ovals and characterize its skin tone for a potential face.
- Extract a sub-image of the oval and run a ratio template over it for candidate eye locations (Sinha 1994), (Scassellati 1998).

Figure 6-5: Sequence of foveal images with eye detection. The eye detector actually looks for the region between the eyes. It has decent performance over a limited range of distances and face orientations. The box indicates a possible face has been detected (being both skin toned and oval in shape). The small cross locates the region between the eyes.

- For each candidate eye location, run a pixel based multi-layer perceptron (previously trained) on the region to recognize shading characteristic of the eyes and the bridge of the nose.

By doing so, the set of possible eye-locations in the image is reduced from the previous level based on a feature filter. This allows the eye detector to run in real time on a 400MHz PC. The methodology assumes that the lighting conditions allow the eyes to be distinguished as dark regions surrounded by highlights of the temples and the bridge of the nose, that human eyes are largely surrounded by regions of skin color, that the head is only moderately rotated, that the eyes are reasonably horizontal, and that people are within interaction distance from the robot (3 to 7 feet).

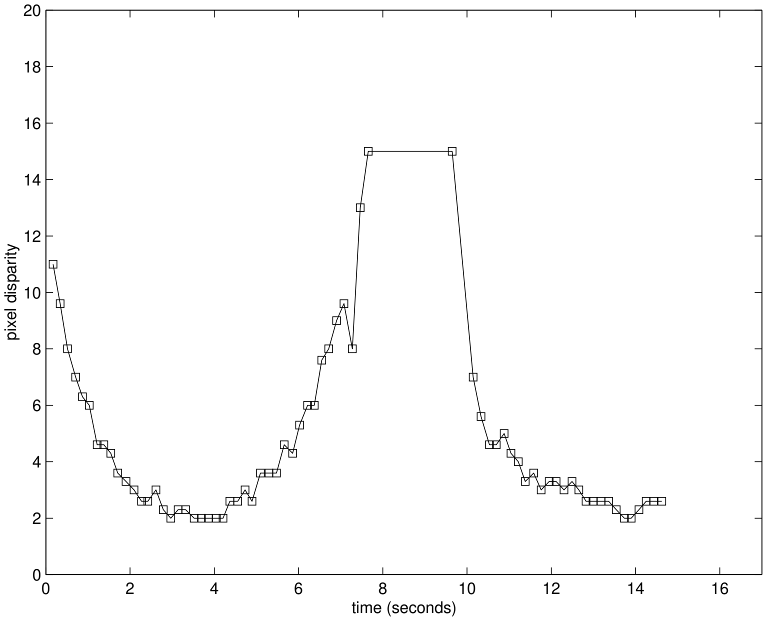


Figure 6-6: This plot illustrates how the target proximity measure varies with distance. The subject begins by standing approximately two feet away from the robot ($t = 0$). He then steps back to a distance of about seven feet ($t = 4$). This is on the outer periphery of the robot's interaction range. Beyond this distance, the robot does not reliably attend to the person as the target of interest as other things are often more salient. The subject then approaches the robot to a distance of 3 inches from its face ($t = 8$ to $t = 10$). The loom detector is firing, which is the plateau in the graph. At $t = 10$ the subject then backs away and leaves the scene.

6.9.2 Proximity Estimation

Given a target in the visual field, proximity is computed from a stereo match between the two wide cameras. The target in the central wide camera is located within the lower wide camera by searching along epipolar lines for a sufficiently similar patch of pixels, where similarity is measured using normalized cross-correlation. This matching process is repeated for a collection of points around the target to confirm that the correspondences have the right topology. This allows many spurious matches to be rejected. Figure 6-6 illustrates how this metric changes with distance from the robot. It is reasonably monotonic, but subject to noise. It is also quite sensitive to the orientations of the two wide center cameras.

6.9.3 Loom Detection

The loom calculation makes use of the two cameras with wide fields of view. These cameras are parallel to each other, so when there is nothing in view that is close to the cameras (relative to the distance between them), their output tends to be very similar. A close object, on the other hand, projects very differently on to the two cameras, leading to a large difference between the two views.

By simply summing the pixel-by-pixel differences between the images from the two cameras, we extract a measure which becomes large in the presence of a close object. Since Kismet's wide cameras are quite far from each other, much of the room and furniture is close enough to introduce a component into the measure which will change as Kismet looks around. To compensate for this, the measure is subject to rapid habituation. This has the side-effect that a slowly approaching object will not be detected - which is perfectly acceptable for a loom response.

6.9.4 Threat Detection

A nearby object (as computed above) along with large but concentrated movement in the wide fov is treated as a threat by Kismet. The amount of motion corresponds to the amount of activation of the motion map. Since the motion map may also become very active during ego-motion, this response is disabled for the brief intervals during which Kismet's head is in motion. As an additional filtering stage, the ratio of activation in the peripheral part of the image versus the central part is computed to help reduce the number of spurious threat responses due to ego-motion. This filter thus looks for concentrated activation in a localized region of the motion map, whereas self induced motion causes activation to smear evenly over the map.

6.10 Results and Evaluation

The overall attention system runs at 20 Hz on several 400 MHz processors. In this section, we evaluate its behavior with respect to directing Kismet's attention to task-relevant stimuli. We also examine how easy it is people to direct the robot's attention to a specific target stimulus, and to determine when they have been successful in doing so.

6.10.1 Effect of Gain Adjustment on Saliency

In section 6.5, we described how the active behavior can manipulate the relative contributions of the bottom-up processes to benefit goal achievement. Figure 6-7 illustrates how the skin tone, motion, and color gains are adjusted as a function of drive intensity, the active behavior, and the nature and quality of the perceptual stimulus.

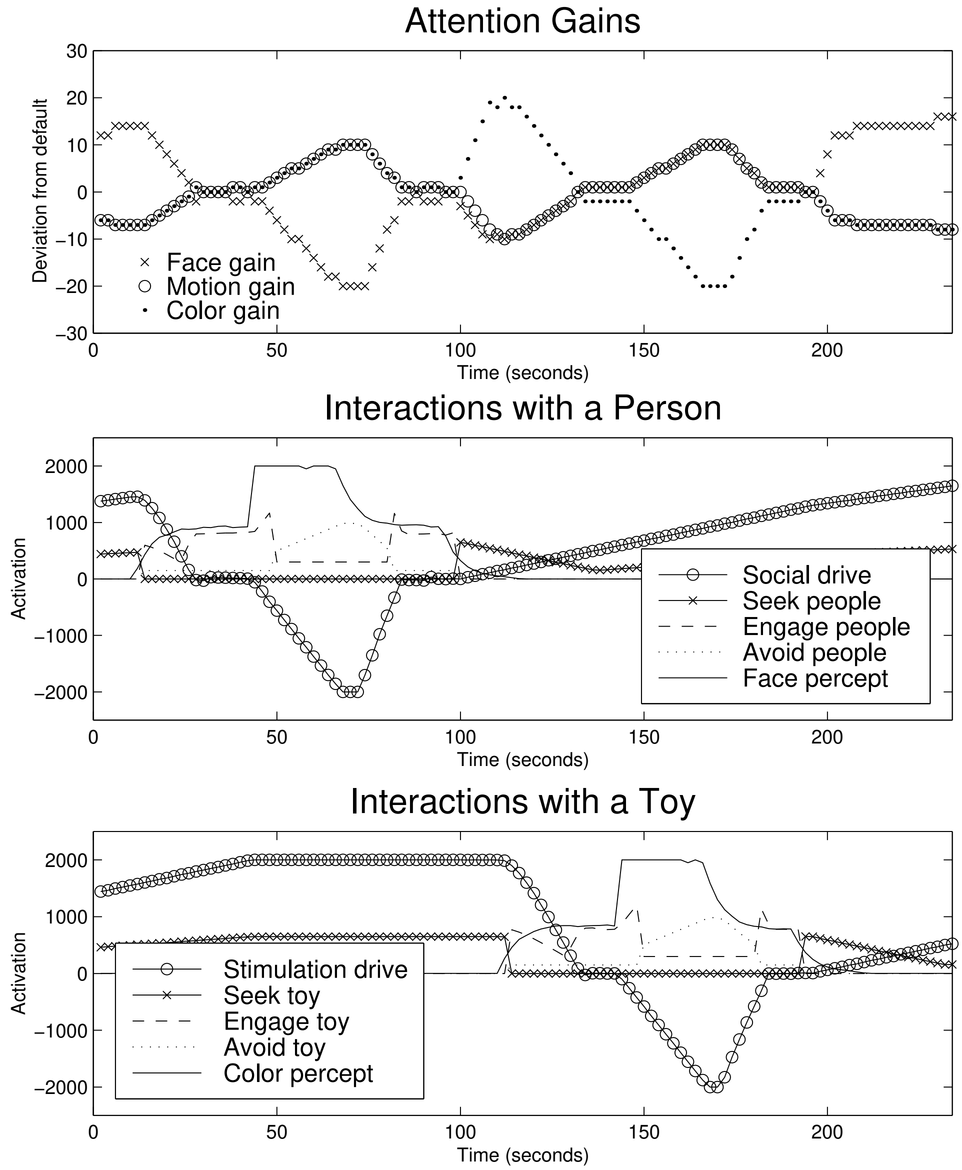


Figure 6-7: Changes of the skin tone (face), motion, and color gains from top-down motivational and behavioral influences (top). On the left half of the top figure, the gains change with respect to person-related behaviors (middle). On the right half of the top figure, the gains change with respect to toy-related behaviors (bottom).

As shown in figure 6-7, when the `social-drive` is activated by face stimuli (middle), the skin tone gain is influenced by the `seek-people` and `avoid-people` behaviors. The effects on the gains are shown on the left side of the top plot. When the `stimulation-drive` is activated by color stimuli (bottom), the color gain is influenced by the `seek-toys` and `avoid-toys` behaviors. This is shown to the right of the top plot. Seeking people out results in enhancing the face gain and avoiding people results in suppressing the face gain. The color gain is adjusted in a similar fashion when toy-oriented behaviors are active (enhancement when seeking out, suppression during avoidance). The middle plot shows how the `social-drive` and the quality of social stimuli determine which people-oriented behavior is activated. The bottom plot shows how the `stimulation-drive` and the quality of toy stimuli determine which toy-oriented behavior is active. All parameters shown in these plots were recorded during the same 4 minute period.

The relative weighting of the attention gains are empirically set to satisfy behavioral performance as well as to satisfy social interaction dynamics. For instance, when engaging in visual search, the attention gains are set so that there is a strong preference for the target stimulus (skin tone when searching for social stimuli like people, saturated color when searching for non-social stimuli like toys). As shown in left in figure 6-3, a distant face has greater overall saliency than a nearby toy if the robot is actively looking for skin toned stimuli. Similarly, as shown to the right in figure 6-3, a distant toy has greater overall saliency than a near by face when the robot is actively seeking out stimuli of highly saturated color.

Behaviorally, the robot will continue to search upon encountering a static object of high raw saliency but of the wrong feature. Upon encountering a static object possessing the right saliency feature, the robot successfully terminates search and begins to visually engage the object. However, the search behavior sets the attention gains to allow Kismet to attend to a stimulus possessing the wrong saliency feature if it is also supplemented with motion. Hence, if a person really wants to attract the robot's attention to a specific target, which the robot is not actively seeking out, then he/she is still able to do so.

During engagement, the gains are set so that Kismet slightly prefers those stimuli possessing the favored feature. However, if a stimulus of the favored feature is not present, a stimulus possessing the unfavored feature is sufficient to attract the robot's attention. Hence, while in engagement, the robot can satiate other motivations in an opportunistic manner when the desired stimulus is not present. However, if the robot is unable to satiate a specific motivation for a prolonged time, the motive to engage that stimuli will increase until the robot eventually breaks engagement to preferentially search for the desired stimulus.

6.10.2 Effect of Gain Adjustment on Looking Preference

Figure 6-8 illustrates how top-down gain adjustments combine with bottom-up habituation effects to bias the robot's gaze. When the `seek-people` behavior is active, the skin tone gain is enhanced and the robot prefers to look at a face over a colorful toy. The robot eventually habituates to the face stimulus and switches gaze briefly to the toy stimulus. Once the robot has moved its gaze away from the face stimulus, the

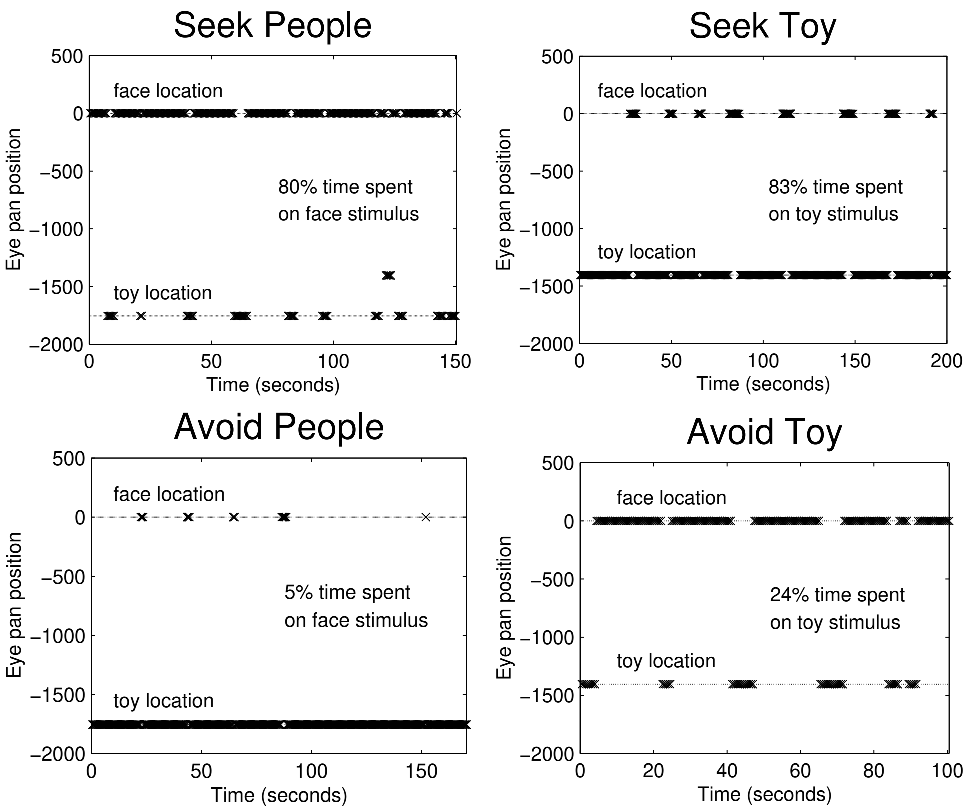


Figure 6-8: Preferential looking based on habituation and top-down influences. These plots illustrate how Kismet's preference for looking at different types of stimuli (a person's face verses a brightly colored toy) varies with top-down behavior and motivational factors. See text.

habituation is reset and the robot rapidly re-acquires the face. In one set of behavioral trials when `seek-people` was active, the robot spent 80% of the time looking at the face. A similar affect can be seen when the `seek-toy` behavior is active — the robot prefers to look at a toy over a face 83% of the time.

The opposite effect is apparent when the `avoid-people` behavior is active. In this case, the skin tone gain is suppressed so that faces become less salient and are more rapidly affected by habituation. Because the toy is relatively more salient than the face, it takes longer for the robot to habituate. Overall, the robot looks at faces only 5% of the time when in this behavioral context. A similar scenario holds when the robot's `avoid-toy` behavior is active — the robot looks at toys only 24% of the time.

6.10.3 Socially Manipulating Attention

Figure 6-9 shows an example of the attention system in use, choosing stimuli in a complex scene that are potentially behaviorally relevant. The attention system runs all the time, even when it is not controlling gaze direction, since it determines the perceptual input to which the motivational and behavioral systems respond. Because the robot attends to a subset of the same cues that humans find interesting, people

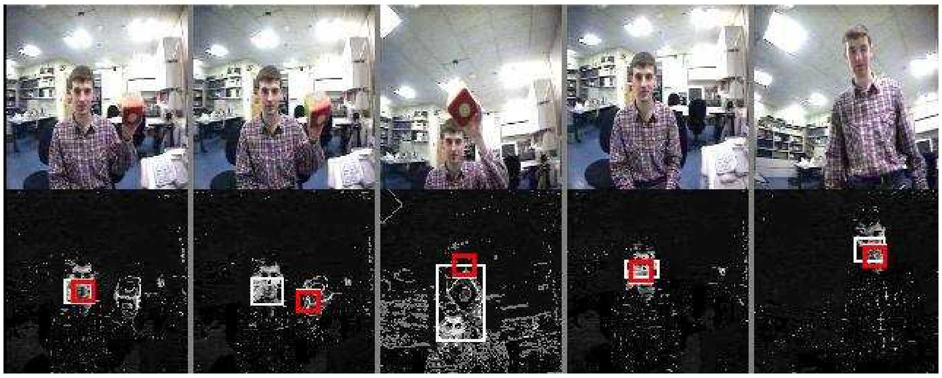


Figure 6-9: Manipulating the robot's attention. Images on the top row are from Kismet's upper wide camera. Images on the bottom summarize the contemporaneous state of the robot's attention system. Brightness in the lower image corresponds to salience; rectangles correspond to regions of interest. The thickest rectangles correspond to the robot's locus of attention. The robot's motivation here is such that stimuli associated with faces and stimuli associated with toys are equally weighted. In the first pair of images, the robot is attending to a face and engaging in mutual regard. By shaking the colored block, its salience increases enough to cause a switch in the robot's attention. The third pair shows that the head tracks the toy as it moves, giving feedback to the human as to the robot's locus of attention. The eyes are also continually tracking the target more tightly than the neck does. In the fourth pair, the robot's attention switches back to the human's face, which is tracked as it moves.

naturally and intuitively direct the robot's gaze to a desired target.

We invited three naive subjects to interact with Kismet. The subjects ranged in age from 25 to 28 years old. All used computers frequently, but were not computer scientists by training. All interactions were video recorded. The robot's attention gains were set to their default values so that there would be no strong preference for one saliency feature over another.

The subjects were asked to direct the robot's attention to each of the target stimuli.. There were seven target stimuli used in the study. Three were saturated color stimuli, three were skin toned stimuli, and the last was a pure motion stimulus. Each target stimulus was used more than once per subject We list them below:

- A highly saturated colorful block
- A bright yellow stuffed dinosaur with multi-color spines
- A bright green cylinder
- A bright pink cup (which is actually detected by the skin tone feature map)
- The person's face
- The person's hand
- A black and white plush cow (which is only salient when moving)

The video was later analyzed to determine which cues the subjects used to attract the robot's attention, which cues they used to determine when they had been successful, and the length of time required to do so. They were also interviewed at the end of the session about which cues they used, which cues they read, and about how long they thought it took to direct the robot's attention. The results are summarized in figure 6-10.

To attract the robot's attention, the most frequently used cues include bringing the target close and in front of the robot's face, shaking the object of interest, or moving it slowly across the centerline of the robot's face. Each cue increases the saliency of a stimulus by making it appear larger in the visual field, or by supplementing the color or skin tone cue with motion. Note, that there was an inherent competition between the saliency of the target and the subject's own face as both could be visible from the wide fov camera. If the subject did not try to direct the robot's attention to the target, the robot tended to look at the subject's face.

The subjects also effortlessly determined when they had successfully re-directed the robot's gaze. Interestingly, it is not sufficient for the robot to orient to the target. People look for a *change* in visual behavior, from ballistic orientation movements to smooth pursuit movements, before concluding that they had successfully re-directed the robot's attention. All subjects reported that eye movement was the most relevant cue to determine if they had successfully directed the robot's attention. They all reported that it was easy to direct the robot's attention to the desired target. They estimated the mean time to direct the robot's attention at 5 to 10 seconds. This turns out to be the case; the mean time over all trials and all targets is 5.8 seconds.

<table>
<tr><th>stimulus category</th><th>stimulus</th><th>presentations</th><th>average time (s)</th><th>commonly used cues</th><th>commonly read cues</th></tr>
<tr><td rowspan="3">color & movement</td><td>yellow dinosaur</td><td>8</td><td>8.5</td><td rowspan="8">motion across center line

shaking motion

bringing target close to robot</td><td rowspan="8">eye behavior, especially tracking

facial expression, especially raised eyebrows

body posture, especially forward lean or withdraw</td></tr>
<tr><td>multi-colored block</td><td>8</td><td>6.5</td></tr>
<tr><td>green cylinder</td><td>8</td><td>6.0</td></tr>
<tr><td>motion only</td><td>b/w cow</td><td>8</td><td>5.0</td></tr>
<tr><td rowspan="3">skin-toned & movement</td><td>pink cup</td><td>8</td><td>6.5</td></tr>
<tr><td>hand</td><td>8</td><td>5.0</td></tr>
<tr><td>face</td><td>8</td><td>3.5</td></tr>
<tr><td>Total</td><td></td><td>56</td><td>5.8</td></tr>
</table>

Figure 6-10: Summary from attention manipulation interactions. Each subject was asked to direct the robot's attention to each of the target stimuli listed in the "stimulus" column of the table. In switching between different test cases, each stimulus was used more than once. There were a total of eight presentations for each target stimuli. The time required to direct the robot's attention to the target was recorded. Each subject signaled to the experimenter when he/she had been successful in doing so. The commonly reported cues used to direct the robot's attention, and to determine when one had been successful, are also shown. The attention system is well matched to these cues.

6.11 Limitations and Extensions

There are a number of ways the current implementation can be improved and expanded upon. Some of these recommendations involve supplementing the existing framework, others involve integrating this system into a larger framework.

One interesting way this system can be improved is by adding a stereo depth map. Currently, the system estimates the proximity of the selected target. However a depth map would be very useful as a bottom-up contribution. For instance, regions corresponding to closer proximity to the robot should be more salient than those further away. A stereo map would also be very useful for scene segmentation to separate stimuli of interest from background. We are currently working towards this using the two central wide fov cameras.

Another interesting feature map to incorporate would be edge orientation. Wolfe, Triesman, among others argue in favor of edge orientation as a bottom-up feature map in humans. Currently, Kismet has no shape metrics to help it distinguish objects from each other (such as its block from its dino). Adding features to support this is an important extension to the existing implementation.

There are no auditory bottom-up contributions. A sound localization feature map would be a nice multi-modal extension (Irie 1995). Currently, Kismet assumes that the most salient person is the one who is talking to it. Often there are multiple people talking around and to the robot. It is important that the robot knows who is addressing it and when. Sound localization would be of great benefit here. Fortunately, there are stereo microphones on Kismet's ears that could be used for this purpose.

Another interesting extension would be to separate the color saliency map into individual color feature maps. Kismet can preferentially direct its attention to saturated color, but not specifically to green, blue, red, or yellow. Humans are capable of directing search based on a specific color channel. Although Kismet has access to the average r, g, b, y components of the target stimulus, it would be nice if it could keep these colors segmented (so that it can distinguish a blue circle on a green background, for instance). Computing individual color feature maps would be a step towards these extensions.

Currently there is nothing that modifies the decay rate of the habituation feature map. The habituation contribution implements a primitive attention span for the robot. It would be an interesting extension to have motivational factors, such as fatigue or arousal, influence the habituation decay rate. Caregivers continually adjust the arousal level of their infant so that the infant remains alert but not too excited (Bullowa 1979). For Kismet, it would be interesting if the human could adjust the robot's attention span by keeping it at a moderate arousal level. This could benefit the robot's learning rate by maintaining a longer attention span when people are around and the robot is engaged in interactions with high learning potential.

Kismet's visual perceptual world only consists of what is in view of the cameras. Ultimately, the robot should be able to construct an ego-centered saliency map of interaction space. In this representation, the robot could keep track of where interesting things are located, even if they are not currently in view. This will prove to be a very important representation for social referencing (Siegel 1999). If Kismet could engage in social referencing, then it could look to the human for the affective assessment and then back to the event that it queried the caregiver about. Chances are, the event in question and the human's face will not be in view at the same time. Hence, a representation of where interesting things are, even when out of view, is an important resource.

6.12 Summary

There are many interesting ways in which Kismet's attention system can be improved and extended. This should not overshadow the fact that the existing attention system is an important contribution to autonomous robotics research.

Other researchers have developed bottom-up attention systems (Itti et al. 1998), (Wolfe 1994). Many of these systems work in isolation and are not embedded in a behaving robot. Kismet's attention system goes beyond raw perceptual saliency to incorporate top-down task-driven influences that vary dynamically over time with its goals. By doing so, the attention system is tuned to benefit the task the robot is currently engaged in.

There are far too many things that the robot could be responding to at any time. The attention system gives the robot a locus of interest that it can organize its behavior around. This contributes to perceptual stability since the robot is not inclined to flit its eyes around randomly from place to place, changing its perceptual input at a pace too rapid for behavior to keep up. This in turn contributes to behavioral stability since the robot has a target that it can direct its behavior towards and respond to. Each target (people, toys) has a physical persistence that is well matched to the robot's behavioral time scale. Of course, the robot can respond to different targets sequentially in time, but this occurs at a slow enough time scale that the behaviors have time to self organize and stabilize into a coherent goal-directed pattern before a switch to a new behavior is made.

There is no prior art in incorporating a task-dependent attentional system into a robot. Some side step the issue by incorporating an implicit attention mechanism into the perceptual conditions that release behaviors (Blumberg 1994), (Velasquez 1998). Others do so by building systems that are hardwired to perceive one type of stimulus tailored to the specific task (Schall 1997), (Mataric, Williamson, Demiris & Mohan 1998), or use very simple sensors (Hayes & Demiris 1994), (Billard & Dautenhahn 1997). However, the complexity of Kismet's visual environment, the richness of its perceptual capabilities, and its time-varying goals required an explicit implementation.

The social dimension of Kismet's world adds additional constraints that prior robotic systems have not had to deal with. As argued earlier, the robot's attention system needed to be tuned to the attention system of humans. In this way, both robot and humans are more likely to find the same sorts of things interesting or attention grabbing. As a result, people can very naturally and quickly direct the robot's attention. The attention system coupled with gaze direction provides people with a powerful and intuitive social cue. The readability and interpretation of the robot's behavior is greatly enhanced since the person has an accurate measure of what the robot is responding to.

The ability for humans to easily influence the robot's attention and to read its cues has a tremendous benefit to various forms of social learning and is an important form of scaffolding. When learning a task, it is difficult for a robotic system to learn what perceptual aspects matter. This only gets worse as robots are expected to perform more complex tasks in more complex environments. However, this challenging learning issue can be addressed in an interesting way if the robot learns the task with a human instructor who can explicitly direct the robot's attention to the salient aspects, and can determine from the robot's social cues whether or not the robot is attending to the relevant features. This doesn't solve the problem, but it could facilitate a solution in a new and interesting way that is natural and intuitive for people.

In the big picture, low level feature extraction and visual attention are components of a larger visual system. We present how the attention system is integrated with other visual behaviors in chapter 13.

Chapter 7

Recognition of Affective Intent in Robot-Directed Speech

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 chapter focuses on the problem of recognizing affective intent in robot-directed speech. The work presented in this chapter was carried out in collaboration with Lijin Aryananda, and is reported in (Breazeal & Aryananda 2000).

7.1 Emotion Recognition in Speech

When extracting the affective message of a speech signal, there are two related yet distinct questions one can ask. The first is: "*What is the emotion being expressed?*". In this case, the answer describes an emotional quality – such as sounding angry, or frightened, or disgusted, etc.. Each emotional state causes changes in the autonomic nervous system. This, in turn, influences heart rate, blood pressure, respiratory rate, sub-glottal pressure, salivation, and so forth. These physiological changes produce global adjustments to the acoustic correlates of speech – influencing pitch, energy, timing, and articulation. 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, Polzin & A. 1996), (Nakatsu, Nicholson & Tosa 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 are far from perfect in recognizing emotion from speech alone. Some have attempted to use multi-modal cues (facial expression with expressive speech) to improve recognition performance (Chen & Huang 1998).

However, for the purposes of training a robot, the raw emotional content of the speaker's voice is only part of the message. This leads us to the second, related question: "*What is the affective intent of the message?*". Answers to this question may be that the speaker was praising, prohibiting, alerting, etc. the recipient of the message. A few researchers have developed recognition systems that can recognize speaker approval versus speaker disapproval from child-directed speech (Roy & Pentland 1996), or recognize praise, prohibition, and attentional bids from infant-directed speech (Slaney & McRoberts 1998). For the remainder of this chapter, we discuss how this idea could be extended to serve as a useful training signal for Kismet. Note that Kismet does not learn from humans yet, but this is an important capability that could support socially situated learning.

7.2 Affective Intent in Speech

Developmental psycholinguists have extensively studied how affective intent is communicated to preverbal infants (Fernald 1989), (Grieser & Kuhl 1988). Infant-directed speech is typically quite exaggerated in pitch and intensity(Snow 1972). From the results of a series of cross-cultural studies, Fernald suggests that much of this information is communicated through the "melody" of infant-directed speech. In particular, there is evidence for at least four distinctive prosodic contours, each of which communicates a different affective meaning to the infant (approval, prohibition, comfort, and attention). Maternal exaggerations in infant-directed speech seem to be particularly well matched to the innate affective responses of human infants (Mumme, Fernald & Herrera 1996).

Inspired by this work, we have implemented a recognizer to distinguish the four affective intents for praise, prohibition, comfort, attentional bids. Of course, not everything a human says to Kismet will have an affective meaning, so we also distinguish neutral robot-directed speech. These affective intents are well matched to teaching a robot since praise (positive reinforcement), prohibition (negative reinforcement), and directing attention, could be intuitively used by a human instructor to facilitate the robot's learning process. Within the AI community, a few researchers have already demonstrated how affective information can be used to bias learning at both goal-directed and affective levels for robots (Velasquez 1998) and synthetic characters (Yoon, Blumberg & Schneider 2000).

For Kismet, output of the vocal classifier is interfaced with the emotion subsystem (see chapter 8) where the information is appraised at an affective level and then used to directly modulate the robot's own affective state[1]. In this way, the affective meaning of the utterance is communicated to the robot through a mechanism similar to the one Fernald suggests. As with human infants, socially manipulating the robot's affective system is a powerful way to modulate the robot's behavior and to elicit an appropriate response.

In the rest of this chapter 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 the most useful for recognizing affective intent, as well as how it may be used by human infants to organize their behavior. We then outline a series of design issues for integrating this competence into Kismet. We present a detailed description of our approach and how we have integrated it into Kismet's affective circuitry. The performance of the

[1]Typically, "affect" refers to positive and negative qualities. For our work with Kismet, we also include arousal levels and the robot's willingness to approach or withdraw, when talking about Kismet's affective state

system is evaluated with naive subjects as well as the robot's caregivers. We discuss our results, suggest future work, and summarize our findings.

7.3 Affect and Meaning in Infant-directed Speech

Developmental psycholinguists have studied the acoustic form of adult speech directed to preverbal infants and have discovered an intriguing relation between voice pitch and affective intent. (Fernald 1989), (Papousek, Papousek & Bornstein 1985), (Grieser & Kuhl 1988). When mothers speak to their preverbal infant, their prosodic patterns (the contour of the fundamental frequency and modulations in intensity) are exaggerated in characteristic ways. Even with newborns, mothers use higher mean pitch, wider pitch range, longer pauses, shorter phrases, and more prosodic repetition when addressing infants than when speaking to an adult. This exaggerated manner of speaking (i.e., motherese) serves to engage infant's attention and prolong interaction.

Attentional Bids, Approval, Prohibition, and Comfort

Maternal intonation is finely tuned to the behavioral and affective state of the infant. Further, mothers intuitively use selective prosodic contours to express different affective 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 & Kuhl 1988), (Fernald 1993), (McRoberts, Fernald & Moses 2000). 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 to encourage a response (Ferrier 1985), and to use bell shaped contours to maintain attention once it has been established (Stern, Spieker & MacKain 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. Figure 7-1 illustrates these prototypical contours.

Exaggerated Prosodic Cues Convey Meaning

It is interesting that even though preverbal infants do not understand the linguistic content of the message, they appear to understand the affective content and respond appropriately. This may comprise some of infants earliest communicated meanings of maternal vocalizations. The same patterns can be found when communicating these same intents to adults, but in a significantly less exaggerated manner (Fernald 1989). By eliminating the linguistic content of infant-directed and adult-directed utterances for the categories described above (only preserving the "melody" of the message) Fernald found that adult listeners were more accurate in recognizing these affective categories in infant-directed speech than in adult-directed speech. This suggests that the relation of prosodic form to communicative function is made uniquely salient

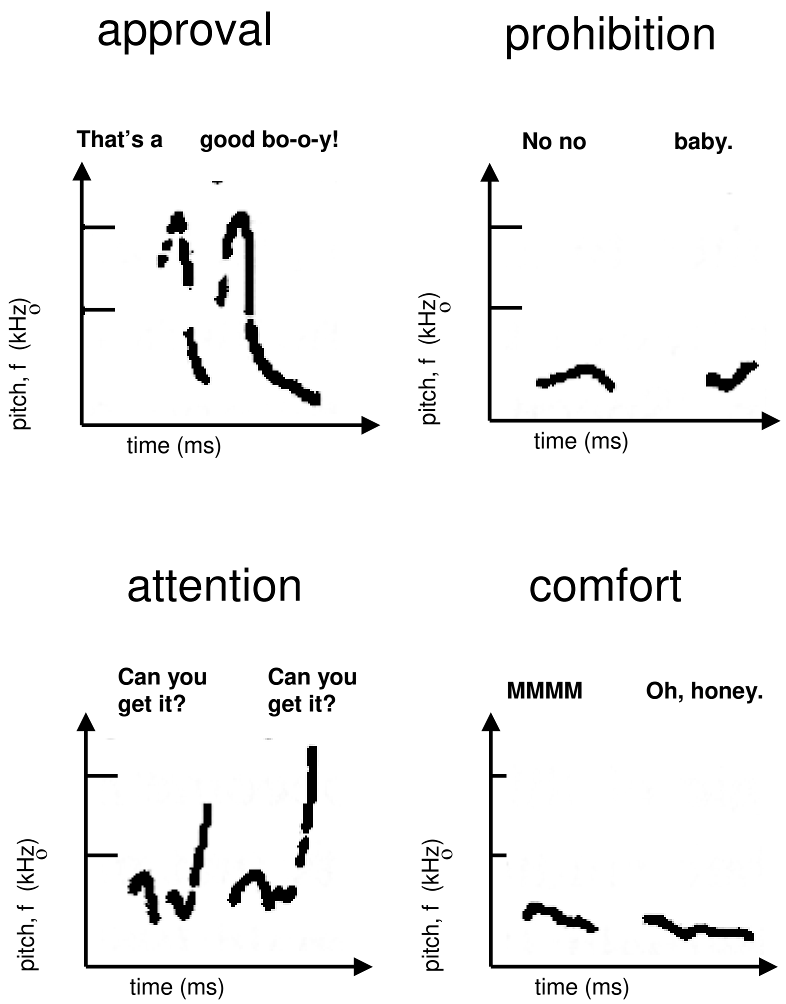


Figure 7-1: Fernald's prototypical contours for approval, prohibition, attention, and soothing. These affective contours have been found to exist in several cultures. It is argued that they are well matched to saliency measures hardwired into the infant's auditory processing system. Caregivers use these contours intuitively to modulate the infant's arousal level.

in the melodies of mother's speech, and that these intonation contours provide the listener with reliable acoustic cues to the speaker's intent.

Matching Acoustic Structure to Communicative Function

Fernald has used the results of such studies to argue for the adaptive significance of prosody in child language acquisition, as well as in the development and strength of the parent-offspring relationship. She 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. Auditory signals with high frequency and rising pitch have been found to be more alerting to human listeners than those signals lower in frequency and falling pitch (Ferrier 1985). Hence, the acoustic design of attentional bids would appear to be appropriate to the goal of eliciting attention. Similarly, low mean pitch, narrow pitch range, and low intensity (all characteristics of comfort vocalizations) have been found to be correlated with low arousal (Papousek et al. 1985). Given that the mother's goal in soothing her infant is to decrease arousal, comfort vocalizations are well suited to this function. Speech having a sharp, loud, staccato contour, low pitch mean, and narrow pitch range tend to startle the infant (tending to halt action or even induce withdraw) and are particularly effective as warning signals (Fernald 1989). Infants show a listening preference for exaggerated pitch contours. They respond with more positive affect to wide range pitch contours than to narrow range pitch contours. Hence the exaggerated bell-shaped prosody contour for approval is effective for sustaining the infant's attention and engagement (Stern et al. 1982).

Development of Meaning

By anchoring the message in the melody, there may be a facilitative effect on "pulling" the word out of the acoustic stream and causing it to be associated with an object or event. This development is argued to occur in four stages (Fernald 1989). In the first stage, certain acoustic features of speech have intrinsic perceptual salience for the infant. Certain maternal vocalizations function as unconditioned stimuli in alerting, soothing, pleasing, and alarming the infant. In stage two, the melodies of maternal speech become increasingly more effective in directing the infant's attention, and in modulating the infant's arousal and affect. The communication of intention and emotion takes place in the third stage. Vocal and facial expressions give the infant initial access to the feelings and intentions of others. Stereotyped prosodic contours occurring in specific affective contexts come to function as the first regular sound-meaning correspondences for the infant. In the fourth stage, prosodic marking of focused words helps the infant to identify linguistic units within the stream of speech. Words begin to emerge from the melody.

7.4 Design Issues

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.

Robot Aesthetics

As discussed above, the perceptual task of recognizing affective 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. It will be a while 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 (chapter 4), how it moves (chapters 13, 10), and how it expresses itself (chapters 12, 11). 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 seem quite appropriate. We have found this typically to be the case for both men and women when interacting with Kismet.

Real-time Performance

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

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 interaction otherwise.

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). For this reason, we have designed the recognizer to only categorize sufficiently exaggerated prosody such as as praise, prohibition, attention, and soothing (i.e., the caregiver has to say it as if he/she *really* means it). Vocalizations with insufficient exaggeration are classified as neutral.

Acceptable vs Unacceptable Misclassification

Given that humans are not perfect at recognizing the affective content in speech, the robot is sure to 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.

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 system, the robot's ability to express itself through facial expression, voice quality, and body posture conveys the robot's affective interpretation of the message. This allows 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 (perhaps what the robot just did was *very* good, and the robot should be *really* happy about it).

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.

7.5 The Algorithm

As shown in figure 7-2, the affective speech recognizer receives robot-directed speech as input. The speech signal is analyzed by the low level speech processing system, producing time-stamped pitch (Hz), percent periodicity (a measure of how likely a frame is a voiced segment), energy (dB), and phoneme values [2] all in real-time. The next module performs filtering and pre-processing to reduce the amount of noise in the data. The pitch value of a frame is simply set to zero if the corresponding percent periodicity indicates that the frame is more likely to correspond to unvoiced speech. The resulting pitch and energy data are then passed through the feature extractor, which calculates a set of selected features (F_1 to F_n). Finally, based on the trained model, the classifier determines whether the computed features are derived from an approval, an attentional bid, a prohibition, soothing speech, or a neutral utterance.

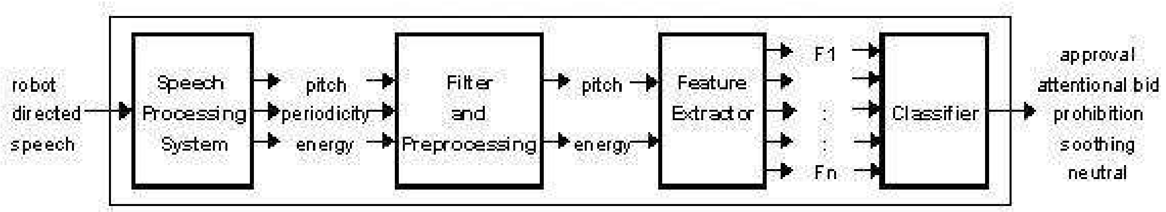


Figure 7-2: The spoken affective intent recognizer.

7.5.1 Training the System

We made recordings of two female adults who frequently interact with Kismet as caregivers. The speakers were asked to express all five affective intents (approval, attentional bid, prohibition, comfort, and, neutral) during the interaction. Recordings were made using a wireless microphone, and the output signal was sent to the low-level speech processing system running on Linux. For each utterance, this phase produced a 16-bit single channel, 8 kHz signal (in a `.wav` format) as well as its corresponding real-time pitch, percent periodicity, energy, and phoneme values. All recordings were performed in Kismet's usual environment to minimize variability of environment-specific noise. We then eliminated samples containing extremely loud noises (door slams, etc.) and labeled the remaining data set according to the speakers' affective intents during the interaction. There were a total of 726 utterances in the final data set – approximately 145 utterances per class.

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

Data Preprocessing

The pitch value of a frame was set to zero if the corresponding percent periodicity was lower than a threshold value. This indicates that the frame is more likely to correspond to unvoiced speech. Even after this procedure, observation of the resulting pitch contours still indicated the presence of substantial noise. Specifically, a significant number of errors were discovered in the high pitch value region (above 500 Hz). Therefore, additional preprocessing was performed on all pitch data. For each pitch contour, a histogram of ten regions was constructed. Using the heuristic that the pitch contour was relatively smooth, 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 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 a Gaussian mixture model, updated with the EM algorithm and a Kurtosis-based approach for dynamically deciding the appropriate number of kernels (Vlassis & Likas 1999). Due to the limited set of training data, we performed cross-validation in all classification processes. Specifically, we held out a subset of data and trained a classifier using the remaining data. We then tested the classifier's performance on the held out test set. This process was repeated 100 times per classifier. We calculated the mean and variance of the percentage of correctly classified test data to estimate the classifier's performance.

7.5.2 The Single Stage Classifier: First Pass

As shown in figure 7-3, the preprocessed pitch contour of 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 figure 7-4). All pitch features were measured using only non-zero pitch values. We hypothesized that although none of these features directly encoded any temporal information about the pitch contour, they would still be useful in distinguishing some classes. For example, approval and attentional bids were expected to generate high pitch variance while prohibition should have a lower pitch mean and a high energy level.

Using this feature set, we applied a sequential forward feature selection process to construct a single stage classifier. The classification performance of each possible pair of features was measured. The sixty-six feature pairs were then sorted based on their respective performance, from highest to lowest. Successively, a feature pair from the sorted list was added into the selected feature set to determine the best n features for this classifier.

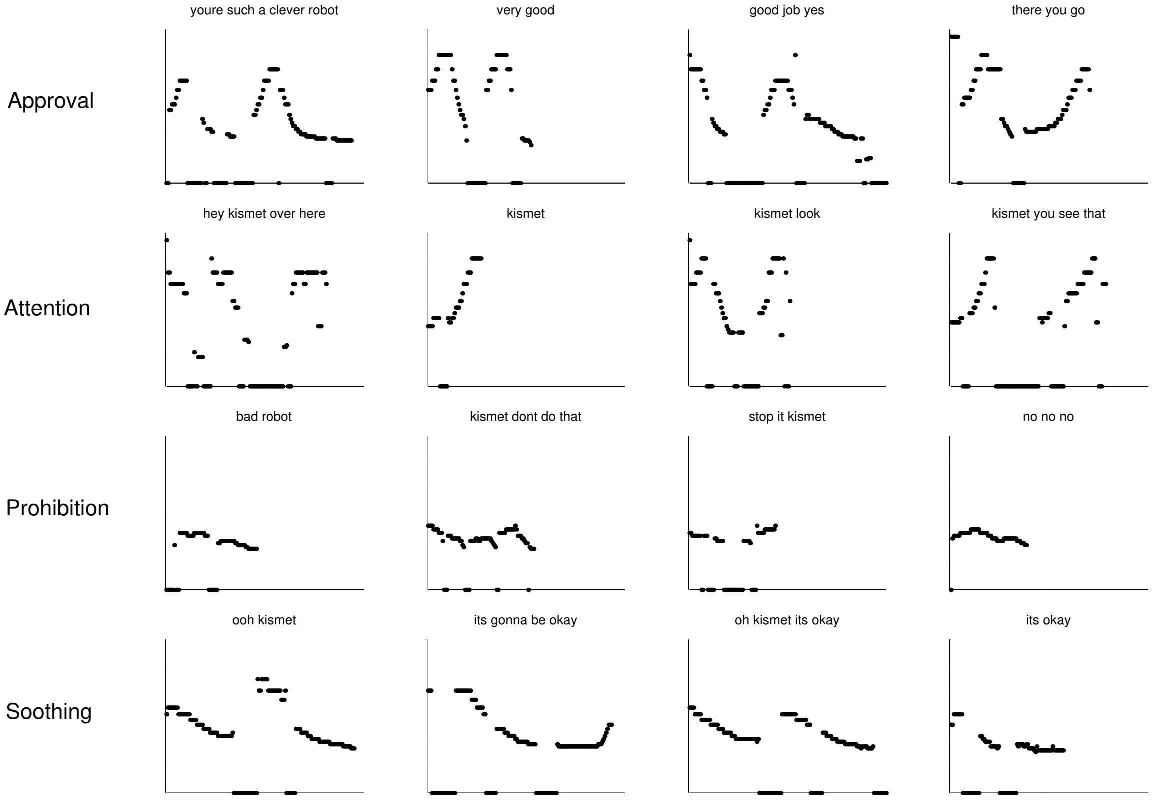


Figure 7-3: Fernald's prototypical prosodic contours found in the preprocessed data set. Notice the similarity to those shown in figure 7-1.

	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

Figure 7-4: Features extracted in the single stage classifier. These features are measured over the non-zero values over the entire utterance. Feature F_6 measures the steepness of the slope of the pitch contour.

	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
F1	65.47	58.06	61.22	60.32	56.77	59.21	63.60	72.09	70.96	70.03	60.51
F2		40.23	59.16	48.82	51.31	55.23	52.86	68.79	62.11	63.28	56.27
F3			47.66	46.43	46.69	46.72	51.45	64.04	57.43	56.94	51.61
F4				32.68	48.54	50.75	49.81	63.42	59.93	55.23	53.48
F5					47.36	51.86	49.42	63.49	59.73	58.74	52.07
F6						37.33	44.66	54.39	49.41	49.64	47.48
F7							44.19	59.56	53.65	55.41	49.57
F8								57.28	50.31	45.52	49.94
F9									58.61	59.88	62.47
F10										59.47	59.08
F11											59.35

Figure 7-5: Feature pair performance (%). The feature pair F_1, F_9 give the best classification performance, which corresponds to pitch mean and energy variance, respectively.

Single Stage Classifier Results

Figure 7-5 illustrates each feature pair's classification performance. The combination of F_1 (pitch mean) and F_9 (energy variance) produces the best performance of all the feature pairs (72.1%) while combining F_4 (pitch range) and F_5 (delta pitch mean) results in the worst performance (32.7%). These feature pairs were sorted based on performance. Figure 7-6 shows the classification results as each of the top pairs in the sorted list are added sequentially into the feature set. Classification performance increases as more features are added, reaching a maximum (78.8%) with five features in the set, and then levels off above 60% with six or more features. Table 7-8 provides a closer look at the classifier constructed using these best eight feature pairs. It is clear that all seven classifiers perform best in recognizing prohibition, but not as well in classifying the other classes. Figure 7-9 plots the feature space of the first classifier (F_1 and F_9), which explains why a high number of approval, attention, soothing, and neutral samples were incorrectly classified. There are three clusters in the feature space. The prohibition class forms the first cluster, which is well separated from the rest. Approval and attention samples form the second cluster, with some overlap between the two classes. Soothing and neutral class form the last cluster, also with some overlap.

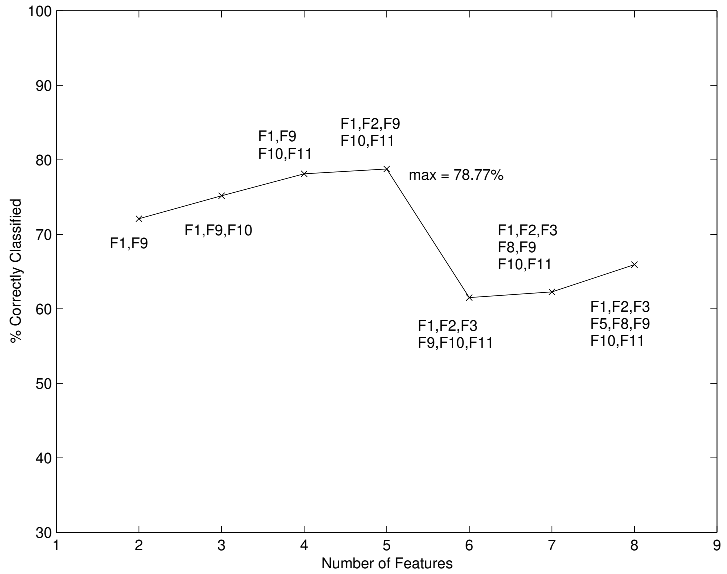


Figure 7-6: Classification performance using sequential forward selection. The best performance is given by the combination of features F_1, F_2, F_9, F_{10}, and F_{11}. These correspond to pitch mean, pitch variance, energy variance, energy range, and maximum energy.

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

Figure 7-7: Classification results of the ten best feature pairs for the single stage classifier. We use these findings to design the first stage of our multi-stage classifier.

7.5.3 The Multi-Stage Classifier: Second Pass

Results obtained for the single stage classifier revealed that the global pitch and energy features were useful for separating some classes from the rest, but not sufficient for constructing a high performance 5-way classifier. In our second attempt, instead of having one single stage classifier that simultaneously classifies all five classes, we implemented several mini-classifiers that classifies the data in stages. In the first stage, the classifier uses global pitch and energy features to separate some classes (high arousal versus low arousal) as well as possible. The remaining clustered classes are then passed to subsequent classification stages. Obviously, we had to consider new features in order to build these additional classifiers. Utilizing prior information, we included a new set of features that encoded the *shape* of the pitch contour. We found these features to be useful in separating the *difficult* classes.

Multi-Stage Classifier Results

Figure 7-7 illustrates the classification results of the best ten feature pairs obtained in the single stage classifier attempt, including the number of incorrectly classified samples in each class. 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 figure 7-9) 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

Feature pair		Feature set	Performa nce mean (%)	Performa nce 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

Figure 7-8: A closer look at classification results for the single stage classifier. The performance (the percent correctly classified) is shown for the best pair-wise set having eight features. The pair-wise performance was ranked for the best ten pairs. As each successive feature was added, we see performance peaks with five features (78.8%), but then drops off.

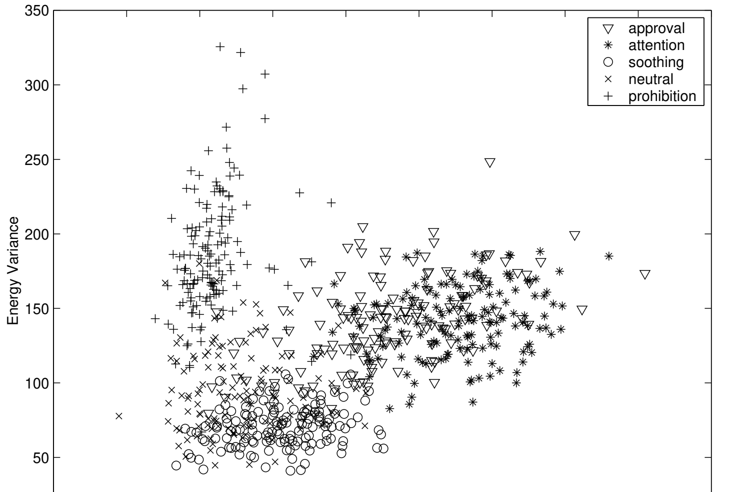


Figure 7-9: Feature space of all five classes with respect to energy variance, F_9, and pitch mean, F_1. We see three distinguishable clusters for prohibition, soothing and neutral, and approval and attention.

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 with respect to 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 those 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 energy-related features should be used to classify soothing and neutral speech (having low intensity) from the other higher intensity classes (see figure 7-10). In the second stage, we execute another classifier to decide if a low intensity utterance corresponds to either soothing or neutral speech. If the utterance exhibits high intensity, then we use the $F_1 - F_9$ pair to distinguish among prohibition, the approval-attention cluster, or high intensity neutral. An additional stage would be required to classify approval versus attention if the utterance happened to fall within the approval-attention cluster.

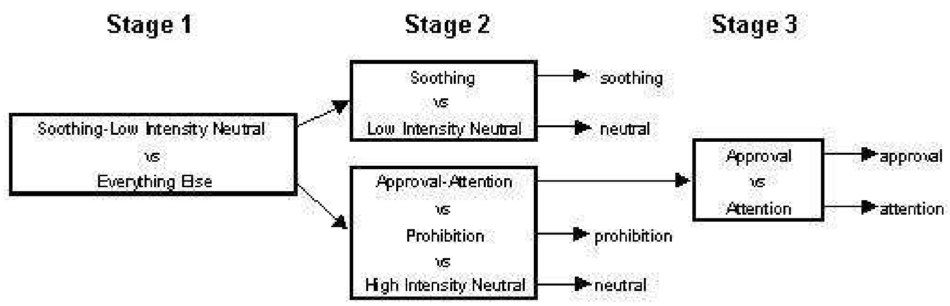


Figure 7-10: The classification stages of the multi-stage classifier.

Stage 1: Soothing-Low Intensity Neutral vs Everything Else

The first two columns in table 7-11 show the classification performance of the top four feature pairs (sorted according to 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 (energy variance, maximum energy, and energy range), with an average performance of 93.6%. The resulting feature space is shown in figure 7-12.

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

Figure 7-11: Classification results in stage 1.

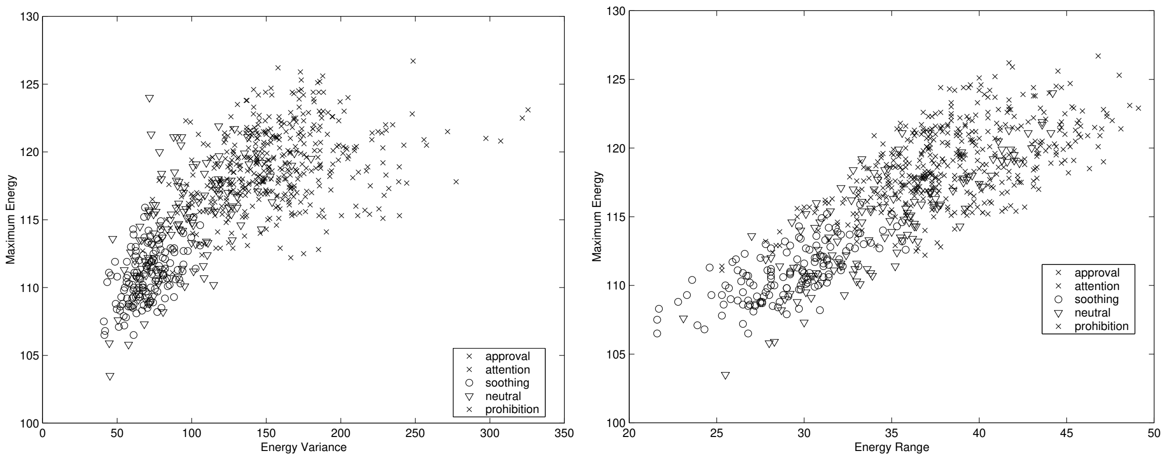


Figure 7-12: Feature space: soothing vs neutral vs rest.

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 suggest looking for a smooth pitch contour exhibiting a frequency down-sweep. Visual observations of the neutral samples in the data set indicated that neutral speech generated flatter and choppier pitch contours as well as less modulated energy contours. Based on these postulations, 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 the pitch contour indicated whether or not the contour contained a down-sweep segment. It was calculated by performing a linear fit on the contour segment starting at the maximum peak. This classifier's average performance is 80.3%.

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 90.0%. Based on Fernald's prototypical prosodic patterns and the feature space shown in figure 7-13, 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 92.1%.

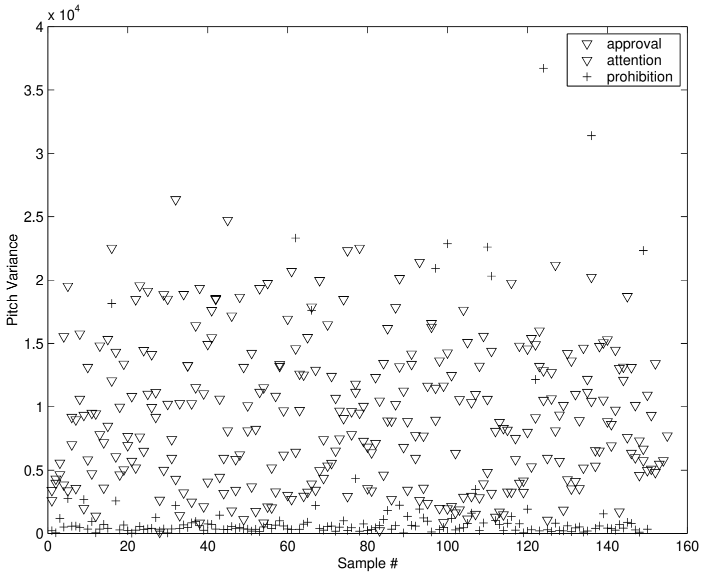


Figure 7-13: Feature space: approval-attention vs prohibition.

Stage 3: Approval vs Attention

Since the approval class and attention class span across the same region in the global pitch vs. 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 fit 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 longest pitch segment that contributed 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 figure 7-14. 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.

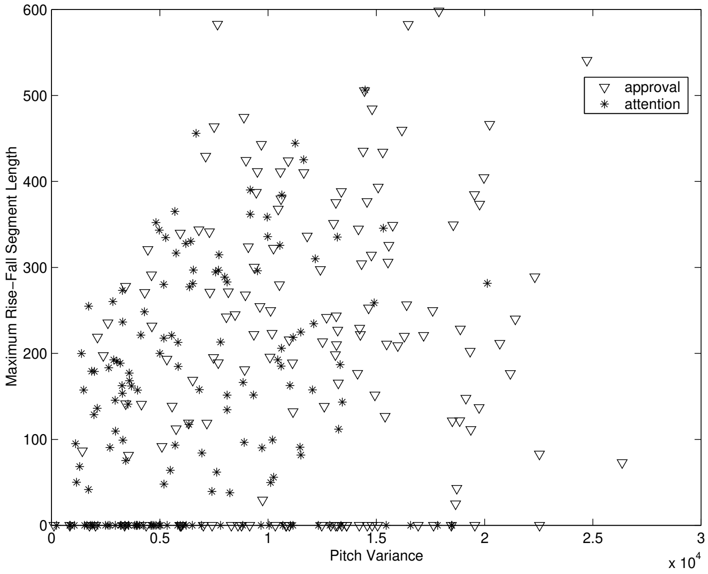


Figure 7-14: Feature space: approval versus attentional bid.

7.5.4 Overall Performance

The final classifier was evaluated using a new test set generated by the same female speakers, containing 371 utterances. Because each mini-classifier was trained using different portions of the original database (for the single stage classifier), we had to gather a new data set to ensure that we would not be testing any mini-classifier stage on data that was used to train it. Figure 7-15 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 multi-stage classifier performs significantly better in classifying the *difficult* classes, i.e., approval versus attention and soothing versus neutral. This verifies that the features encoding the shape of the pitch contours (derived from prior knowledge provided by Fernald's prototypical prosodic patterns) were very useful.

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

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

Figure 7-15: Overall classification performance.

7.6 Integration with the Emotion System

The output of the recognizer is integrated into the rest of Kismet's synthetic nervous system as shown in figure 7-16. Please refer to chapter 8 for a detailed description of the design of the emotion system. In this chapter, we briefly present only those aspects of the emotion system as they are related to integrating recognition of vocal affective intent into Kismet.

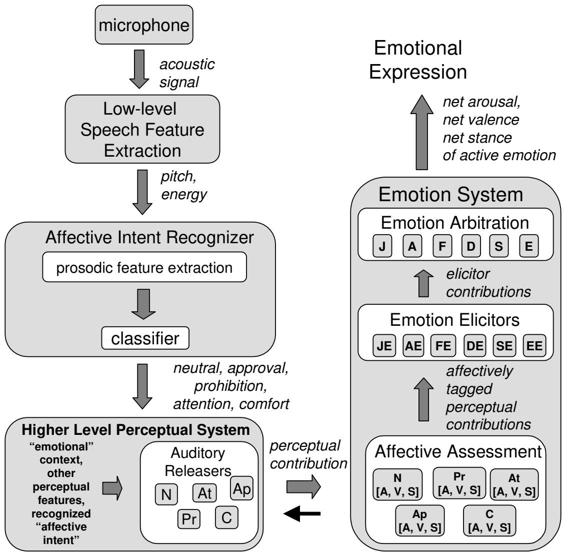


Figure 7-16: System architecture for integrating vocal classifier input to Kismet's emotion system. See text.

The entry point for the classifier's result is at the auditory perceptual system. Here, it is fed into an associated releaser process. In general, there are many different kinds of releasers defined for Kismet, each combining different contributions from a variety of perceptual and motivational systems. For our purposes here, we only discuss those releasers related to the input from the vocal classifier. The output of each vocal affect releaser represents its perceptual contribution to the rest of the SNS. Each releaser combines the incoming recognizer signal with contextual information (such as the current "emotional" state) and computes its level of activation according to the magnitude of its inputs. If its activation passes above threshold, it passes its output on to the emotion system. The emotion system is presented in chapter 8.

Within the emotion system, the output of each releaser must first pass through the affective assessment subsystem in order to influence emotional behavior. Within this assessment subsystem, each releaser is evaluated in affective terms by an associated *somatic marker* (SM) process. This mechanism is inspired by the *Somatic Marker Hypothesis* of Damasio (1994) where incoming perceptual information is "tagged" with affective information. Table 7-17 summarizes how each vocal affect releaser is somatically tagged.

There are three classes of tags that the affective assessment phase uses to affectively characterize its perceptual, motivational, and behavioral input. Each tag has an associated intensity that scales its contribution to the overall affective state. The *arousal* tag, A, specifies how arousing this percept is to the emotional system. Positive values correspond to a high arousal stimulus whereas negative values correspond to a low arousal stimulus. The *valence* tag, V, specifies how good or bad this percept is to the emotional system. Positive values correspond to a pleasant stimulus whereas negative values correspond to an unpleasant stimulus. The *stance* tag, S, specifies how approachable the percept is. Positive values correspond to advance whereas negative values correspond to retreat. Because there are potentially many different kinds of factors that modulate the robot's affective state (e.g., behaviors, motivations, perceptions), this tagging process converts the myriad of factors into a common currency that can be combined to determine the *net* affective state.

	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

Figure 7-17: Table mapping $[A, V, S]$ to classified affective intents. Praise biases the robot to be "happy", prohibition biases it to be "sad", comfort evokes a "content, relaxed" state, and attention is arousing. See text.

For Kismet, the $[A, V, S]$ trio is the currency the emotion system uses to determine which emotional response should be active. This occurs in two phases: First, all somatically marked inputs are passed to the *emotion elicitor* stage. Each emotion process has an elicitor associated with it that filters each of the incoming $[A, V, S]$ contributions. Only those contributions that satisfy the $[A, V, S]$ criteria for that emotion process are allowed to contribute to its activation. This filtering is done independently for each class of affective tag. For instance, a valence contribution with a large negative value will not only contribute to the `sorrow` emotion process, but to the `fear`, `anger`, and `distress` processes as well. Given all these factors, each elicitor computes its net $[A, V, S]$ contribution and activation level, and passes them to the associated emotion process within the *emotion arbitration* subsystem.

In the second stage, the emotion processes within the emotion arbitration subsystem compete for activation based on their activation level. There is an emotion process for each of Ekman's six basic emotions (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 ensues. 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 (chapter 11), tone of voice (chapter 12), or posture (chapter 11). The robot's facial expression also sets up the human's expectation of what behavior will soon follow. As a result, the human observing the robot can see its behavior, in addition to having an understanding of why the robot is behaving in that manner. As we have argued previously, readability is an important issue for social interaction with humans.

7.7 Use of Behavioral Context to improve interpretation

Most affective speech recognizers are not integrated into robots equipped with emotion systems that are also 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).

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 `joy` 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 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 emotion processes and to excite the negatively valenced emotion 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.

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 is 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 (slightly arousing, slightly positive).

7.8 Experiments

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 naive 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?

7.8.1 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). One of the subjects was a caregiver of Kismet, who spoke to the robot in Indonesian. Subjects were instructed to express each affective 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.

7.8.2 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 figure 7-18), 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 linear fit on these contours generates a flat slope, resulting in a neutral classification. A few soothing utterances were confused for neutral despite having the down-sweep frequency characteristic because they contained too many words and coarse pitch contours. Attentional bids generated the worst classification performance for the strong utterances (it performed better than most for the weak utterances). A careful observation of the classification errors revealed that many of the misclassified attentional bids contained the word "kismet" spoken with a bell-shaped pitch contour. The classifier recognized this as the characteristic rise-fall pitch segment found in approvals. We also found that many other common words used in attentional bids, such as "hello" (spoken as "hel-lo-o"), also generated a bell-shaped pitch contour. 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.

First, a high number of utterances are perceived to carry a *strong* affective message, which implies the use of exaggerated prosody during the interaction session (as we hoped for). The remaining question is whether or not the classifier will generalize to the naive speakers' exaggerated prosodic patterns. Except for the two special cases discussed above, the experimental results indicate that the classifier performs very well in recognizing the naive speakers' prosodic contours even though it was trained only on utterances from the primary caregivers. Moreover, the same failure modes occur in the naive speaker test set. No strongly valenced intents were misclassified as those with opposite valence. It is very encouraging to discover that the classifier not only generalizes to perform well on naive speakers (using either English or other languages), but it also makes very few unacceptable misclassifications.

7.8.3 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

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 speak ers	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

Figure 7-18: Classification performance on naive speakers. The subjects spoke to the robot directly and received expressive feedback. Their utterances were recorded during the interaction. An objective scorer ranked each utterance as strong, medium, or weak. We expect to see (and do) the best performance for the caregivers and for strong utterances from naive subjects.

to be “cute”, which encourages this manner of address. Naive subjects appear to enjoy interacting with Kismet and are often impressed at how life-like it is. 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 an exaggerated prosody characteristic of infant-directed speech. It is quite different from the manner in which they spoke to 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 less inhibited, whereas adult males are often more reserved. For those who are relatively uninhibited, their styles for conveying affective 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.

The subjects in the study made ready use of Kismet’s expressive feedback to assess when the robot “understood” them. The robot’s expressive repertoire is quite rich, including both facial expressions 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”.

Kismet’s expression through face and body posture becomes more intense as the activation level of the corresponding emotion process increases. For instance, small smiles versus large grins were often used to discern how happy the robot appeared. Small ear perks versus widened eyes with elevated ears and craning the neck forward were often used to discern growing levels of interest and attention. The subjects could discern these intensity differences, and several modulated their own speech to influence them.

During the 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 were enough to arouse a strong sense of empathy. The subject would often immediately stop and look to the experimenter with an anguished expression on her face, claiming to feel “terrible” or “guilty”. In this emotional feedback cycle, the robot’s own affective response to the subject’s vocalizations evoked a strong and similar emotional response in the subject as well.

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 in an even more foreboding manner. 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. This technique was employed to modulate the degree to which the strength of the message was communicated to the robot.

7.9 Limitations and Extensions

The ability of naive subjects to interact with Kismet in this affective and dynamic manner suggests that its response rate is acceptable. 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, and waiting long enough during that pause to make sure that speech has completed. There is another delay of approximately one second associated with interpreting the classifier in affective terms and feeding it through to 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 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?". A performance of seventy to eighty percent 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, in our approach we have taken care to minimize the number of "bad" misclassifications. We also exploit the social context to reduce misclassifications further (such as soothing versus neutral). Finally, we 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 naive 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.

7.10 Summary

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 affective 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 naive 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.

Chapter 8

The Motivation System

In general, animals are in constant battle with many different sources of danger. They must make sure that they get enough to eat, that they do not become dehydrated, that they do not overheat or freeze, that they do not fall victim to a predator, and so forth. The animal's behavior is beautifully adapted to survive and reproduce in this hostile environment. Early ethologists used the term *motivation* to broadly refer to the apparent self-direction of an animal's attention and behavior (Tinbergen 1951), (Lorenz 1973).

In more evolutionary advanced species, the following features appear to become more prominent: the ability to process more complex stimulus patterns in the environment, the simultaneous existence of a multitude of motivational tendencies, a highly flexible behavioral repertoire, and social interaction as the basis of social organization. Within an animal of sufficient complexity, there are multiple motivating factors that contribute to the its observed behavior. Modern ethologists, neuroscientists, and comparative psychologists continue to discover the underlying physiological mechanisms, such as internal clocks, hormones, and internal sense organs, that serve to regulate the animal's interaction with the environment and promote its survival. For the purposes of this chapter, we focus on two classes of motivation systems: homeostatic regulation and emotion.

8.0.1 Homeostatic Regulation

To survive, animals must maintain certain critical parameters within a bounded range. For instance, an animal must regulate its temperature, energy level, amount of fluids, etc.. Maintaining each critical parameter requires that the animal come into contact with the corresponding satiatory stimulus (shelter, food, water, etc.) at the right time. The process by which these critical parameters are maintained is generally referred to as *homeostatic regulation* (Carver & Scheier 1998). In a simplified view, each satiatory stimulus can be thought of as an innately specified need. In broad terms, there is a desired fixed point of operation for each parameter, and an allowable bounds of operation around that point. As the critical parameter moves away from the desired point of operation, the animal becomes more strongly motivated to behave in ways that will restore that parameter. The physiological mechanisms that serve to regulate these needs, driving the animal into contact with the needed stimulus at the appropriate time, are quite complex and distinct (Gould 1982), (McFarland & Bosser

1993).

8.0.2 Emotion

Emotions are another important motivation system for complex organisms. They seem to be centrally involved in determining the behavioral reaction to environmental (often social) and internal events of major significance for the needs and goals of a creature (Plutchik 1991), (Izard 1977). For instance, Frijda (1994*a*) suggests that positive emotions are elicited by events that satisfy some motive, enhance one's power of survival, or demonstrate the successful exercise of one's capabilities. Positive emotions often signal that activity toward the goal can terminate, or that resources can be freed for other exploits. In contrast, many negative emotions result from painful sensations or threatening situations. Negative emotions motivate actions to set things right or prevent unpleasant things from actually occurring.

Several theorists argue that a few select emotions are *basic* or *primary* – they are endowed by evolution because of their proven ability to facilitate adaptive responses to the vast array of demands and opportunities a creature faces in its daily life (Ekman 1992), (Izard 1993). The emotions of anger, disgust, fear, joy, sorrow and surprise are often supported as being basic from evolutionary, developmental, and cross-cultural studies (Ekman & Oster 1982). Each basic emotion is posited to serve a particular function (often biological or social), arising in particular contexts, to prepare and motivate a creature to respond in adaptive ways. They serve as important reinforcers for learning new behavior. In addition, emotions are refined and new emotions are acquired throughout emotional development. Social experience is believed to play an important role in this process (Ekman & Oster 1982).

Several theorists argue that emotion has evolved as a relevance detection and response preparation system. They posit an appraisal system that assesses the perceived antecedent conditions with respect to the organism's well being, its plans, and its goals (Levenson 1994), (Izard 1994), (Frijda 1994*c*), (Lazarus 1994). Scherer has studied this assessment process in humans and suggests that people affectively appraise events with respect to novelty, intrinsic pleasantness, goal/need significance, coping, and norm/self compatibility. Hence the level of cognition required for appraisals can vary widely (Scherer 1994).

These appraisals (along with other factors such as pain, hormone levels, drives, etc.) evoke a particular emotion which recruits response tendencies within multiple systems. These include physiological changes (such as modulating arousal level via the autonomic nervous system), adjustments in subjective experience, elicitation of behavioral response (such as approach, attack, escape, etc.), and displaying expression. The orchestration of these systems represents a generalized solution for coping with the demands of the original antecedent conditions. Plutchik (1991) calls this stabilizing feedback process *behavioral homeostasis*. Through this process, emotions establish a desired relation between the organism and the environment – pulling it towards certain stimuli and events and pushing it away from others. Much of the relational activity can be social in nature, motivating proximity seeking, social avoidance, chasing off offenders, etc. (Frijda 1994*b*).

The expressive characteristics of emotion in voice, face, gesture, and posture serve an important function in communicating emotional state to others. Levenson (1994) argues that this benefits people in two ways. First by allowing others to know how we feel, and second by influencing their behavior. For instance, the crying of an infant has a powerful mobilizing influence in calling forth nurturing behaviors of adults. Darwin argued that emotive signaling functions were selected for during the course of evolution because of their communicative efficacy. For members of a social species, the outcome of a particular act usually depends partly on the reactions of the significant others in the encounter. As argued by Scherer, the projection of how the others will react to these different possible courses of action largely determines the creature's behavioral choice. The signaling of emotion communicates the creature's evaluative reaction to a stimulus event (or act) and thus narrows the possible range of behavioral intentions that are likely to be inferred by observers. Darwin stressed the major significance of emotional expression as signals of behavioral intention and their role in social interaction.

8.1 Overview of the Motivation System

Kismet's motivations establish its nature by defining its "needs" and influencing how and when it acts to satisfy them. The nature of Kismet is to socially engage people, and ultimately to learn from them. Kismet's drive and emotion processes are designed such that the robot is in homeostatic balance, and an alert and mildly positive affective state, when it is interacting well with people, and when the interactions are neither overwhelming nor under-stimulating. This corresponds to an environment that affords high learning potential as the interactions slightly challenge the robot yet also allow Kismet to perform well.

Kismet's motivation system consists of two related subsystems, one which implements `drives` and a second which implements `emotions`.[1] There are several processes in the emotion system that model different arousal states (such as "interest", "calm", or "boredom"). These do not correspond to the *basic* emotions, such as the six proposed by Ekman (anger, disgust, fear, joy, sorrow, and surprise). Nonetheless, they have a corresponding expression and a few have an associated behavioral response. For our purposes, we will treat these arousal states as "emotions" in our system. Each subsystem serves a regulatory function for the robot (albeit in different ways) to maintain the robot's "well being". Each `drive` is modeled as an idealized homeostatic regulation process that maintains a set of critical parameters within a bounded range. There is one `drive` assigned to each parameter. Kismet's `emotions` are idealized models of basic emotions, where each serves a particular function (often social), each arises in a particular context, and each motivates Kismet to respond in an adaptive manner. They tend to operate on shorter, more immediate, and specific

[1] As a convention, we will use the boldface to distinguish parts of the architecture of this particular system from the general uses of those words. In this case, "`drives`" refers to the particular computational processes that are active in the system, while "drives" refers to the general uses of that word.

circumstances than the `drives` (which operate over longer time scales).

8.2 The Homeostatic Regulation Subsystem

Kismet's `drives` serve four purposes. First, they indirectly influence the attention system as described in chapter 6. Second, they influence behavior selection by preferentially passing activation to some behaviors over others. Third, they influence the affective state by passing activation energy to the `emotion` processes. Since the robot's expressions reflect its affective state, the `drives` indirectly control the affective cues the robot displays to people. Last, they provide a functional context that organizes behavior and perception. This is of particular importance for emotive appraisals.

The design of Kismet's homeostatic regulation subsystem is heavily inspired by ethological views of the analogous process in animals (McFarland & Bosser 1993). However, it is a simplified and idealized model of those discovered in living systems. One distinguishing feature of a `drive` is its temporally cyclic behavior. That is, given no stimulation, a `drive` will tend to increase in intensity unless it is satiated. This is analogous to an animal's degree of hunger or level of fatigue, both following a cyclical pattern.

Another distinguishing feature is its homeostatic nature. Each acts to maintain a level of intensity within a bounded range (neither too much nor too little). Its change in intensity reflects the ongoing needs of the robot and the urgency for tending to them. There is a desired operational point for each `drive` and acceptable bounds of operation around that point. We call this range the *homeostatic regime*. As long as a `drive` is within the homeostatic regime, the robot's needs are being adequately met. For Kismet, maintaining its `drives` within their homeostatic regime is a never-ending process. At any point in time, the robot's behavior is organized about satiating one of its `drives`.

Each `drive` is modeled as a separate process shown in figure 8-1. Each has a temporal input to implement its cyclic behavior. The activation energy A_{drive} of each `drive` ranges between $[A_{drive}^{-max}, A_{drive}^{+max}]$, where the magnitude of the A_{drive} represents its intensity. For a given A_{drive} intensity, a large positive magnitude corresponds to being under stimulated by the environment, whereas a large negative magnitude corresponds to being over stimulated by the environment. In general, each A_{drive} is partitioned into three regimes: an *under-stimulated regime*, an *overwhelmed regime*, and a *homeostatic regime*. A drive remains in its homeostatic regime when it is encountering its satiatory stimulus and that stimulus is of appropriate intensity. In the absence of the satiatory stimulus (or if the intensity is too low), the `drive` tends toward the under-stimulated regime. Alternatively, if the satiatory stimulus is too intense, the `drive` tends toward the overwhelmed regime. Hence to remain in balance, it is not sufficient that the satiatory stimulus be present, it must also be of a good quality.

In the current implementation there are three `drives`. They are:

- *Social*

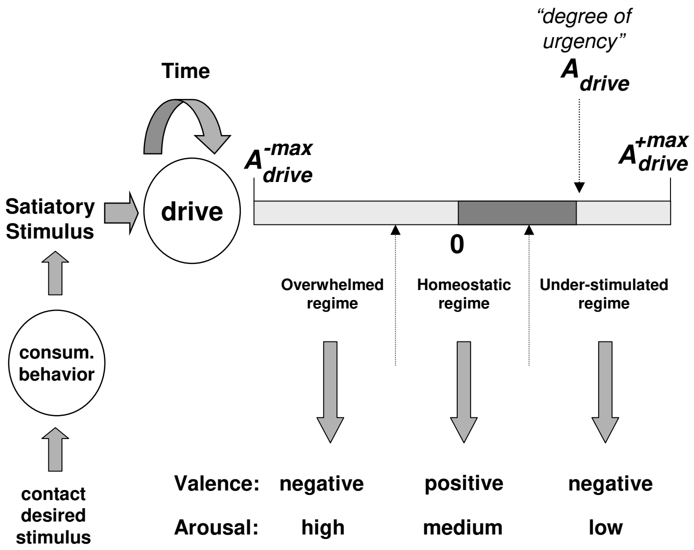


Figure 8-1: The homeostatic model of a drive process. See text.

- *Stimulation* (from environment)
- *Fatigue*

The Social drive

One `drive` is to be social, that is, to be in the presence of people and to be stimulated by people. This is important for biasing the robot to learn in a social context. On the under-stimulated extreme the robot is "lonely"; it is predisposed to act in ways to establish face-to-face contact with people. If left unsatiated, this `drive` will continue to intensify toward the under-stimulated end of the spectrum. On the overwhelmed extreme, the robot is "asocial"; it is predisposed to act in ways to avoid face-to-face contact. The robot tends toward the overwhelmed end of the spectrum when a person is over-stimulating the robot. This may occur when a person is moving too much or is too close to the robot's eyes.

The Stimulation drive

Another `drive` is to be stimulated, where the stimulation is generated externally by the environment, typically by engaging the robot with a colorful toy. This drive is important so that Kismet has an innate bias to interact with objects. This encourages the caregiver to draw the robot's attention to toys and events around the robot. On the under-stimulated end of this spectrum, the robot is "bored". This occurs if Kismet has been unstimulated over a period of time. On the overwhelmed part of the spectrum, the robot is "over-stimulated". This occurs when the robot receives more stimulation than its perceptual processes can handle well. In this case, the robot is biased to reduce its interaction with the environment, perhaps by closing its eyes or turning its head away from the stimulus. This `drive` is important for social learning as it encourages the caregiver to challenge the robot with new interactions.

The Fatigue drive

This `drive` is unlike the others in that its purpose is to allow the robot to shut out the external world instead of trying to regulate its interaction with it. While the robot is "awake", it receives repeated stimulation from the environment or from itself. As time passes, this `drive` approaches the "exhausted" end of the spectrum. Once the intensity level exceeds a certain threshold, it is time for the robot to "sleep". While the robot "sleeps", *all* `drives` return to their homeostatic regimes. After this, the robot awakens.

The `drives` spread activation energy to the `emotion` processes. In this manner, the robot's ability to satisfy its drives and remain in a state of "well being" is reflected by its affective state. When in the homeostatic regime, a `drive` spreads activation to those processes characterized by positive valence and balanced arousal. This corresponds to a "contented" affective state. When in the under-stimulated regime, a `drive` spreads activation to those processes characterized by negative valence and low arousal. This corresponds to a "bored" affective state that can eventually build to "sorrow". When in the overwhelmed regime, a `drive` spreads activation to those processes characterized by negative valence and high arousal. This corresponds to an affective state of "distress".

The `emotion` subsystem influences the robot's facial expression. The caregiver can read the robot's facial expression to interpret whether the robot is "distressed" or "content", and can adjust his/her interactions with the robot accordingly. The caregiver accomplishes this by adjusting either the type (social verses non-social) and/or the quality (low intensity, moderate intensity, or high intensity) of the stimulus presented to Kismet. These emotive cues are critical for helping the human work with the robot to establish and maintain a suitable interaction where the robot's drives are satisfied, where it is sufficiently challenged, yet where it is largely competent in the exchange.

We leave our discussion of `drives` for the moment. In chapter 9 we will present a detailed example of how the robot's `drives` influence behavior arbitration. In this way, they motivate which behavior the robot performs to bring the robot into contact with needed stimuli.

8.3 The Emotion Subsystem

The organization and operation of the *emotion subsystem* is strongly inspired by various theories of emotions in humans (as summarized in section 8.0.2). It is designed to be a flexible system that mediates between both environmental and internal stimulation to elicit an adaptive behavioral response that serves either social or self-maintenance functions. The `emotions` are triggered by various events which are evaluated as being of significance to the "well being" of the robot. Once triggered, each `emotion` serves a particular set of functions to establish a desired relation between the robot to its environment. They motivate the robot to come into contact with things that promote its "well being", and to avoid those that don't.

Antecedent conditions	Emotion	Behavior	Function
delay, difficulty in achieving goal of adaptive behavior	anger, frustration	complain	show displeasure to caregiver to modify his/her behavior
presence of an undesired stimulus	disgust	withdraw	signal rejection of presented stimulus to caregiver
presence of a threatening, overwhelming stimulus	fear, distress	escape	move away from a potentially dangerous stimuli
prolonged presence of a desired stimulus	calm	engage	continued interaction with a desired stimulus
success in achieving goal of active behavior, or praise	joy	display pleasure	reallocate resources to the next relevant behavior, (eventually to reinforce behavior)
prolonged absence of a desired stimulus, or prohibition	sorrow	display sorrow	evoke sympathy and attention from caregiver, (eventually to discourage behavior)
a sudden, close stimulus	surprise	startle response	alert
appearance of a desired stimulus	interest	orient	attend to new, salient object
need of an absent and desired stimulus	boredom	seek	explore environment for desired stimulus

Figure 8-2: Summary of the antecedents and behavioral responses that comprise Kismet's emotive responses. The antecedents refer to the eliciting perceptual conditions for each "emotion". The behavior column denotes the observable response that becomes active with the emotion. For some this is simply a facial expression. For others, it is a behavior such as "escape". The column to the right describes the function each emotive response serves Kismet.

8.3.1 Emotive Responses

We begin with a high level discussion of the emotional responses implemented in Kismet. Table 8-2 summarizes, under what conditions certain "emotions" and behavioral responses arise, and what function they serve the robot. This table is derived from the evolutionary, cross-species, and social functions hypothesized by Plutchik (1991), Darwin (1872), and Izard (1977). The table includes the six primary emotions proposed by Ekman (1992) along with three arousal states ("boredom", "interest", and "calm").

By adapting these ideas to Kismet, the robot's emotional responses mirror those of biological systems and therefore should seem plausible to a human. As discussed in section 8.0.2 this is very important for social interaction. Under close inspection, we also note that the four categories of proto-social responses from chapter 2 (affective, exploratory, protective, and regulatory) are represented within this table.

Each of the entries in this table has a corresponding affective display. For instance, the robot exhibits sadness upon the prolonged absence of a desired stimulus. This may occur if the robot has not been engaged with a toy for a long time. The sorrowful expression is intended to elicit attentive acts from the human caregiver. Another class of affective responses relates to behavioral performance. For instance, a successfully accomplished goal is reflected by a smile on the robot's face, whereas delayed progress is reflected by a stern expression. Exploratory responses include visual search for desired stimulus and/or maintaining visual engagement of a desired stimulus. Kismet currently has several protective responses, the strongest of which is to close its eyes and turn away from threatening or overwhelming stimuli. Many of these emotive responses serve a regulatory function. They bias the robot's behavior to bring it into contact with desired stimuli (orientation or exploration), or to avoid poor quality or dangerous stimuli (protection or rejection). In addition, the expression on the robot's face is a social signal to the human caregiver, who responds in a way to further promote the robot's "well being". Taken as a whole, these affective responses encourage the human to treat Kismet as a socially aware creature and to establish meaningful communication with it.

8.3.2 Components of Emotion

Several theories posit that emotional reactions consist of several distinct but interrelated facets (Scherer 1984), (Izard 1977). In addition, several appraisal theories hypothesize that a characteristic appraisal (or meaning analysis) triggers the emotional reaction in a context sensitive manner (Frijda 1994*b*), (Lazarus 1994), (Scherer 1994). Summarizing these ideas, an "emotional" reaction for Kismet consists of:

- A precipitating event,
- an affective appraisal of that event,
- a characteristic expression (face, voice, posture),
- and action tendencies that motivate a behavioral response

Two factors that we do not directly address with Kismet are:

- A subjective feeling state, or
- a pattern of physiological activity.

Kismet is not conscious, so it does not have feelings [2]. Nor does it have internal sensors that might sense something akin to physiological changes due to autonomic nervous activity. However, Kismet does have a parameter that maps to arousal level, so in a very simple fashion Kismet has a correlate to autonomic nervous system activity.

In living systems, it is believed that these individual facets are organized in a highly inter-dependent fashion. Physiological activity is hypothesized to physically prepare the creature to act in ways motivated by action tendencies. Furthermore, both the physiological activities and the action tendencies are organized around the adaptive implications of the appraisals that elicited the emotions. From a functional perspective, Smith (1989) and Russell (1997) suggest that the individual components of emotive facial expression are also linked to these emotional facets in a highly systematic fashion.

In the remainder of this chapter we discuss the relation between the eliciting condition(s), appraisal, action tendency, behavioral response, and observable expression in our implementation. An overview of the system is shown in figure 8-3. Some of these aspects are covered in greater depth in other chapters. For instance, detailed presentations of the expression of affect in Kismet's face, posture, and voice are covered in chapters 11 and 12. A detailed description of how the behavioral responses are implemented is given in chapter 9.

8.3.3 Emotive Releasers

We begin our discussion with the input to the emotion system. The input originates from the high level perceptual system, where it is fed into an associated *releaser* process. Each releaser can be thought of as a simple "cognitive" assessment that combines lower level perceptual features into behaviorally significant perceptual categories.

There are many different kinds of releasers defined for Kismet, each hand-crafted, and each combining different contributions from a variety of factors. Each releaser is evaluated with respect to the robot's "well being" and its goals. This evaluation is converted into an activation level for that releaser. If the perceptual features and evaluation are such that the activation level is above threshold (i.e., the conditions specified by that releaser hold), then its output is passed to its corresponding behavior process in the behavior system. It is also passed to the affective appraisal stage where it can influence the emotion system. There are a number of factors that contribute to the assessment made by each releaser. They are as follows:

[2] Several emotion theorists posit that consciousness is a requirement for an organism to experience feelings (see *The Feeling of what Happens: Body and Emotion in the Making of Consciousness* by Antonio Damasio). That Kismet is not conscious (at least not yet) is the author's phiosophical position

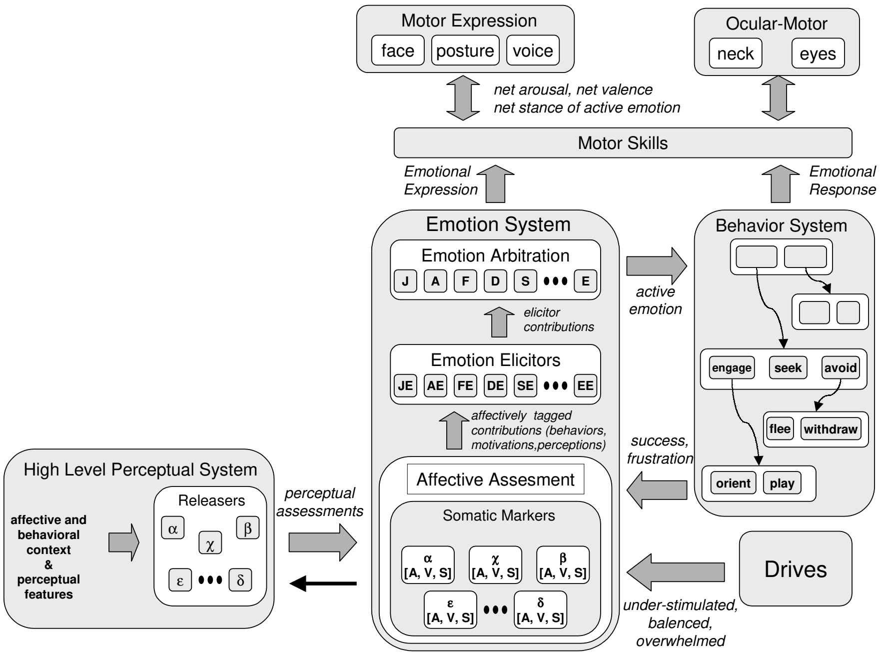


Figure 8-3: An overview of the emotion system. The antecedent conditions come through the high-level perceptual system where they are assessed with respect to the robot's "well being" and active goals. The result is a set of behavior and emotional response specific releasers. The emotional response releasers are passed to an affective appraisal phase where each active releaser is tagged with arousal, valence, and stance markers. In general, behaviors and `drives` can also send influences to this affective appraisal phase. All active contributions are filtered through the emotion elicitors for each emotion process. Each elicitor computes the relevance of its corresponding emotion process. In the emotion arbitration phase, the emotion processes compete for activation in a winner-take-all scheme. The winner can evoke its corresponding behavioral response (such as "flee" in the case of `fear`). It also evokes a corresponding facial expression, body posture, and vocal quality. These multi-modality expressive cues are arbitrated by the motor skill system.

- *Drives*: The active `drive` provides important context for many releasers. In general, it determines whether a given type of stimulus is either "desired" or "undesired". For instance, if the `social-drive` is active, then skin-toned stimuli are desirable, but colorful stimuli are undesirable (even if they are of good quality). Hence, this motivational context plays an important role in determining whether the emotional response will be one of incorporation or rejection of a presented stimulus.

- *Affective State*: The current affective state provides important context for certain releasers. A good example is the `soothing-speech` releaser described in chapter 7. Given a "soothing" classification from the affective intent recognizer, the `soothing-speech` releaser only becomes active if Kismet is distressed. Otherwise, the `neutral-speech` releaser is activated. This second stage of processing reduces the number of misclassifications between "soothing" speech verses "neutral" speech.

- *Active behavior(s)*: The behavioral state also plays an important role in disambiguating certain perceptual conditions. For instance, a `no-face` perceptual condition could correspond to several different possibilities. The robot could be engaged in a `seek-people` behavior, in which case a skin toned stimulus is a desired but absent stimulus. Initially this would encourage exploration. However, over time, this could contribute to an state of deprivation due to a long term loss. Alternatively, the robot could be engaged in an `escape` behavior. In this case, `no-face` corresponds to successful escape, a rewarding circumstance.

- *Perceptual state(s)*: The incoming percepts can contribute to the affective state on their own (such as a looming stimulus, for instance), or in combination with other stimuli (such as combining skin-tone with distance to perceive a distant person). An important assessment is how intense the stimulus is. Stimuli that are closer to the robot, move faster, or are larger in the field of view are more intense than stimuli that are further, slower, or smaller. This is an important measure of the quality of the stimulus, and to determine if the stimulus is a threat or not.

8.3.4 Affective Appraisal

Within the appraisal phase, each releaser with activation above threshold is appraised 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, behavioral, or motivational information is "tagged" with affective information. There are three classes of tags the SM uses to affectively characterize a given releaser. Each tag has an associated intensity that scales its contribution to the overall affective state. The *arousal* tag, A, specifies how arousing this factor is to the emotional system. It very roughly corresponds to the activity of the autonomic nervous system. Positive values correspond to a high arousal stimulus whereas negative values correspond to a low arousal stimulus. The *valence* tag, V, specifies how favorable or unfavorable this percept is to the emotional system. Positive values correspond to a pleasant stimulus whereas negative values correspond to an unpleasant stimulus. The *stance* tag, S, specifies how approachable the percept is. Positive values correspond to advance whereas negative values correspond to retreat. There are four types of appraisals considered:

- *Intensity*: The intensity of the stimulus generally maps to arousal. Threatening or very intense stimuli are tagged with high arousal. Absent or low intensity stimuli are tagged with low arousal. Soothing speech has a calming influence on the robot, so it also serves to lower arousal if initially high.

- *Relevance*: The relevance of the stimulus (whether or not it addresses the current goals of the robot) influences valence and stance. Stimuli that are relevant are "desirable" and are tagged with positive valence and approaching stance. Stimuli that are not relevant are "undesirable" and are tagged with negative arousal and withdrawing stance.

- *Intrinsic Pleasantness*: Some stimuli are hardwired to influence the robot's affective state in a specific manner. Praising speech is tagged with positive valence and slightly high arousal. Scolding speech is tagged with negative valence and low arousal (tending to elicit "sorrow"). Attentional bids alert the robot and are tagged with medium arousal. Looming stimuli startle the robot and are tagged with high arousal. Threatening stimuli elicit `fear` and are tagged with high arousal, negative valence, and withdrawing stance.

- *Goal Directedness*: Each behavior specifies a goal, i.e., a particular relation the robot wants to maintain with the environment. Success in achieving a goal promotes `joy` and is tagged with positive valence. Prolonged delay in achieving a goal results in "frustration" and is tagged with negative valence and withdrawing stance. The stance component increase slowly over time to transition from "frustration" to "anger".

As initially discussed in chapter 3, because there are potentially many different kinds of factors that modulate the robot's affective state (e.g., behaviors, motivations, perceptions), this tagging process converts the myriad of factors into a common currency that can be combined to determine the net affective state. For Kismet, the $[A, V, S]$ trio is the currency the emotion system uses to determine which emotional response should be active. In the current implementation, the affective tags for each releaser are specified by the designer. These may be fixed constants, or linearly varying quantities. In all, there are three contributing factors to the robot's net affective state:

- *Drives*: Recall that each `drive` is partitioned into three regimes: homeostatic, overwhelmed or under stimulated. For a given `drive`, each regime potentiates arousal and valence differently, which contribute to the activation of different `emotion` processes.

- *Behavior*: The success or delayed progress of the active behavior can directly influence the affective state. Success contributes to positive emotive responses, whereas delayed progress contributes to negative emotive responses such as frustration.

- *Releasers*: The external environmental factors that elicit emotive responses.

8.3.5 Emotion Elicitors

All somatically marked inputs are passed to the *emotion elicitor* stage. Each `emotion` process has as 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. Figure 8-4 summarizes how $[A, V, S]$ values map onto each `emotion` process. We show three 2-D slices through a 3-D space. This filtering is done independently for each type of affective tag. For instance, a valence contribution with a large negative value will not only contribute to the `sad` process, but to the `fear`, `distress`, `anger`, and `disgust` processes as well. Given all these factors, each elicitor computes its average $[A, V, S]$ from all the individual arousal, valence, and stance values that pass through its filter.

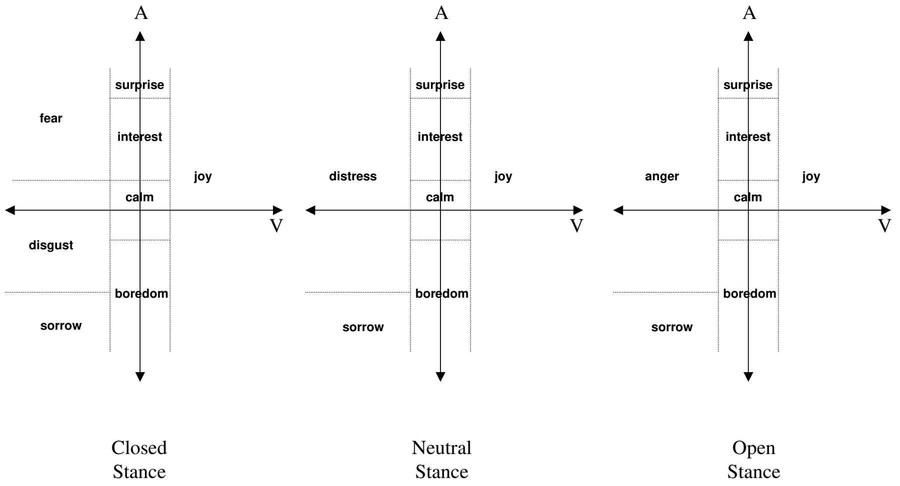


Figure 8-4: Mapping of arousal, valence, and stance dimensions, [A, V, S], to `emotions`. This figure shows three 2-D slices through this 3-D space.

Given the net $[A, V, S]$ of an elicitor, the activation level is computed next. Intuitively, the activation level for an elicitor corresponds to how "deep" the point specified by the net $[A, V, S]$ lies within the arousal, valence, and stance boundaries that define the corresponding "emotion" region shown in figure 8-4. This value is scaled with respect to the size of the region so as to not favor the activation of some processes over others in the arbitration phase. The contribution of each dimension to each elicitor is computed individually. If any one of the dimensions is not represented, then the activation level is set to zero. Otherwise, the A, V, and S contributions are summed together to arrive at the activation level of the elicitor. This activation level is passed on to the corresponding `emotion` process in the arbitration phase.

There are many different processes that contribute to the overall affective state. Influences are sent by `drives`, the active behavior, and releasers. We have tried several different schemes for computing the net contribution to a given `emotion` process, but found this one to have the nicest properties. In an earlier version, we simply averaged all the incoming contributions. This tended to "smooth" the net affective state to an unacceptable degree. For instance, if the robot's `fatigue-drive` is high (biasing a low arousal state) and a threatening toy appears (contributing to a strong negative valence and high arousal), the averaging technique could result in a slightly negative valence and neutral arousal. This is insufficient to evoke `fear` and an escape response when the robot should protect itself. As an alternative, we could hard-wire certain releasers directly to `emotion` processes. However, it is not clear how this approach supports the influence of `drives` and behaviors, whose affective contributions change as a function of time. For instance, a given `drive` contributes to `fear`, `sorrow`, or `interest` processes depending on its current activation regime. Our current approach balances the constraints of having certain releasers contribute heavily and directly to the appropriate emotive response, while accommodating those influences that contribute to different `emotions` as a function of time. The end result also has nice properties for generating facial expressions that reflect this assessment process in a rich way. This is important for social interaction as originally argued by Darwin. This expressive benefit is discussed in further detail in chapter 11).

8.3.6 Emotion Activation

Next, the activation level of each `emotion` process is computed. There is a process defined for each `emotion` listed in table 8-2: `joy`, `anger`, `disgust`, `fear`, `sorrow`, `surprise`, `interest`, `boredom`, and `calm`.

Numerically, the activation level $A_{emotion}$ of each `emotion` process can range between $[0, A^{max}_{emotion}]$ where $A^{max}_{emotion}$ is an integer value determined empirically. Although these processes are always active, their intensity must exceed a threshold level before they are expressed externally. The activation of each process is computed by the equation:

$$A_{emotion} = \sum(E_{emotion} + B_{emotion} + P_{emotion}) - \delta_t$$

where $E_{emotion}$ is the activation level of its affiliated elicitor process, $B_{emotion}$ is a DC bias that can be used to make some `emotion` processes easier to activate than others. $P_{emotion}$ adds a level of persistence to the active emotion. This introduces a form of inertia so that different `emotion` processes don't rapidly switch back and forth. Finally, δ_t is a decay term that restores an `emotion` to its bias value once the `emotion` becomes active. Hence, unlike `drives` (which contribute to the robot's longer term "mood"), the `emotions` have an intense expression followed by decay to a baseline intensity. The decay takes place on the order of seconds.

8.3.7 Arbitration

Next, the `emotion` processes compete for control in a winner-take-all arbitration scheme based on their activation level. The activation level of an `emotion` process is a measure of its relevance to the current situation. Each of these processes is distinct from the others and regulates the robot's interaction with its environment in a distinct manner. Each becomes active in a different environmental (or internal) situation. Each motivates a different observable response by spreading activation to a specific behavior process in the behavior system. If this amount of activation is strong enough, then the active emotion can "seize" temporary control and force the behavior to become expressed. In a process of behavioral homeostasis as proposed by Plutchik, the emotive response maintains activity through feedback until the correct relation of robot to environment is established.

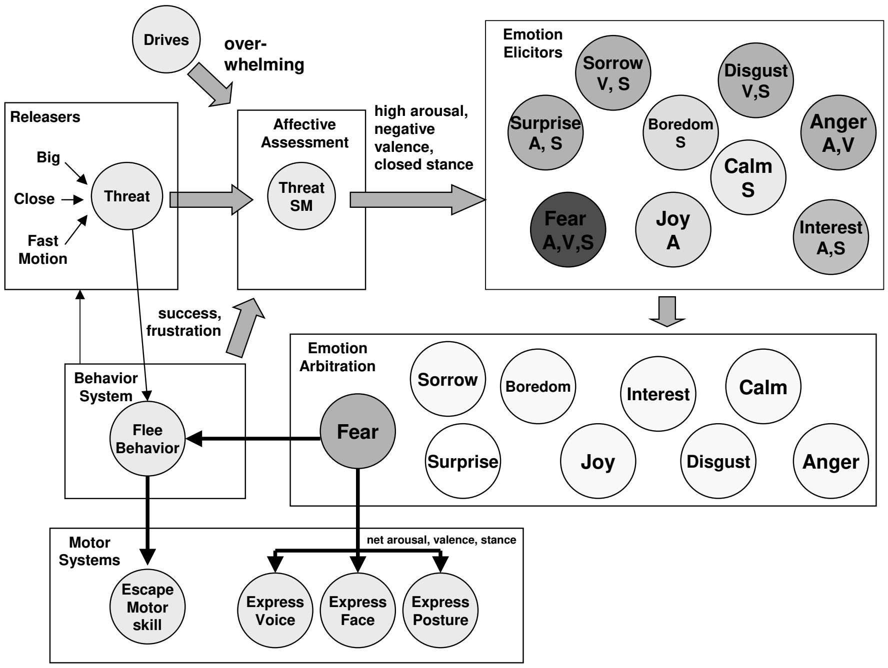


Figure 8-5: The implementation of the `fear` emotion. The releaser for `threat` is passed to the affective assessment phase. It is tagged with high arousal, negative valence, and closed stance by the corresponding somatic marker process. This affective information is then filtered by the corresponding elicitor of each emotion process. Darker shading corresponds to a higher activation level. Note that only the `fear elicitor` process has each of the arousal, valence, and stance conditions matched (hence, it has the darkest shading). As a result, it is the only one that passes activation to its corresponding `emotion` process.

Concurrently, the net $[A, V, S]$ of the active process is sent to the expressive components of the motor system, causing a distinct facial expression, vocal quality, and body posture to be exhibited. The strength of the facial expression reflects the level of activation of the `emotion`. Figure 8-5 illustrates the emotional response network for the `fear` emotion process. Affective networks for the other responses in table 8-2 are defined in a similar manner. By modeling Kismet's emotional responses after those of living systems, people have a natural and intuitive understanding of Kismet's emotional behavior and how to influence it.

There are two threshold levels for each `emotion` process: one for expression and one for behavioral response. 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, if the caregiver shakes a toy in a threatening manner near the robot's face, Kismet will first exhibit a fearful expression and then activate the escape response. By staging the response in this manner, the caregiver gets immediate expressive feedback that he/she is frightening the robot. If this was not the intent, then the caregiver has an intuitive understanding of why the robot is frightened and modifies behavior accordingly. The facial expression also sets up the human's expectation of what behavior will soon follow. As a result, the caregiver not only sees what the robot is doing, but has an understanding of why.

8.4 Regulating Playful Interactions

Kismet's design relies on the ability of people to interpret and understand the robot's behavior. If this is the case, then the robot can use expressive feedback to tune the caregiver's behavior in a manner that benefits the interaction.

In general, when a `drive` is in its homeostatic regime, it potentiates positive valenced `emotions` such as `joy` and arousal states such as `interest`. The accompanying expression tells the human that the interaction is going well and the robot is poised to play (and ultimately learn). When a `drive` is not within the homeostatic regime, negative valenced `emotions` are potentiated (such as `anger`, `fear`, or `sorrow`) which produces signs of distress on the robot's face. The particular sign of distress provides the human with additional cues as to what is "wrong" and how he/she might correct for it. For example, overwhelming stimuli (such as a rapidly moving toy) produce signs of `fear`. Infants often show signs of anxiety when placed in a confusing environment.

Note that the same sort of interaction can have a very different "emotional" effect on the robot depending on the motivational context. For instance, playing with the robot while all `drives` are within the homeostatic regime elicits `joy`. This tells the human that playing with the robot is a good interaction to be having at this time. However, if the `fatigue-drive` is deep into the under stimulated end of the spectrum, then playing with the robot actually prevents the robot from going to "sleep". As a result, the `fatigue-drive` continues to increase in intensity. When high enough, the `fatigue-drive` begins to potentiate `anger` since the goal of "going to sleep" is blocked. The human may interpret this as the robot acting "cranky" because it is "tired".

In this section we present a couple of interaction experiments to illustrate how the robot's motivations and facial expressions can be used to regulate the nature and quality of social exchange with a person. Several chapters in this thesis give other examples of this process (chapters 7 and 13 in particular). Whereas the examples in this chapter focus on the interaction of `emotions`, `drives` and expression, these other chapters focus on the perceptual conditions of eliciting different emotive responses.

Each experiment involves a caregiver interacting with the robot using a colorful toy. Data was recorded on-line in real-time during the exchange. Figures 8-6 and 8-7 plot the activation levels of the appropriate `emotions`, `drives`, behaviors, and percepts. `Emotions` are always plotted together with activation levels ranging from 0 to 2000. Percepts, behaviors, and `drives` are often plotted together. Percepts and behaviors have activation levels that also range from 0 to 2000, with higher values indicating stronger stimuli or higher potentiation respectively. `Drives` have activation ranging from -2000 (the overwhelmed extreme) to 2000 (the under-stimulated extreme). The perceptual system classifies the toy as a non-face stimuli, hence it serves to satiate the `stimulation drive`. The motion generated by the object gives a rating of the stimulus intensity. The robot's facial expressions reflect its ongoing motivational state and provides the human with visual cues as to how to modify the interaction to keep the robot's `drives` within homeostatic ranges.

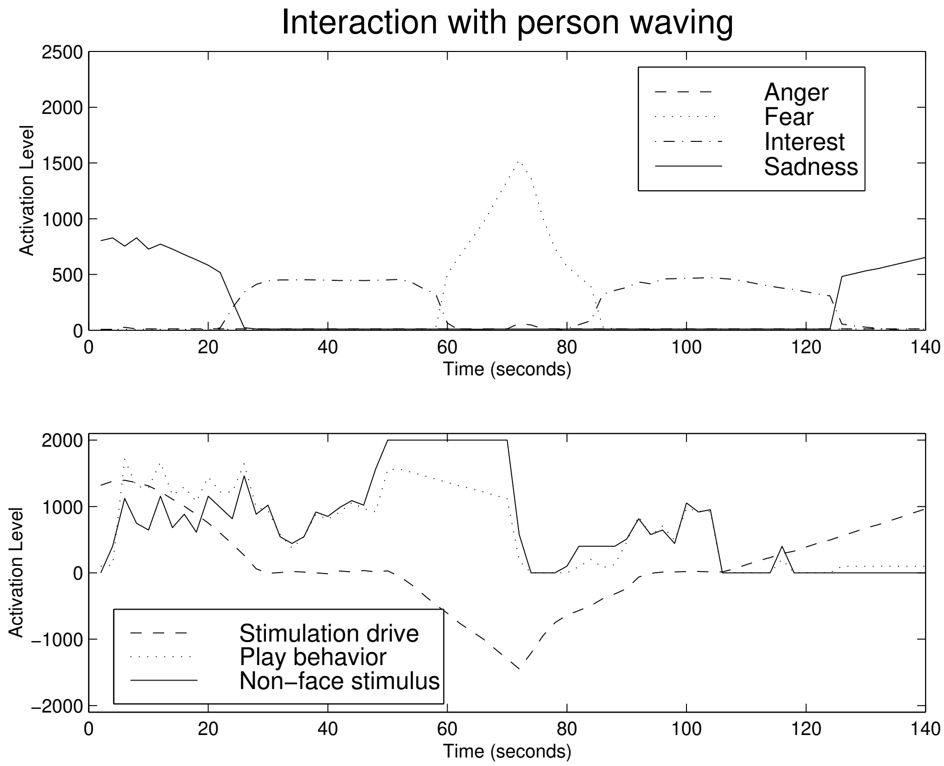


Figure 8-6: Experimental results for the robot interacting with a person waving a toy. The top chart shows the activation levels of the `emotions` involved in this experiment as a function of time. The bottom chart shows the activation levels of the `drives`, behaviors, and percepts relevant to this experiment. So long as the waving continues at a reasonable intensity, the robot remains `interested`. When the stimulus intensity becomes too great, the robot begins to show `fear`.

For the waving toy experiment, a lack of interaction before the start of the run

($t \leq 0$) places the robot in a "sad" emotional state as the `stimulation-drive` lies in the `under-stimulated` end of the spectrum for activation $A_{stimulation} \geq 400$. This corresponds to a long-term loss of a desired stimulus. From $5 \leq t \leq 25$ a salient toy appears and stimulates the robot within the acceptable intensity range ($400 \leq A_{nonFace} \leq 1600$) on average. This corresponds to waving the toy gently in front of the robot. This amount of stimulus causes the `stimulation-drive` to diminish until it resides within the homeostatic range, and a look of interest appears on the robot's face. From $25 \leq t \leq 45$ the stimulus maintains a desirable intensity level, the `drive` remains in the homeostatic regime, and the robot maintains "interest". At $45 \leq t \leq 70$ the toy stimulus intensifies to large, sweeping motions which threaten the robot ($A_{nonFace} \geq 1600$). This causes the `stimulation-drive` to migrate toward the overwhelmed end of the spectrum and the `fear` process to become active. As the `drive` approaches the overwhelmed extreme, the robot's face displays an intensifying expression of fear. Around $t = 75$ the expression peaks at an emotional level of $A_{fear} = 1500$ and experimenter responds by stopping the waving stimulus before the escape response is triggered. With the threat gone, the robot "calms" down somewhat as the `fear` process decays. The interaction then resumes at an acceptable intensity. Consequently, the `stimulation-drive` returns to the homeostatic regime and the robot displays interest again. At $t \geq 105$ the waving stimulus stops for the remainder of the run. Because of the prolonged loss of the desired stimulus, the robot is under-stimulated and an expression of sadness reappears on the robot's face.

Figure 8-7 illustrates the influence of the `fatigue-drive` on the robot's motivational and behavioral state when interacting with a caregiver. Over time, the `fatigue-drive` increases toward the `under-stimulated` end of the spectrum. As the robot's level of "fatigue" increases, the robot displays stronger signs of being "tired". At time step $t = 95$, the `fatigue-drive` moves above the threshold value of 1600 which is sufficient to activate the `sleep` behavior when no other interactions are occurring. The robot remains "asleep" until all `drives` are restored to their homeostatic ranges. Once this occurs, the activation level of the "sleep" behavior decays until the behavior is no longer active and the robot "wakes up" in an calm state. However, at time step $t = 215$, the plot shows what happens if a human continues to interact with the robot despite its "fatigued" state. The robot cannot fall asleep as long as the `play-with-toy` behavior wins the competition and inhibits the `sleep` behavior. If the `fatigue-drive` exceeds threshold and the robot cannot fall "asleep", the robot begins to show signs of "frustration". Eventually the robot's level of "frustration" increases until the robot appears angry (at t=1800). Still the human persists with the interaction. Eventually the robot's fatigue level reaches near maximum and the `sleep` behavior wins out.

These experiments illustrate a few of the emotive responses of table 8-2 that arise when engaging a human. It demonstrates how the robot's emotive cues can be used to regulate the nature and intensity of the interaction, and how the nature of the interaction influences the robot's behavior. The result is an ongoing "dance" between robot and human aimed at maintaining the robot's `drives` within homeostatic bounds and maintaining a good "emotive" state. If the robot and human are good partners, the robot remains "interested" most of the time. These expressions indicate that the

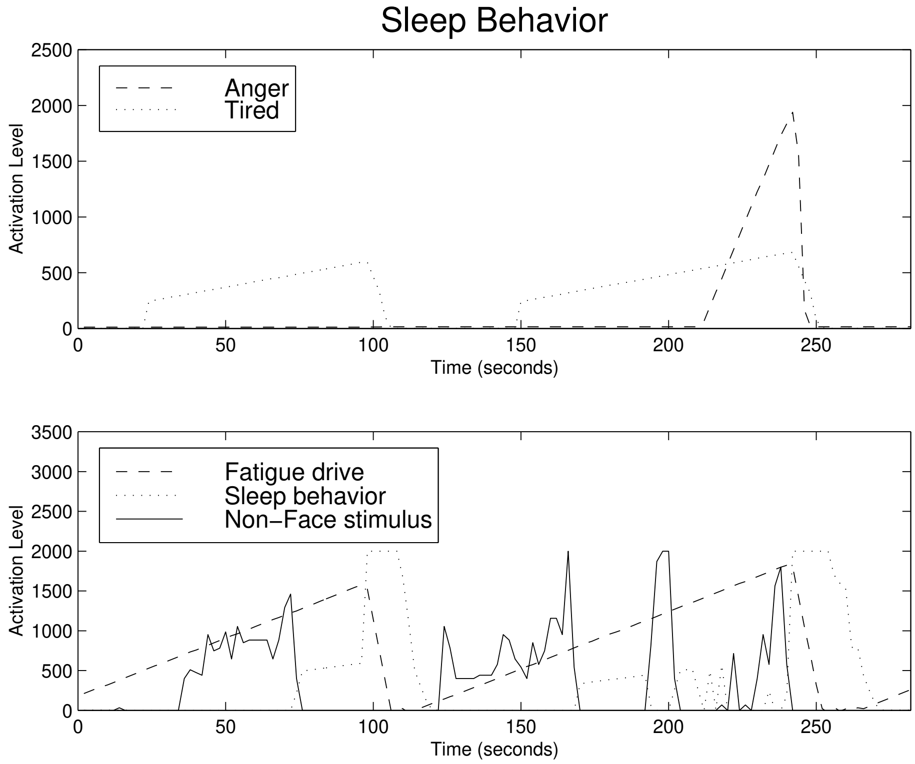


Figure 8-7: Experimental results for long-term interactions of the `fatigue-drive` and the `sleep` behavior. The `fatigue-drive` continues to increase until it reaches an activation level that potentiates the `sleep` behavior. If there is no other stimulation, this will allow the robot to activate the `sleep` behavior.

interaction is of appropriate intensity for the robot.

8.5 Limitations and Extensions

Kismet's motivation system appears adequate for generating infant-like social exchanges with a human caregiver. To incorporate social learning, or to explore socio-emotional development, a number of extensions could be made.

Extension to Drives

To support social learning, we could incorporate new drives into the system. For instance, a *self-stimulation* drive could motivate the robot to play by itself, perhaps by modulating its vocalizations to learn how to control its voice to achieve specific auditory effects. A *mastery/curiosity* drive might motivate the robot to balance exploration verses exploitation when learning new skills. This would correlate to the amount of novelty the robot experiences over time. If its environment is too predictable, this drive could bias the robot to prefer novel situations. If the environment is highly unpredictable for the robot, it could show distress which would encourage the caregiver to slow down.

Ultimately, the `drives` should provide the robot with a reinforcement signal as Blumberg (1996) has done. This could be used to motivate the robot to learn communication skills that satisfy its `drives`. For instance, the robot may discover that making a particular vocalization results in having a toy appear. This has the additional effect that the `stimulation-drive` becomes satiated. Over time, through repeated games with the caregiver, the caregiver could treat that particular vocalization as a request for a specific toy. Given enough of these consistent, contingent interactions during play, the robot may learn to utter that vocalization with the *expectation* that its `stimulation-drive` be reduced. This would constitute a simple act of meaning.

Extensions to Emotions

Kismet's `drives` relate to a hardwired preference for certain kinds of stimuli. The power of the emotion system is its ability to associate affective qualities to different kinds of events and stimuli. As discussed in chapter 7, the robot could have a learning mechanism by which it uses the caregiver's affective assessment (praise or prohibition) to affectively tag a particular object or action. This is of particular importance if the robot is to learn something novel – i.e., something for which it does not already have an explicit evaluation function. Through a process of social referencing (discussed in chapter 2) the robot could learn how to organize its behavior by using the caregiver's affective assessment. Human infants continually encounter novel situations, and social referencing plays an important role in their cognitive, behavioral, and social development.

Another aspect of learning involves learning new emotions. These are termed *secondary* emotions (Damasio 1994). Many of these are socially constructed through interactions with others.

One might pose the question "what would it take to give Kismet genuine emotions?". This question is posed in Picard (1997). Kismet's emotion system addresses some of the aspects of emotions in simple ways. For instance, the robot carries out some simple "cognitive" appraisals. The robot expresses its "emotional" state. It also uses analogs of emotive responses to regulate its interaction with the environment to promote its "well being". However, there are many aspects of human emotions that the system does not address, nor does it address any to an adult human level.

For instance, many of the appraisals proposed by Scherer are highly cognitive and require substantial social knowledge and self awareness. The robot does not have any "feeling" states. It is unclear if consciousness is required for this or not, or what consciousness would even mean for a robot. Kismet does not reason about the emotional state of others. There have been a few systems that have been designed for this competence that employ symbolic models (Ortony, Clore & Collins 1988), (Elliot 1992), or (Reilly 1996). The ability to recognize, understand, and reason about another's emotional state is an important ability for having a theory of mind about other people, which is considered by many to be a requisite of adult-level social intelligence (Dennett 1987).

Another aspect we have not addressed is the relation between emotional behavior and personality. Some systems tune the parameters of their emotion systems to produce synthetic characters with different personalities. For instance, characters who are quick to anger, more timid, friendly, and so forth (Yoon et al. 2000). In a similar manner, Kismet has its own version of a synthetic personality, but we have tuned it to this particular robot and have not tried to experiment with different synthetic personalities. This could be an interesting set of studies.

This leads us to a discussion of both an important feature and limitation of the motivation system – the number of parameters. Motivation systems of this nature are capable of producing rich, dynamic, compelling behavior at the expense of having many parameters that must be tuned. For this reason, systems of the complexity that rival Kismet are hand-crafted. If learning is introduced, it is done so in limited ways. This is a trade-off of the technique and there are no obvious solutions. Designers scale the complexity of these systems by maintaining a principled way of introducing new releasers, appraisals, elicitors, etc. The functional boundaries and interfaces between these stages must be honored.

8.6 Summary

Kismet's emotive responses enable the robot to use social cues to tune the caregiver's behavior so that both perform well during the interaction. Kismet's motivation system is explicitly designed so that a state of "well being" for the robot corresponds to an environment that affords a high learning potential. This often maps to having a caregiver actively engaging the robot in a manner that is neither under-stimulating nor overwhelming. Furthermore, the robot actively regulates the relation between itself and its environment, to bring itself into contact with desired stimuli and to avoid undesired stimuli. All the while, the cognitive appraisals leading to these actions are displayed on the robot's face. Taken as a whole, the observable behavior that results from these mechanisms conveys intentionality to the observer. This is not surprising as they are well matched to the proto-social responses of human infants. In numerous examples presented throughout this thesis, people interpret Kismet's behavior as the product of intents, beliefs, desires, and feelings. They respond to Kismet's behaviors in these terms. This produces natural and intuitive social exchange on a physical and affective level.

Chapter 9

The Behavior System

With respect to social interaction, Kismet's behavior system must be able to support the kinds of behaviors that infants engage in. Furthermore, it should be initially configured to emulate those key action patterns observed in an infant's initial repertoire that allow him/her to interact socially with the caregiver. Because the infant's initial responses are often described in ethological terms, the architecture of the behavior system adopts several key concepts from ethology regarding the organization of behavior (Tinbergen 1951), (Lorenz 1973), (McFarland & Bosser 1993), (Gould 1982).

From the literature on pre-speech communication of infants, we can extract several key action patterns that serve to foster social interaction between infants and their caregivers (Bullowa 1979), (de Boysson-Bardies 1999). In chapter 2, we discussed these action patterns, the role they play in establishing social exchanges with the caregiver, and the importance of these exchanges for learning meaningful communication acts. Chapter 8 presented how the robot's homeostatic regulation mechanisms and emotional models take part in many of these proto-social responses. This chapter presents the contributions of the behavior system to these responses.

9.0.1 Infant Social Responses

Infants are born with innate reflexes and responses that are elicited by the presence of the caregiver. The infant's action patterns are close enough to the adult form to be recognizable and interpretable by the mother. They are highly organized, often consisting of a chain of discrete temporal episodes (lasting only a few seconds). Within each episode, the behavior is well coordinated and synchronized, often consisting of facial expressions, vocalizations, and body movement (Newson 1979). These responses compel her to treat her infant as a sentient and communicating human being, to nurture him, and to teach him. Over and over again, events which are at first only the results of automatic action patterns, or which are spontaneous or accidental and beyond the control of either mother or infant, are endowed with significance because of the way the mother reacts towards the baby in the light of the event and its effect upon him. Without this, the human dialog cannot take place.

Establishing a dialog

Within each session with her infant, the mother makes constant micro-adjustments to changes in her infant's behavior. To make these adjustments, she takes into account her infant's current abilities, his attention span, and his level of arousal (Garvey 1974). Aware of her infant's limitations, her responses are aimed toward establishing and maintaining his interest. Often she tries to re-orient his eyes and face towards her so that they hold each other in mutual gaze. Once in mutual regard, she exaggerates, slows down, and simplifies her behavioral displays to fit within her infant's information processing abilities, which are slower and more limited than her own (Heckhausen 1987). She adjusts the timing of her responses, introduces variations about a common theme to the interaction, and tries to balance his agenda with her own agenda for him (Kaye 1979). She will purposely leave spaces between her own repetitious utterances and gestures for the infant to fill. In the meantime, she is constantly watching and listening for new initiatives from him. Over time, the baby's ability to take turns becomes more flexible and regular (Ekerman 1993), (Rutter & Durkin 1987). It is a critical skill for social learning and leads to dynamic exchanges between caregiver and infant (Eckerman & Stein 1987), (Ross & Lollis 1987).

Dynamics of Social Interaction

Tronick et al. (1979) identify five phases that characterize social exchanges between three-month-old infants and their caregivers: *initiation, mutual-orientation, greeting, play-dialog,* and *disengagement*. Each phase represents a collection of behaviors which mark the state of the communication. Not every phase is present in every interaction. For example, initiation is not necessary of both partners are already in mutual regard. A greeting does not ensue if mutual orientation is not established. Furthermore, a sequence of phases may appear multiple times within a given exchange, such as repeated greetings before the play-dialog phase begins, or cycles of disengagement to mutual orientation to disengagement. We summarize these phases below:

- *Initiation*: In this phase one of the partners is involved but the other is not. Frequently it is the mother who tries to actively engage her infant. She typically moves her face into an in-line position, modulates her voice in a manner characteristic of attentional bids, and generally tries to get the infant to orient towards her. Chapters 6 and 7 present how these cues are naturally and intuitively used by naive subjects to get Kismet's attention.

- *Mutual Orientation*: Here both partners attend to the other. Their faces may be either neutral or bright. The mother often smoothes her manner of speech, and the infant may make isolated sounds. Kismet's ability to locate eyes in its visual field and direct its gaze towards them is particularly powerful during this phase.

- *Greeting*: Both partners attend to the other as smiles are exchanged. Often, when the baby smiles, its limbs go into motion and the mother becomes increasingly animated. This is the case for Kismet's greeting response where the robot's smile is accompanied by small ear motions. Afterwards, both move to neutral or bright faces. Now they may transition back to mutual orientation, initiate another greeting, enter into a play dialog, or disengage.

- *Play Dialog*: During this phase, the mother speaks in a burst-pause pattern and the infant vocalizes during the pauses (or makes movements of intention to do so). The mother responds with a change in facial expression or a single burst of vocalization. In general, this phase is characterized by mutual positive affect conveyed by both partners. Over time the affective level decreases and the infant looks away. This chapter discusses Kismet's turn-taking behavior.
- *Disengagement*: Finally, one of the partners looks away while the other is still oriented. Both may then disengage, or one may try to reinitiate the exchange.

Proto-Social Skills for Kismet

In chapter 2, we categorized a variety of infant proto-social responses into four categories. With respect to Kismet, the *affective responses* are important because they allow the caregiver to attribute feelings to the robot, which encourages the human to modify the interaction to bring Kismet into a positive emotional state. The *exploratory responses* are important because they allow the caregiver to attribute curiosity, interest, and desires to the robot. The human can use these responses to direct the interaction towards things and events in the world. The *protective responses* are important to keep the robot from damaging stimuli, but also to elicit concern and caring responses from the caregiver. The *regulatory responses* are important for pacing the interaction at a level that is suitable for both human and robot.

In addition, Kismet needs skills that allow it to engage the caregiver in tightly coupled dynamic interactions. Turn-taking is one such skill that is critical to this process (Garvey 1974). It enables the robot to respond to the human's attempts at communication in a tightly temporally correlated and contingent manner. If the communication modality is facial expression, then the interaction may take the form of an imitative game (Eckerman & Stein 1987). If the modality is vocal, then proto-dialogs can be established (Rutter & Durkin 1987). This dynamic is a cornerstone of the social learning process that transpires between infant and adult.

9.1 Views from Ethology on the Organization of Behavior

For Kismet to engage a human in this dynamic, natural, and flexible manner, its behavior needs to be robust, responsive, appropriate, coherent, and directed. We can learn much from the behavior of animals, who must behave effectively in a complex dynamic environment in order to satisfy their needs and maintain their well being. This entails having the animal apply its limited resources (finite number of sensors, muscles and limbs, energy, etc.) to perform numerous tasks. Given a specific task, the animal exhibits a reasonable amount of persistence. It works to accomplish a goal, but not at the risk of ignoring other important tasks if the current task is taking too long.

For ethologists, the animal's observable behavior attempts to satisfy its competing physiological needs in an uncertain environment. Animals have multiple needs that must be tended to, but typically only one need can be satisfied at a time (hunger, thirst, rest, etc.). Ethologists strive to understand how animals organize their behaviors and arbitrate between them to satisfy these competing goals, how animals decide what to do for how long, and how they decide which opportunities to exploit (Gallistel 1980).

By observing animals in their natural environment, ethologists have made significant contributions to understanding animal behavior and providing descriptive models to explain its organization and characteristics. In this section, we present several key ideas from ethology which have strongly influenced the design of the behavior system. These theories and concepts specifically address the issues of relevance, coherence, and concurrency which are critical for animal behavior as well as the robot's behavior. The behavior system we have constructed is similar in spirit to that of (Blumberg 1996), which has also drawn significant insights from animal behavior.

Behaviors

Ethologists such as Lorenz (1973) and Tinbergen (1951) viewed behaviors as being complex, temporally extended patterns of activity that address a specific biological need. In general, the animal can only pursue one behavior at a time such as feeding, defending territory, or sleeping. As such, each behavior is viewed as a self-interested goal-directed entity that competes against other behaviors for control of the creature. They compete for expression based on a measure of relevance to the current internal and external situation. Each behavior determines its own degree of relevance by taking into account the creature's internal motivational state and its perceived environment.

Perceptual Contributions

For the perceptual contribution to behavioral relevance, Tinbergen and Lorenz posited the existence of innate and highly schematic perceptual filters called *releasers*. Each releaser is an abstraction for the minimal collection of perceptual features that reliably identify a particular object or event of biological significance in the animal's natural environment. Each releaser serves as the perceptual elicitor to either a group of behaviors or to a single behavior. The function of each releaser is to determine if all perceptual conditions are right for its affiliated behavior to become active. Because each releaser is not overly specific or precise, it is possible to "fool" the animal by devising a mock stimulus that has the right combination of features to elicit the behavioral response. In general, releasers are conceptualized to be simple, fast, and just adequate. When engaged in a particular behavior, the animal tends to only attend to those features that characterize its releaser.

Motivational Contributions

Ethologists have long recognized that an animal's internal factors contribute to behavioral relevance. We discussed two examples of motivating factors in chapter 8, namely homeostatic regulatory mechanisms and emotions. Both serve regulatory functions for the animal to maintain its state of well being. The homeostatic mechanisms often work on slower time-scales and bring the animal into contact with innately specified needs such as food, shelter, water, etc. The emotions operate on faster time scales and regulate the relation of the animal with its (often social) environment. An active emotional response can be thought of as temporarily "seizing" control of the behavior system to force the activation of a particular observable response in the absence of other contributing factors. By doing so, the emotion addresses the antecedent conditions that evoked it. Emotions bring the animal close to things that benefit its survival, and motivate it to avoid those circumstances that are detrimental to its well being. Emotional responses are also highly adaptive, and the animal can learn how to apply them to new circumstances.

Overall, motivations add richness and complexity to an animal's behavior, far beyond a stimulus-response or reflexive sort of behavior which might occur if only perceptual inputs were considered, or if there were a simple hardwired mapping. Motivations determine the internal agenda of the animal which changes over time. As a result, the same perceptual stimulus may result in a very different behavior. Or conversely, very different perceptual stimuli may result in an identical behavior given a different motivational state. The motivational state will also affect the strength of perceptual stimuli required to trigger a behavior. If the motivations heavily predispose a particular behavior to be active, then a weak stimulus might be sufficient to activate the behavior. Conversely, if the motivations contribute minimally, then a very strong stimulus is required to activate the behavior. Scherer (1994a) discusses the advantages of having emotions decouple the stimulus from the response in emotive reactions. For members in a social species, one advantage is the latency this decoupling introduces between affective expression and ensuing behavioral response. This makes an animal's behavior more readable and predictable to the other animals that it is in close contact.

Behavior Groups

Up to this point, we have taken a rather simplified view of behavior. In reality, a behavior to reduce hunger may be composed of collections of related behaviors. Within each group behaviors are activated in turn, which produces a *sequence* of distinguishable motor acts. For instance, one behavior may be responsible for eating while the others are responsible for bringing the animal near food. In this case, eating is the *consummatory behavior* because it serves to directly satiate the affiliated hunger drive when active. It is the last behavior activated in a sequence simply because once the drive is satiated, the motivation for engaging in the eating behavior is no longer present. This frees the animal's resources to tend to other needs. The other behaviors in the group are called *appetitive behaviors*. The appetitive behaviors represent separate behavioral strategies for bringing the animal to a relationship with its environment where it can directly activate the desired consummatory behavior. Lorenz considered the consummatory behavior to constitute the "goal" of the preceding appetitive behaviors. The appetitive behaviors "seek out" the appropriate releaser that will ultimately result in eliciting the desired consummatory behavior.

Given that each behavior group is composed of competing behaviors, a mechanism is needed to arbitrate between them. For appropriately persistent behavior, the arbitration mechanism should have some "inertia" term which allows the currently active behavior enough time to achieve its goal. However, the active behavior should eventually allow other behaviors to become active if its rate of progress is too slow. Some behaviors (such as feeding) might have a higher priority than other behaviors (such as preening), however both are important for the creature. Sometimes it is important for the preening behavior to be preferentially activated even if it has inherently lower priority than feeding. Hence, the creature must perform "time-sharing" where lower priority activities are given a chance to execute despite the presence of a higher priority activity.

Behavior Hierarchies

Tinbergen's *hierarchy of behavior centers* (an example is shown in figure 9-1) is a more general explanation of behavioral choice that incorporates many of the ideas mentioned above (Tinbergen 1951). It accounts for behavioral sequences that link appetitive behaviors to the desired consummatory behavior. It also factors in both perceptual and internal factors in behavior selection.

In Tinbergen's hierarchy, the nodes stand for *behavior centers* and the links symbolize *transfer* of energy between nodes. Behaviors are categorized according function (i.e., which biological need it serves the animal). Each class of behavior is given a separate hierarchy. For instance, behaviors such as feeding, defending territory, procreation, etc. are placed at the pinnacle of its respective hierarchy. These top level centers must be "motivated" by a form of energy. Hence drive factors direct behavior at the top level of the hierarchy. Figure 9-1 is Tinbergen's proposed model to explain the procreating behavior of the male stickleback fish

Activation energy is specific to the entire category of behavior (its respective hierarchy) and can "flow" down the hierarchy to motivate the behavior centers (i.e., groups of behaviors) below. Paths leading down from the top-level center pass the energy to subordinate centers. However, each of these conduits is "blocked" unless the correct perceptual conditions for that group of behaviors (i.e., the behavior center) is present. This is represented as the rectangles under each node in figure 9-1. Hence, a behavior center under the block is prevented from being executed until the appropriate stimulus is encountered. When this occurs, the block is removed and the flow of energy allows the behaviors within the group to execute and subsequently to pass activation to lower centers.

The hierarchical structure of behavior centers ensures that the creature will perform the sort of activity that will bring it face-to-face with the appropriate stimulus to release the lower level of behavior. Downward flow of energy allows appetitive behaviors to be activated in the correct sequence. Fairly recently, several computational

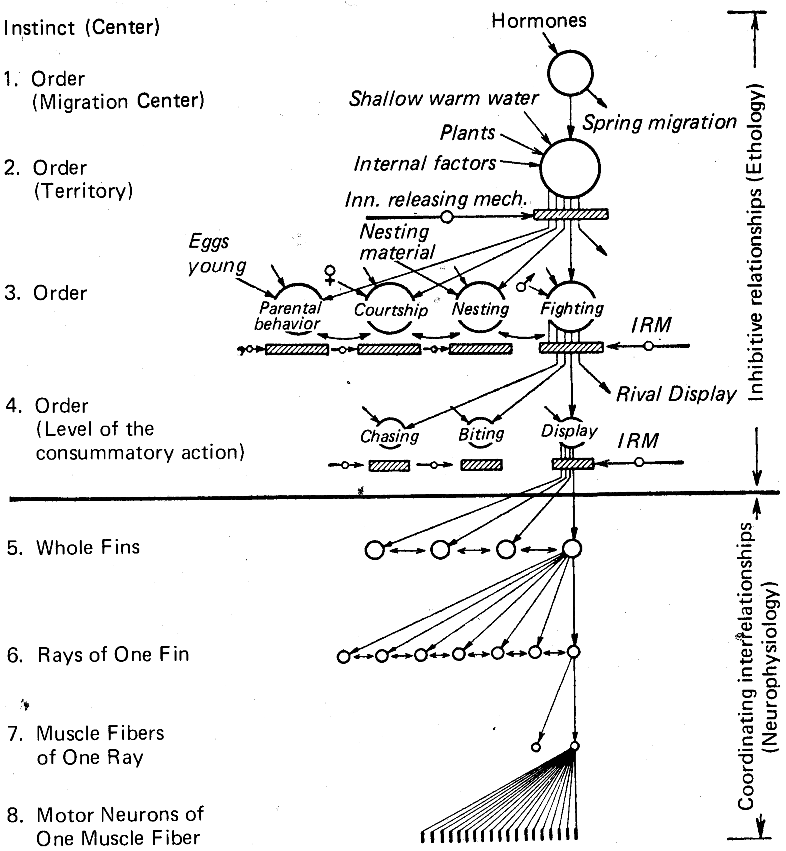


Figure 9-1: Tinbergen's proposed hierarchy to model the procreation behavior of the male stickleback fish (taken from Tinbergen (1951)). The motivational influences (hormones, etc.) operate at the top level. Behaviors of increasing specificity are modeled at deeper levels in the hierarchy. The motor responses are at the bottom.

models of behavior selection have used a similar mechanism such as Tyrrell (1994), and Blumberg (1994). Implicit in this model is that at every level of the hierarchy, a "decision" is being made among several alternatives, of which one is chosen. At the top, the decisions are very general (feed versus drink) and become increasingly more specific as one moves down a hierarchy.

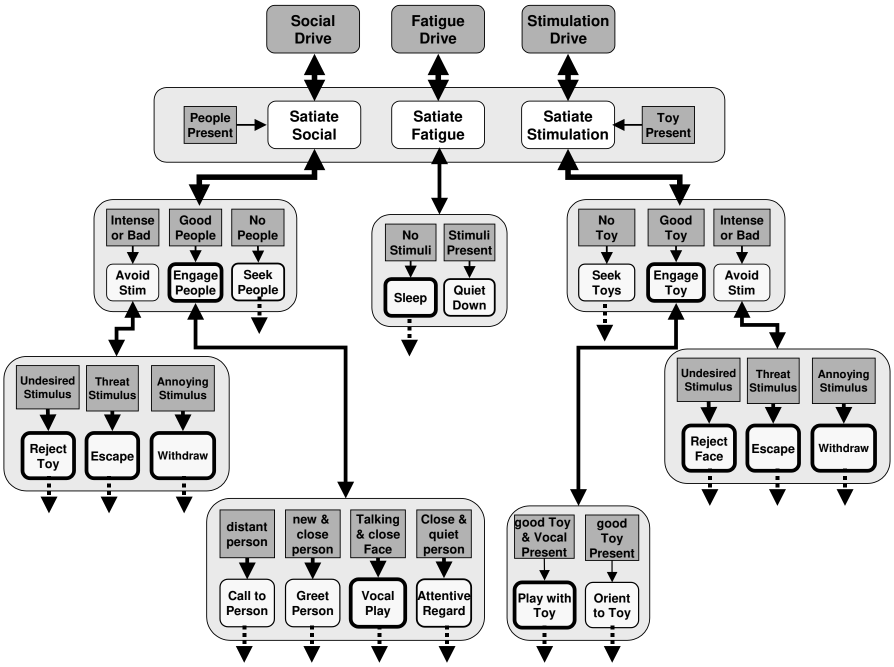


Figure 9-2: Kismet's behavior hierarchy. Bold nodes correspond to consummatory behavior(s) of the behavior group. Solid lines pass activation to other behaviors. Dashed lines send requests to the motor system. The emotional influences are not shown at this scale. See text.

9.2 Organization of Kismet's Behavior System

Following an ethological perspective and previously noted works, Kismet's behavior system organizes the robot's *goals* into a coherent structure (see figure 9-2). Each behavior is viewed as a self-interested, goal directed entity that competes with other behaviors to *establish the current task* of the robot. Given that the robot has multiple time-varying goals that it must tend to, and different behavioral strategies that it can employ to achieve them, an arbitration mechanism is required to determine which behavior(s) to activate and for how long. The main responsibility of the behavior system is to carry out this arbitration. By doing so, it addresses the issues of relevancy, coherency, concurrency, persistence, and opportunism as discussed in chapter 3.

Note, that to *perform* the behavior, the behavior system must work in concert with the motor systems (see chapters 10, 11, 13, and 12). The motor systems are responsible for figuring out how to control robot's motor modalities to carry out the stated goal of the behavior system.

The behavior system is organized into loosely layered, heterogeneous hierarchies of behavior groups (Blumberg 1994). Each group contains behaviors that compete for activation with one another. At the highest level, behaviors are organized into competing *functional groups* (the primary branches of the hierarchy) where each group is responsible for maintaining one of the three homeostatic functions (i.e., to be social, to be stimulated by the environment, and to occasionally rest). We will return to this level again in section 9.4.1.

Only one functional group can be active at a time. The influence of the robot's `drives` is strongest at the top level of the hierarchy, biasing which functional group should be active. This motivates the robot to come into contact with the satiatory stimulus for that `drive`. The intensity level of the `drive` being tended to biases behavior in a way to establish homeostatic balance. This is described in more detail in section 9.4.2.

The emotional influence on behavior activation is more direct and immediate. As discussed in chapter 8, each emotional response is mapped to a distinct behavioral response. Instead of influencing behavior only at the top level of the hierarchy (as is the case with `drives`), an active `emotion` can directly activate the behavioral response to which it maps. It accomplishes this by sending sufficient activation energy to its affiliated behavior(s) and behavior groups such that the desired behavior wins the competition among other behaviors and becomes active. In this way, an emotion can "hijack" behavior to suit its own purposes.

Each functional group consists of an organized hierarchy of behavior groups. At each level in the hierarchy, each behavior group represents a competing strategy (a collection of behaviors) for satisfying the goal of its parent behavior. In turn, each behavior within a behavior group is viewed as a task-achieving entity that pursues its particular goal in order to carry out the strategy of its behavior group. The behavior groups are akin to Tinbergen's *behavioral centers*. They are represented as *container nodes* in the hierarchy (because they "contain" the competing behaviors of that group). They are similar in spirit to the behavior groups of Blumberg's system. However, whereas Blumberg (1994) uses mutual inhibition between competing behaviors within a group to determine the winner, the container node compares the activation levels of its behaviors to determine the winner.

Each behavior group consists of a consummatory behavior and one or more appetitive behaviors. The goal of a behavior group is to activate the consummatory behavior of that group. Each appetitive behavior within the group is designed to bring the robot into a relationship with the environment so that the consummatory behavior can become active. When the consummatory behavior is carried out, the task of that behavior group is achieved. For a given appetitive behavior in the group, carrying out its task might require the performance of other more specific tasks. In this case, these more specific tasks are represented as a child behavior group of the appetitive behavior. Each child behavior group represents a different strategy for achieving the parent (Blumberg 1996).

Hence, at the behavioral category level, the functional groups compete to determine which need is to be met (socializing, playing, or sleeping). At the strategy level, those behavior groups belonging to the winning functional group compete against each other for expression. Finally, on the task level, the behaviors belonging to the winning behavior group compete for expression. Hence the observed behavior of the robot is the result of competition at the functional, strategy, and task levels.

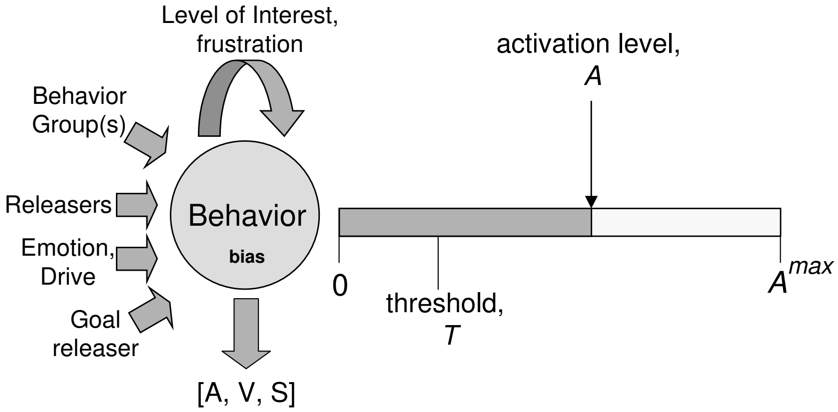


Figure 9-3: The model of a behavior. See text.

9.3 The Model of a Behavior

The individual behaviors within a group compete for activation based on their computed relevance to the given situation. Each behavior determines its own relevance by taking into account perceptual factors (as defined by its affiliated releaser and goal releaser) as well as internal factors (see figure 9-3). The internal factors can either arise from an affiliated `emotion` (or `drive` at the top level), from activity of the behavior group to which it belongs (or the child behavior group, if present), or the behavior's own internal state (such as its frustration, current level of interest, or prepotentiated bias). Hence, as was the case with the motivational system, there are many different types of factors that contribute to a behavior's relevance. These influences must be converted into a common currency so that they can be combined. The result is the activation level for the behavior. The activation level represents some measure of the behavior's "value" to the robot at that point in time.

Provided that the behavior group is active, each behavior within the group updates its level of activation by the equation:

$$A_{behavior} = max(A_{child}, \sum_n (releaser_n \cdot gain_n), A_{update}) \quad (9.1)$$

where,

$$\begin{aligned} A_{update} &= \sum_n (releaser_n \cdot gain_n) + \sum_m (motiv_m \cdot gain_m) + \\ & \quad success(\sum_k releaser_{goal,k}) \cdot (LoI - frustration) + bias \end{aligned} \quad (9.2)$$

A_{child} is the activation level of the child behavior group, if present
n is the number of releaser inputs, $releaser_n$
$gain_n$ is the weight for each contributing releaser
m is the number of motivation inputs, $motiv_m$
$gain_m$ is the weight for each contributing `drive` or `emotion`
$success()$ is a function that returns 1 if the goal has not been achieved, and 0 otherwise.
LoI is the level of interest, $LoI = LoI_{initial} - decay(LoI, gain_{decayLoI})$
$LoI_{initial}$ is the default persistence
$frustration$ increases linearly with time, $frustration = frustation + (gain_{frust} \cdot t)$
$bias$ is a constant that pre-potentiates the behavior
and $decay(x, g) = x - \frac{x}{g}$ for $g > 1$ and $x > 0$, and 0 otherwise

When the behavior group is inactive, the activation level is updated by the equation:

$$A_{behavior} = max(A_{child}, \sum_n (releaser_n \cdot gain_n), decay(A_{behavior}, gain_{decayBeh})) \quad (9.3)$$

Internal Measures

The goal of each behavior is defined as a particular relationship between the robot and its environment (a *goal releaser*). Hence, the success condition can simply be represented as another releaser for the behavior that fires when the desired relation is achieved within the appropriate behavioral and motivational context. For instance, the goal condition for the `seek-person` behavior is the presence of a "good quality" person in the visual field. The `found-person` releaser only fires when people are the desired stimulus (the `social-drive` is active), the robot is engaged in a person finding behavior, and there is a visible person (i.e., skin-tone) who is within face-to-face interaction distance of the robot and is not moving in a threatening manner (no excessive motion). Some behaviors, particularly those at the top level of the hierarchy, operate to maintain a desired internal state (keeping its drive in homeostatic balance, for instance). A releaser for this type of process measures the activation level of the affiliated drive.

The active behavior sends information to the high-level perceptual system that may be needed to contextualize the incoming perceptual features. When a behavior is active, it updates its own internal measures of success and progress to its goal. The behavior sends positive valence to the emotion system upon success of the behavior. As time passes with delayed success, an internal measure of frustration grows linearly with time. As this grows, it sends negative valence and withdrawn stance values to the emotion system (however, the arousal and stance values may vary as a function of time for some behaviors). The longer it takes the behavior to succeed, the more frustrated the robot appears. The frustration level reduces the level of interest of the behavior. Eventually, the behavior "gives up" and loses the competition to another.

Specificity of Releasers

Behaviors that are located deeper within the hierarchy are more specific. As a result, both the antecedent conditions that release the behavior, as well as the goal relations that signal success, become more specific. This establishes a hierarchy of releasers, progressing in detail from broad and general to more specific. The broadest releasers simply establish the type of stimulus ("people" verses "toys") and its presence or absence. As one moves deeper in the hierarchy, many of the releasers are the same as those that are passed to the affective tagging process in the emotion system. Hence, these releasers are not just simple combinations of perceptual features. They are contextualized according to the motivational and behavioral state of the robot (see chapter 8). They are analogous to simple "cognitions" in emotional appraisal theories because they specifically relate the perceptual features to the "well being" and goals of the robot.

Adjustment Parameters

Each behavior follows this general model. Several parameters are used to specify the distinguishing properties of each behavior. This amount of flexibility allows rich behaviors to be specified and interesting behavioral dynamics to be established.

- *Activation within a group*: One important parameter is the releaser used to elicit the behavior. This plays an important role in determining when the behavior becomes active. For instance, the absence of a desired toy stimulus is the correct condition to activate the `seek-toy` behavior. However, as discussed previously, it is not a simple one-to-one mapping from stimulus to response. Motivational factors also influence a behavior's relevance.

- *Deactivation within a group*: Another important parameter is the goal signaling releaser. This determines when an appetitive behavior has achieved its goal and can be deactivated. The consummatory behaviors remain active upon success until a motivational switch occurs that biases the robot to tend to a different need. For instance, during the `seek toy` behavior (an appetitive behavior), the behavior is successful when the `found-toy` releaser fires. This releaser is a combination of `toy-present` contexed by the `seek-toy` behavior. It fires for the short period of time between the decay of the `seek-toy` behavior and the activation of `engage-toy` (the consummatory behavior).

- *Temporal dynamics within a group*: The timing of activating and deactivating behaviors within a group is very important. The human and the robot establish a tightly coupled dynamic when in face-to-face interaction. Both are continuously adapting their behavior to the other, and the manner in which they adapt their behavior is often in direct response to the last action the partner just performed. To keep the flow of interaction smooth, the dynamics of behavioral transitions must be well matched to natural human interaction speeds. For instance, the transition from the `call-to-person` behavior to bring a distant person near, to the activation of the `greet-person` response when the person closes to face-to-face interaction distance, to the transition to the `vocal-play` behavior when the person says his/her first utterance, must occur at a pace that the human feels comfortable with. Each of these involves showing the right amount of responsiveness to the new stimulus situation, the right amount of persistence of the active behavior (the motor act must have enough time to be displayed and witnessed), and the right amount of delay before the next behavior becomes active (so that each display is presented as a "purposeful" and distinct act).

- *Temporal dynamics between levels*: A similar issue holds for the dynamics between different levels of the hierarchy. If a child behavior is successfully addressing the goal of its parent, then the parent should remain active longer to support the favorable progress of its child. For instance, if the robot is having a good interaction with a person, then the time spent doing so should be extended – rather than rigidly following a fixed schedule where the robot must switch to look for a toy after a certain amount of time. We do not want good quality interactions to be needlessly interrupted. Hence, the various needs of the robot must still be addressed in reasonable time, but the timing should be flexible and opportunistic. To accomplish this, the parent behaviors are made aware of the progress of their children. The container node of the child passes activation energy up the hierarchy to its parent, and the parent's activation is a combination of its own measure of relevance and that of its child.

- *Affective influence*: Another important set of parameters adjust how strongly the active behaviors should be allowed to influence the net affective state. The amount of valence, arousal, and stance sent to the emotion system can vary from behavior to behavior. Currently, only the leaf behaviors of the hierarchy influence the emotion system. Their magnitude and growth rate determine how quickly the robot displays frustration, how strongly it displays pleasure upon success, etc. The timing of affective expression is important, since it often occurs during the transition between different behaviors. Because these affective expressions are social cues, they must occur at the right time to signal the appropriate event that elicited the expression.

For instance, consider the period of time between successfully finding a toy during the `seek-toy` behavior, and the transition to the `engage-toy` behavior. During this time span, the `seek-toy` behavior signals its success to the emotion system by sending it a positively valenced signal. This increase in net positive valence is usually sufficient to cause `joy` to become active, and the robot smiles. The smile is a social cue to the caregiver that the robot has successfully found what it was looking for.

9.4 Implementation of the Proto-Social Responses

In the current implementation of the behavior system there are three primary branches, each specialized for addressing a different need. Each is comprised of multiple levels, with three layers being the deepest (see figure 9-2). Each level of the hierarchy serves a different function, and addresses a different set of issues. As one moves down in depth, the behaviors serve to more finely tune the relation between the robot and its environment, and in particular, the relation between the robot and the human.

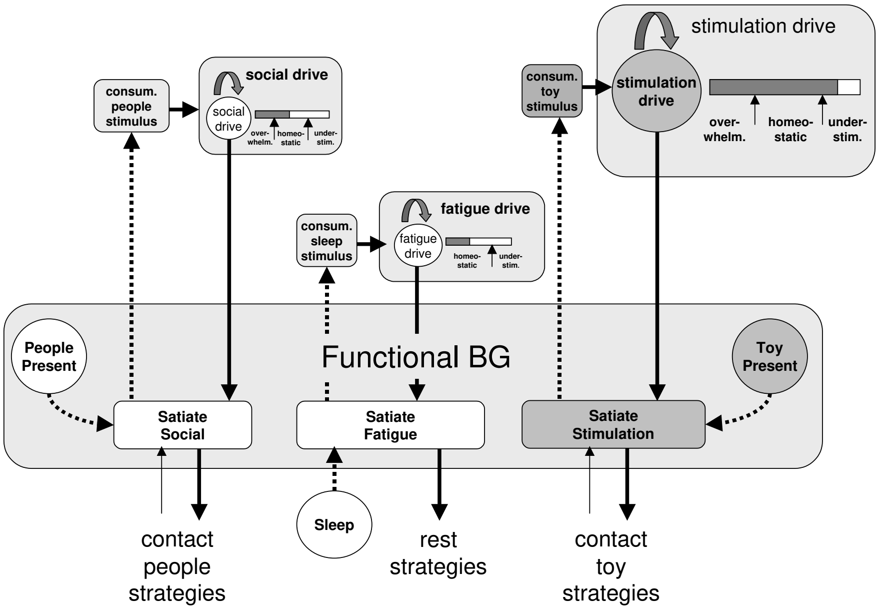


Figure 9-4: The level zero behavior group. This is the functional level that establishes which need Kismet's behavior will be directed towards satiating. Here, the `stimulation-drive` has the greatest intensity of the drives. Furthermore, its satiatory stimulus is present and the `toy-present` releaser is firing. As a result, the `satiate-stimulation` behavior is active and passes the activation from the `toy-present` releaser to satiate the drive. See text.

9.4.1 Level Zero: the Functional Level

The top level of the hierarchy consists of a single behavior group with three behaviors `satiate-social`, `satiate-stimulation`, and `satiate-fatigue` (see figure 9-4). The purpose of this group is to determine which need the robot should address. Specifically, whether the robot should engage people and satiate the `social-drive`, or engage toys and satiate the `stimulation-drive`, or rest and satiate the `fatigue-drive`.

To make this decision, each behavior receives input from its affiliated drive. The larger the magnitude of the drive, the more urgently that need must be addressed, and the greater the contribution the `drive` makes to the activation of the behavior. The `satiate-social` behavior receives input from the `people-present` releaser, and the `satiate-stimulation` behavior receives input from the `toy-present` releaser. The value of each of these releasers is proportional to the intensity of the associated stimulus (for instance, closer objects appear larger in the visual field and have a higher releaser value). The fatigue drive is somewhat different; it receives input from the activation of the `sleep` behavior.

Establish Motivational Context

The winning behavior at this level performs two functions. First, it spreads activation downward to the next level of the hierarchy. This serves to organize behavior around satisfying the affiliated drive. This establishes the motivational context that determines whether a given type of stimulus is either desirable or undesirable. A stimulus is desirable if it satiates the affiliated drive of the active behavior.

Satiate Drive

Second, the top level behaviors act to satiate their affiliated drive. Each satiates their drive when the needed stimulus is encountered and it is of good intensity (neither under-stimulating nor overwhelming). This causes the drive to move to the homeostatic regime. If the stimulus intensity is too intense, it moves the drive to the overwhelmed regime. If the stimulus is not intense enough, it moves the drive toward the under-stimulated regime. These conditions are addressed by behaviors the next level down the hierarchy.

9.4.2 Level One: The Environment Regulation Level

The behaviors at this level are responsible for establishing a good intensity of interaction with the environment. As shown in figure 9-2, `satiate-social` and `satiate-stimulation` pass activation to their behavior group below. At this level, the behavior group consists of three types of behaviors: *searching* behaviors set the current task to explore the environment and bring the robot into contact with the desired stimulus; *avoidance* behaviors set the task to move the robot away from stimuli that are too intense, undesirable, or threatening; and *engagement* behaviors set the task of interacting with desirable, good intensity stimuli.

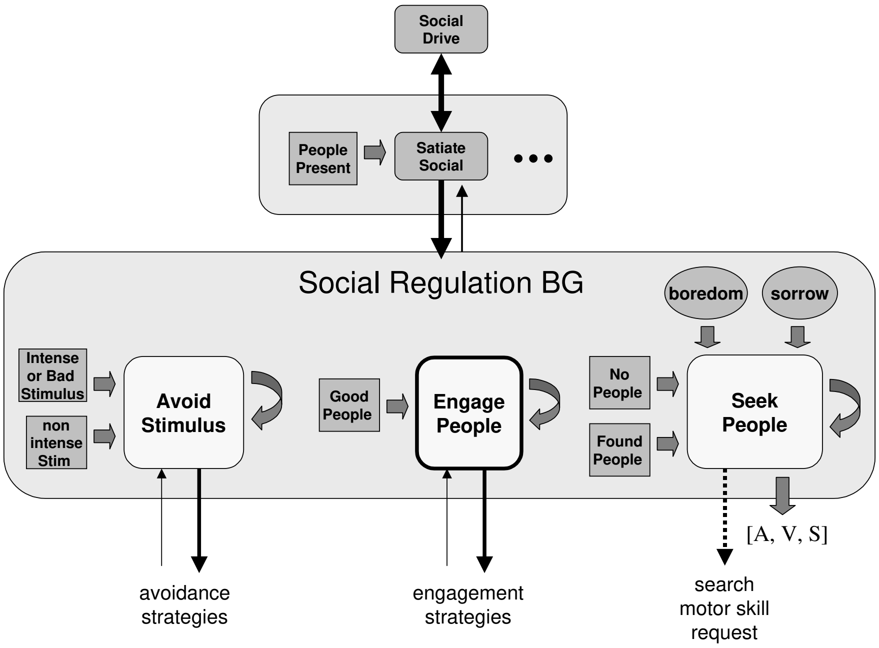


Figure 9-5: Level one behavior group. Only the social hierarchy is shown. This is the environment regulation level that establishes interactions that neither under-stimulate nor overwhelm the robot. See text.

Search Behaviors

Each search behavior establishes the goal of finding the desired stimuli. Thus, the goal of the `seek-people` behavior is to seek out skin-toned stimuli, and the goal of the `seek-toys` behavior is to seek out colorful stimuli. As described in chapter 6, when active each adjusts the gains of the attention system to facilitate these goals. Each search behavior receives contributions from releasers (signaling the current absence of the desired stimulus), or low arousal affective states (such as `boredom`, and `sorrow`) that signal a prolonged absence of the sought after stimulus.

Avoidance Behaviors

Each avoidance behavior, `avoid-stimulus` for both the social and stimulation hierarchies, establishes the goal of putting distance between the robot and the offending stimulus or event. The presence of an offensive stimulus or event contributes to the activation of an avoidance behavior through its releaser. At this level, an offending stimulus is either undesirable (not of the correct type), threatening (very close and moving fast), or annoying (too close or moving too fast to be visually tracked effectively). The behavioral response recruited to cope with the situation depends upon the nature of the offense. The coping strategy is defined within the behavior group one more level down. We discuss the specifics of this in section 9.4.3.

Engagement Behaviors

The goal of the engagement behaviors, `engage-people` or `engage-toys`, is to orient and maintain the robot's attention on the desired stimulus. These are the consummatory behaviors of the level 1 group. With the desired stimulus found, and any offensive conditions removed, the robot can engage in play behaviors with the desired stimulus. These play behaviors are described in section 9.4.4.

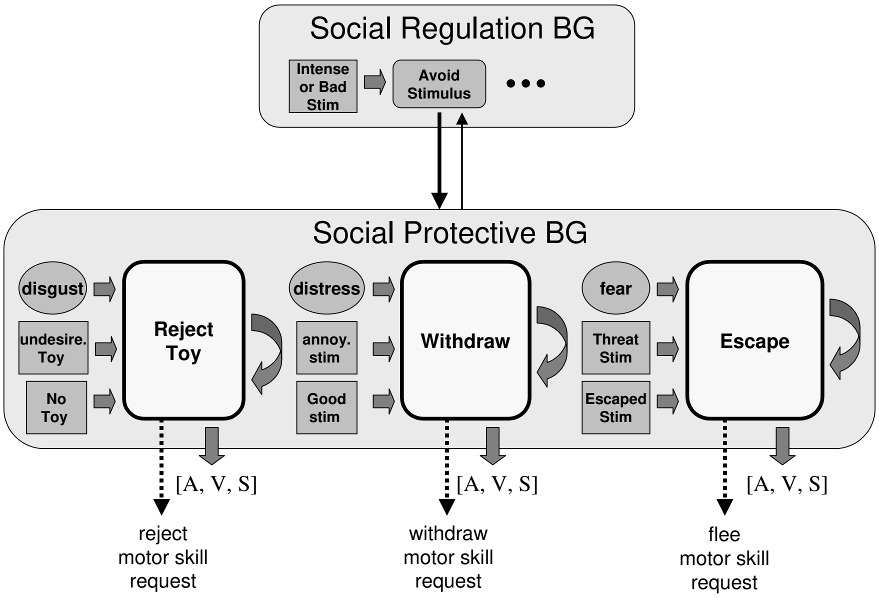


Figure 9-6: Level two protective behavior group. Only the social hierarchy is shown. This is the level two behavior group that allows the robot to avoid offensive stimuli. See text.

9.4.3 Level Two: The Protective Behaviors

As shown in figure 9-6, there are three types of protective behaviors that co-exist within the `protective` behavior group. Each represents a different coping strategy that is responsible for handling a particular kind of offense. Each coping strategy receives contributions from its affiliated releaser as well as from its affiliated emotion process.

Escape Behavior

When active, the goal set by the `escape` behavior is to flee from the offending stimulus. This behavior sends a request to the motor system to perform the fleeing response where the robot closes its eyes, grimaces, and turns its head away from a threatening stimulus. It doesn't matter whether this stimulus is skin-toned or colorful – if anything is very close and moving fast, then it is interpreted as a threat by the low-level visual perception system. There is a dedicated releaser, `threat-stimulus`, that fires whenever a threatening stimulus is encountered. This releaser passes activation to the `escape` behavior as well as to the emotion system. When `fear` is active, it elicits a fearful expression on the robot's face of the appropriate intensity (see chapters 8 and 11). This expression is a social signal that gives advance warning of any behavioral response that may ensue. If the activation level of `fear` is strong enough, it sends sufficient activation to the `escape` behavior to win the competition. The robot then performs the escape maneuver.

Withdraw Behavior

The `withdraw` behavior is active when the robot finds itself in an unpleasant, but not threatening situation. Often this corresponds to a situation where the robot's visual processing abilities are over challenged. For instance, if a person is too close to the robot, the eye-detector has difficulty locating the person's eyes. Alternatively, if a person is waving a toy too fast to be tracked effectively, the excessive amount of motion is classified as "annoying" by the low level visual processes. Either of these conditions will cause the `annoy-stim` releaser to fire. The releaser sends activation energy to the `withdraw` behavior as well as to the emotion system. This causes the `distress` process to become active. Once active, the robot's face exhibits an annoyed appearance. `Distress` also sends sufficient activation to activate the `withdraw` behavior, and a request is made of the motor system to back away from the offending stimulus. The primary function of this response is to send a social cue to the human that they are offending the robot and thereby encourage the person to modify their behavior.

Reject Behavior

The `reject` behavior is active when the robot is being offered an undesirable stimulus. The affiliated emotion process is `disgust`. It is similar to the situation where an infant will not accept the food it is offered. It has nothing to do with the offered stimulus being noxious, it is simply not what the robot is after.

9.4.4 Level Two: The Play Behaviors

Kismet exhibits different play patterns when engaging toys versus people. Kismet will readily track and occasionally vocalize while its attention is drawn to a colorful toy, but it will not evoke its repertoire of envelope displays that characterize `vocal play`. These proto-dialog behaviors are reserved for interactions with people. These social cues are not exhibited when playing with toys. The difference in the manner Kismet interacts with people versus toys provides observable evidence that these two categories of stimuli are distinguished by Kismet.

In this section we focus our discussion on those four behaviors within the `social-play` behavior group. This behavior group encapsulates Kismet's engagement strategies for establishing proto-dialogs during face-to-face exchanges. They finely tune the relation between the robot and the human to support interactive games at a level where both partners perform well.

Calling Behavior

The first engagement task is the `call-to-person` behavior. This behavior is relevant when a person is in view of the robot but too far for face-to-face exchange. The goal of the behavior is to lure the person into face-to-face interaction range (ideally, about three feet from the robot). To accomplish this, Kismet sends a social cue, the `calling` display, directed to the person within calling range.

The releaser affiliated with this behavior combines skin-tone with proximity measures. It fires when the person is four to seven feet from the robot. The actual calling display is covered in detail in chapter 11. It is evoked when the `call-to-person` behavior is active and makes a request to the motor system to exhibit the display. The human observer sees the robot orient towards him/her, crane its neck forward, wiggle its ears with large amplitude movements, and vocalize excitedly. The display is designed to attract a person's attention. The robot then resumes a neutral posture, perks its ears, and raises its brows in an expectant manner. It waits in this posture for a while, giving the person time to approach before the calling sequence resumes. The `call-to-person` behavior will continue to request the display from the motor system until it is either successful and becomes deactivated, or it becomes irrelevant.

Greeting Behavior

The second task is the `greet-person` behavior. This behavior is relevant when the person has just entered into face-to-face interaction range. It is also relevant if the `social-play` behavior group has just become active and a person is already within face-to-face range. The goal of the behavior is to socially acknowledge the human and to initiate a close interaction. When active, it makes a request of the motor system to perform the greeting display. The display involves making eye contact with the person and smiling at them while waving the ears gently. It often immediately follows the success of the `call-to-person` behavior. It is a transient response, only issued once, as its completion signals the success of this behavior.

Attentive Regard Behavior

The third task is `attentive-regard`. This behavior is active when the person has already established a good face-to-face interaction distance with the robot but remains silent. The goal of the behavior is to visually attend to the person and to appear open to interaction. To accomplish this, it sends a request to the motor system to hold gaze on the person, ideally looking into the person's eyes if the eye detector can locate them. The robot watches the person intently and vocalizes occasionally. If the person does speak, this behavior loses the competition to the `vocal-play` behavior.

Turn-taking Behavior

The forth task is `vocal-play`. The goal of this behavior is to carry out a proto-dialog with the person. It is relevant when the person is within face-to-face interaction distance and has spoken. To perform this task successfully, the `vocal-play` behavior must closely regulate turn-taking with the human. This involves a close interaction with the perceptual system to perceive the relevant turn-taking cues from the person (i.e., that a person is present and whether or not there is speech occurring), and with the motor system to send the relevant turn-taking cues back to the person.

There are four turn-taking phases this behavior must recognize and respond to. Each state is recognized using distinct perceptual cues, and each phase involves making specific display requests of the motor system.

- *Relinquish speaking turn*: This phase is entered immediately after the robot finishes speaking. The robot relinquishes its turn by craning its neck forward, raising its brows, and making eye-contact (in adult humans, shifting gaze direction is sufficient, but we exaggerated the display for Kismet to increase its readability). It holds its gaze on the person throughout this phase. However, due to noise in the visual system, in practice the eyes tend to flit about the person's face, perhaps even leaving it briefly and then returning soon afterwards. This display signals that the robot has finished speaking and is waiting for the human to say something. It will time out after a few seconds (approx. 8 seconds) if the person does not respond. At this point, the robot reacquires its turn and issues another vocalization in an attempt to reinitiate the dialog.

- *Attend to human's speech*: Once the perceptual system acknowledges that the human has started speaking, the robot's ears perk. This little feedback cue signals that the robot is listening to the person speak. The robot looks generally attentive to the person and continues to maintain eye contact if possible.

- *Reacquire speaking turn*: This phase is entered when the perceptual system acknowledges that the person's speech has ended. The robot signals that it is about to speak by leaning back to a neutral posture and averting its gaze. The robot is likely to blink its eyes as it shifts posture.

- *Deliver speech*: Soon after the robot shifts its posture back to neutral, the robot vocalizes. The utterances are short babbles, generated by the vocalization system (presented in chapter 12). Sometimes more than one is issued. The eyes migrate back to the person's face, to their eyes if possible. Just before the robot is prepared to finish this phase, it is likely to blink. The behavior transitions back to the relinquish turn phase and the cycle resumes.

The system is designed to maintain social exchanges with a person for about twenty minutes; at this point the other drives typically begin to dominate the robot's motivation. When this occurs, the robot begins to behave in a fussy manner – the robot becomes more distracted by other things around it, and it makes fussy faces more frequently. It is more difficult to engage in proto-dialog. Overall, it is a significant change in behavior. People seem to readily sense the change and try to vary the interaction, often by introducing a toy. The smile that appears on the robot's face, and the level of attention that it pays to the toy, are strong cues that the robot is now involved in satiating its `stimulation drive`.

9.5 Experiments and Analysis

The behavior system implements the four classes of proto-social responses. The robot displays affective responses by changing emotive facial expressions in response to stimulus quality and internal state. These expressions relate to goal achievement, emotive reactions, and reflections of the robot's state of "well being". The exploratory responses include visual search for desired stimuli, orientation, and maintenance of mutual regard. Kismet has a variety of protective responses that serve to distance the robot from offending stimuli. Finally, the robot has a variety of regulatory responses bias the caregiver to provide the appropriate level and kinds of interactions at the appropriate times. These are communicated to the caregiver through carefully timed social displays as well as affective facial expressions. The organization of the behavior system addresses the issues of relevancy, coherency, persistence, flexibility, and opportunism. The proto-social responses address the issues of believability, promoting empathy, expressiveness, and conveying intentionality.

9.5.1 Regulating Interaction

Figure 9-7 shows Kismet responding to a toy with these four response types. The robot begins the trial looking for a toy and displaying sadness (an affective response). The robot immediately begins to move its eyes searching for a colorful toy stimulus (an exploratory response) ($t < 10$). When the caregiver presents a toy ($t \approx 13$), the robot engages in a play behavior and the stimulation drive becomes satiated ($t \approx 20$). As the caregiver moves the toy back and forth ($20 < t < 35$), the robot moves its eyes and neck to maintain the toy within its field of view. When the stimulation becomes excessive ($t \approx 35$), the robot becomes first displeased and then fearful as the stimulation drive moves into the overwhelmed regime. After extreme over-stimulation, a protective escape response produces a large neck movement ($t = 38$) which removes the toy from the field of view. Once the stimulus has been removed, the stimulation drive begins to drift back to the homeostatic regime (one of the many regulatory responses in this example).

9.5.2 Interaction Dynamics

The behavior system produces interaction dynamics that are similar to the five phases of infant social interactions (initiation, mutual-orientation, greeting, play-dialog, and disengagement) discussed in section 9.0.1. These dynamic phases are not explicitly represented in the behavior system, but emerge from the interaction of the synthetic

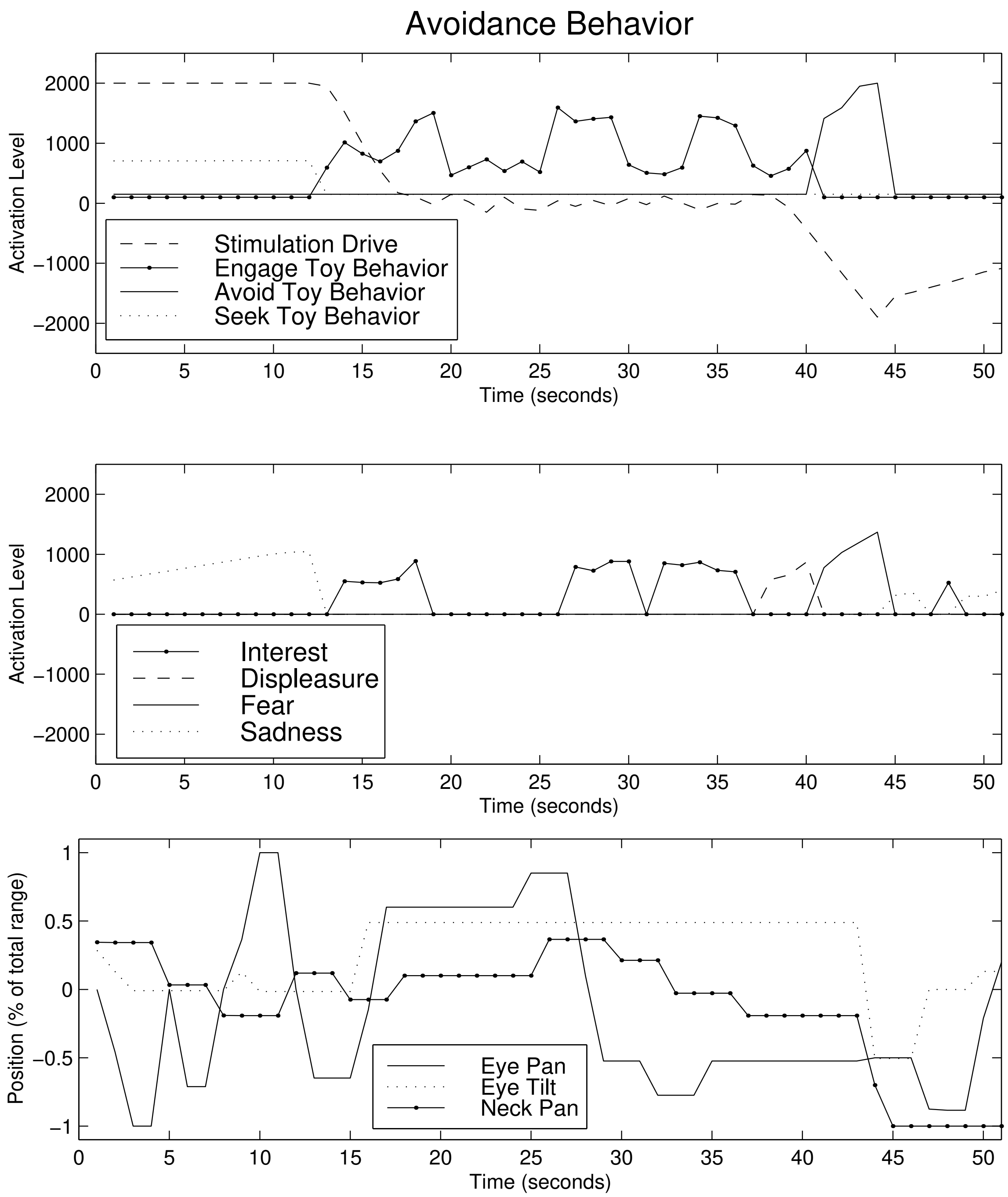


Figure 9-7: Kismet's response to excessive stimulation. Behaviors and drives (top), emotions (middle), and motor output (bottom) are plotted for a single trial of approximately 50 seconds. See text for description.

nervous system with the environment. By producing behaviors that convey intentionality, we exploit the caregivers natural tendencies to treat the robot as a social creature, and thus to respond in characteristic ways to the robot's overtures. This reliance on the external world produces dynamic behavior that is both flexible and

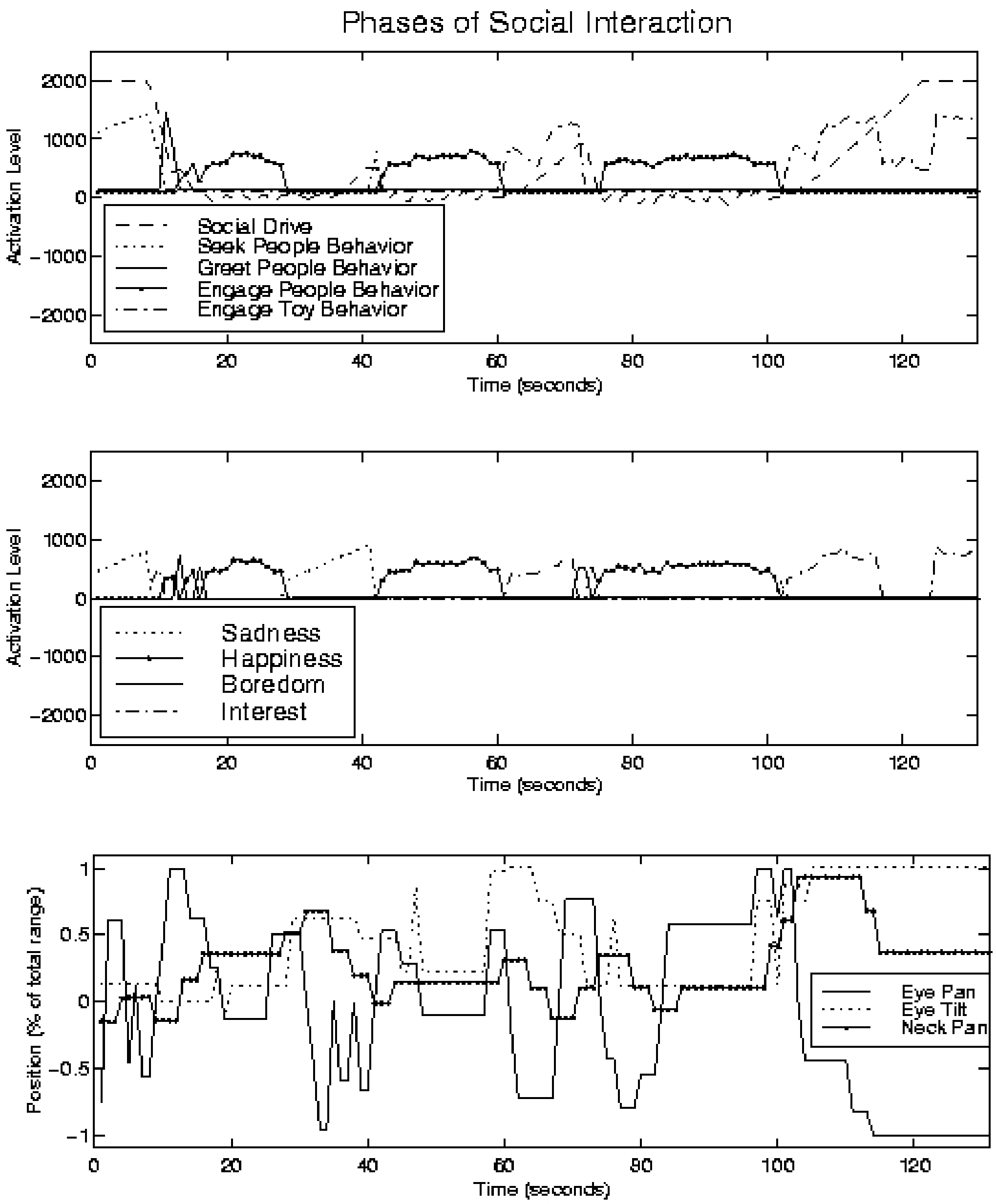


Figure 9-8: Cyclic responses during social interaction. Behaviors and drives (top), emotions (middle), and motor output (bottom) are plotted for a single trial of approximately 130 seconds. See text for description.

robust.

Figure 9-8 shows Kismet's dynamic responses during face-to-face interaction with a caregiver. Kismet is initially looking for a person and displaying sadness (the initiation phase). The sad expression evokes nurturing responses from the caregiver. The robot begins moving its eyes looking for a face stimulus ($t < 8$). When it finds the caregiver's face, it makes a large eye movement to enter into mutual regard ($t \approx 10$).

Once the face is foveated, the robot displays a greeting behavior by wiggling its ears ($t \approx 11$), and begins a play-dialog phase of interaction with the caregiver ($t > 12$). Kismet continues to engage the caregiver until the caregiver moves outside the field of view ($t \approx 28$). Kismet quickly becomes sad, and begins to search for a face, which it re-acquires when the caregiver returns ($t \approx 42$). Eventually, the robot habituates to the interaction with the caregiver and begins to attend to a toy that the caregiver has provided ($60 < t < 75$). While interacting with the toy, the robot displays interest and moves its eyes to follow the moving toy. Kismet soon habituates to this stimulus, and returns to its play-dialog with the caregiver ($75 < t < 100$). A final disengagement phase occurs ($t \approx 100$) when the robot's attention shifts back to the toy.

Regulating Vocal Exchanges

Kismet employs different social cues to regulate the rate of vocal exchanges. These include both eye movements as well as postural and facial displays. These cues encourage the subjects to slow down and shorten their speech. This benefits the auditory processing capabilities of the robot.

To investigate Kismet's performance in engaging people in proto-dialogs, we invited three naive subjects to interact with Kismet. They ranged in age from 25 to 28 years of age. There was one male and two females, all professionals. They were simply asked to talk to the robot. Their interactions were video recorded for further analysis.

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

Before the subjects adapt their behavior to the robot's capabilities, the robot is more likely to interrupt them. There tend to be more frequent delays in the flow of "conversation" where the human prompts the robot again for a response. Often these "hiccups" in the flow appear in short clusters of mutual interruptions and pauses (often over 2 to 4 speaking turns) before the turns become coordinated and the flow smoothes out. However, by analyzing the video of these human-robot "conversations", there is evidence that people entrain to the robot (see figure 9-9). These "hiccps" become less frequent. The human and robot are able to carry on longer sequences of clean turn transitions. At this point the rate of vocal exchange is well matched to the

		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+

Figure 9-9: Data illustrating evidence for entrainment of human to robot.

robot's perceptual limitations. The vocal exchange is reasonably fluid. Table 9-10 shows that the robot is engaged in a smooth proto-dialog with the human partner the majority of the time (about 82%).

9.6 Limitations and Extensions

Kismet can engage a human in compelling social interaction, both with toys and during face-to-face exchange. People seem to interpret Kismet's emotive responses quite naturally and adjust their behavior so that it is suitable for the robot. Furthermore, people seem to entrain to the robot by reading its turn-taking cues. The resulting interaction dynamics are reminiscent of infant-caregiver exchanges. However, there are number of ways in which we could improve the system.

The robot does not currently have the ability to interrupt itself. This will be an important ability for more sophisticated exchanges. When watching video of people talking with Kismet, they are quite resilient to hiccups in the flow of "conversation". If they begin to say something just before the robot, they will immediately pause once the robot starts speaking and wait for the robot to finish. It would be nice if Kismet could exhibit the same courtesy. The robot's babbles are quite short at the

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

Figure 9-10: Kismet's turn taking performance during proto-dialog with three naive subjects. Significant disturbances are small clusters of pauses and interruptions between Kismet and the subject until turn-taking become coordinated again.

moment, so this is not a serious issue yet. But as the utterances become longer, it will become more important.

It is also important for the robot to understand where the human's attention is directed. At the very least, the robot should have a robust way of measuring when a person is addressing it. Currently the robot assumes that if a person is near by, then that person is attending to the robot. The robot also assumes that it is the most salient person who is addressing it. Clearly this is not always the case. This is painfully evident when two people try to talk to the robot and to each other. It would be a tremendous improvement to the current implementation if the robot would only respond when a person addressed it directly (instead of addressing someone else), and if the robot responded to the correct person (instead of the most salient person). Sound localization using the stereo microphones on the ears could help identify the source of the speech signal. This information could also be correlated with visual input to direct the robot's gaze. In general, determining where a person is looking is a computationally difficult problem (Newman & Zelinsky 1998) (Scassellati 1999).

The latency in Kismet's verbal turn-taking behavior needs to be reduced. For humans, the average time for a verbal reply is about 250 ms. For Kismet, its verbal response time varies from 500 ms to 1500 ms. Much of this depends on the length of the person's previous utterance, and the time it takes the robot to shift between turn-taking postures. In the current implementation, the `in-speech` flag is set when the person begins speaking, and is cleared when the person finishes. There is a delay of about 500 ms built into the speech recognition system from the end of speech to accommodate pauses between phrases. Additional delays are related to the length of the spoken utterance – the longer the utterance the more computation is required before the output is produced. To alleviate awkward pauses and to give people immediate feedback that the robot heard them, the ear-perk response is triggered by the `sound-flag`. This flag is sent immediately whenever the speech recognizer receives input (speech or non-speech sounds). Delays are also introduced as the robot shifts posture between taking its turn and relinquishing the floor. However, this also sends important social cues and enlivens the exchange. In watching the video, the turn-taking pace is certainly slower than for conversing adults, but given the lively posturing and facial animation, it appears engaging. The naive subjects readily adapted to this pace and did not seem to find it awkward. However, to scale the performance to adult human performance, the goal of a 250 ms delay between speaking turns should be achieved.

9.7 Summary

Drawing strong inspiration from ethology, the behavior system arbitrates among competing behaviors to address issues of relevance, coherency, flexibility, robustness, persistence, and opportunism. This enables Kismet to behave in a complex, dynamic world. However to socially engage a human, its behavior must address issues of believability – such as conveying intentionality, promoting empathy, being expressive, and displaying enough variability to appear unscripted while remaining consistent. To accomplish this, we have implemented a wide assortment of proto-social responses of human infants. These responses encourage the human caregiver to treat the robot as a young, socially aware creature. Particular attention has been paid to those behaviors that allow the robot to actively engage a human, to call to people if they are too far away, and to carry out proto-dialogs with them when they are near by. The robot employs turn taking cues that humans use to entrain to the robot. As a result, the proto-dialogs become smoother over time. The general dynamics of the exchange share structural similarity with those of three-month old infants with their caregivers. All five phases (initiation, mutual regard, greeting, play dialog, and disengagement) can be observed.

Chapter 10

Overview of the Motor Systems

Whereas the behavior system is responsible for deciding which task the robot should perform at any time, the motor system is responsible for figuring out how to drive the motors in order to carry out the task. In addition, whereas the motivation system is responsible for establishing the affective state of the robot, the motor system is responsible for commanding the actuators in order to convey that emotional state.

There are four distinct motor systems that carry out these functions for Kismet. The *vocalization system* produces expressive babbles that allow the robot to engage humans in proto-dialog. The *face motor system* orchestrates the robot's emotive facial expressions and body posture, its facial displays that serve communicative social functions, those that serve behavioral functions (such as "sleeping"), and lip synchronization with accompanying facial animation. The *oculo-motor system* produces human-like eye movements and head orientations that serve important sensing as well as social functions. Finally, the *motor skills system* coordinates each of these specialized motor systems to produce coherent multi-modal motor acts.

10.1 Levels of Interaction

Kismet's rich motor behavior can be conceptualized on four different levels (as shown in figure 10-1). These levels correspond to the *social level*, the *behavior level*, the *skills level*, and the *primitives level*. This decomposition is motivated by distinct temporal, perceptual, and interaction constraints that exist at each level.

Temporal Constraints

The temporal constraints pertain to how fast the motor acts must be updated and executed. These can range from real-time vision rates (33 frames/sec) to the relatively slow time scale of social interaction (potentially transitioning over minutes).

Perceptual Feedback Constraints

The perceptual constraints pertain to what level of sensory feedback is required to coordinate behavior at that layer. This perceptual feedback can originate from the low level visual processes such as the current target from the attention system, to relatively high-level multi-modal percepts generated by the behavioral releasers.

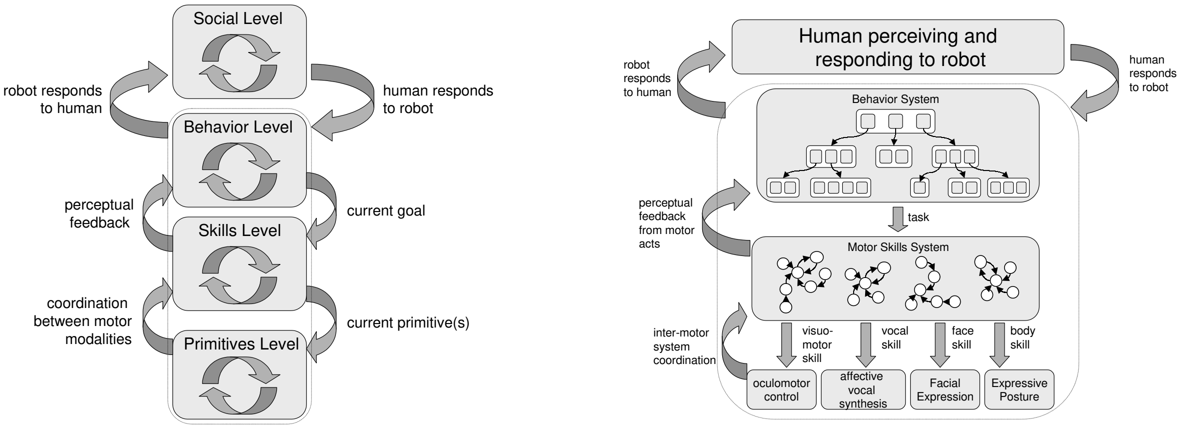


Figure 10-1: Levels of behavioral organization. The primitive level is populated with tightly coupled sensori-motor loops. The skill level contains modules that coordinate primitives to achieve tasks. Behavior level modules deal with questions of relevance, persistence and opportunism in the arbitration of tasks. The social level comprises design-time considerations of how the robot's behaviors will be interpreted and responded to in a social environment.

Interaction Constraints

The interaction constraints pertain to the arbitration of units that compose each layer. This can range from low-level oculo-motor primitives (such as saccades and smooth pursuit), to using visual behavior to regulate turn-taking.

10.1.1 Issues at Each Level

Each level serves a particular purpose for generating the overall observed behavior. As such, each level must address a specific set of issues. The levels of abstraction help simplify the overall control of behavior by restricting each level to address those core issues that are best managed at that level. By doing so, the coordination of behavior at each level (i.e., arbitration), between the levels (i.e., top-down and bottom-up), and through the world is maintained in a principled way.

The Social Level

The social level explicitly deals with issues pertaining to having a human in the interaction loop. This requires careful consideration of how the human interprets and responds to the robot's behavior in a social context. For instance, using visual behavior (making eye contact and breaking eye contact) to help regulate the transition of speaker turns during vocal turn-taking is an example. We presented this in chapter 9. Chapter 7 discussed examples with respect to affect-based interactions during emotive vocal exchanges. Chapter 13 discusses the relationship between animate visual behavior and social interaction. A summary of these findings is presented in chapter 14.

The behavior level

The behavior level deals with issues related to producing relevant, appropriately persistent, and opportunistic behavior. This involves arbitrating between the many possible goal-achieving behaviors that Kismet could perform to establish the current task. Actively seeking out a desired stimulus and then visually engaging it is an example. Other behavior examples are described in chapter 9.

The motor skill level

The motor skill level is responsible for figuring out how to move the motors to accomplish the task specified by the behavior system. Fundamentally, this level deals with the issues of blending of and sequencing between coordinated ensembles of motor primitives (each ensemble is a distinct motor skill). The skills level must also deal with coordinating multi-modal motor skills (e.g., those motor skills that combine speech, facial expression, and body posture). Kismet's searching behavior is an example where the robot alternately performs ballistic eye-neck orientation movements with gaze fixation to the most salient target. The ballistic movements are important for scanning the scene, and the fixation periods are important for locking on the desired type of stimulus. We elaborate upon this system at the end of this chapter.

The motor primitives level

The motor primitives level implements the building blocks of motor action. This level must deal with motor resource allocation and tightly coupled sensori-motor loops. Kismet actually has three distinct motor systems at the primitives level: the *expressive vocal system* (see chapter 12), the *facial animation system* (see chapter 11), the *oculo-motor system* (see chapter 13). Aspects of controlling the robot's body posture are described in chapters 11 and 13).

10.2 The Motor Skills System

Given the current task (as dictated by the behavior system, chapter 9), the *motor skills system* is responsible for figuring out how to carry out the stated goal. Often this requires coordination between multiple motor modalities (speech, body posture, facial display, and gaze control). Requests for these modalities can originate from the top down (e.g. from the emotion system or behavior system), as well as from the bottom-up (the vocal system requesting lip and jaw movements for lip synchronizing). Hence, the motor skills level must address the issue of servicing the motor requests of different systems across the different motor resources.

The motor skills system must deal with the issues of appropriately *blending* the motor actions of concurrently active behaviors. Sometimes concurrent behaviors require completely different sets of actuators (such as babbling while watching a stimulus). In this case there is no direct competition over a shared resource, so the motor skills system should command the actuators to execute both behaviors simultaneously. Other times, two concurrently active behaviors may compete for the same actuators. For instance, the robot may have to smoothly track a moving object while maintaining vergence. These two behaviors are complementary in that each can be carried out without the sacrifice or degradation in the performance of the other. However, the motor skills system must coordinate the motor commands to do so appropriately.

The motor skills system is also responsible for *smoothly transitioning* between sequentially active behaviors. For instance, to initiate a social exchange, the robot must first mutually orient to the caregiver and then exchange a greeting with her before the social exchange can commence. Once started, Kismet may take turns with the caregiver in exchanging vocalizations, facial expressions, etc. After a while, either party can disengage from the other (such as looking away), thereby terminating the interaction. While sequencing between these behaviors, the motor system must figure out how to transition smoothly between them in a timely manner so as to not disrupt the natural flow of the interaction.

Finally, the motor skills system is responsible for moving the robot's actuators to convey the appropriate emotional state of the robot. This may involve performing facial expressions, or adapting the robot's posture. Of course, this affective state must be conveyed while carrying out the active task(s). This is a special case of blending mentioned above, which may or may not compete for the same actuators. For instance, looking at an unpleasant stimulus may be performed by directing the eyes to the stimulus, but orienting the face away from the stimulus and configuring the face into a "disgusted" look.

10.2.1 Motor Skill Mechanisms

It often requires a sequence of coordinated motor movements to satisfy a goal. Each motor movement is a primitive (or a combination of primitives) from one of the base motor systems (the vocal system, the oculo-motor system, etc.). Each of these coordinated series of motor primitives is called a *skill*, and each skill is implemented as a finite state machine (FSM). Each motor skill encodes knowledge of how to move from one motor state to the next, where each sequence is designed to bring the robot closer to the current goal. The motor skills level must arbitrate among the many different FSMs, selecting the one to become active based on the active goal. This decision process is straightforward since there is an FSM tailored for each task of the behavior system.

Many skills can be thought of as *fixed action patterns* (FAPs) as conceptualized by early ethologists (Tinbergen 1951), (Lorenz 1973). Each FAP consists of two components, the *action* component and the *taxis* (or orienting) component. For Kismet, FAPs often correspond to communicative gestures where the action component corresponds to the facial gesture, and the taxis component (to whom the gesture is directed) is controlled by gaze. People seem to intuitively understand that when Kismet makes eye contact with them, they are the locus of Kismet's attention and

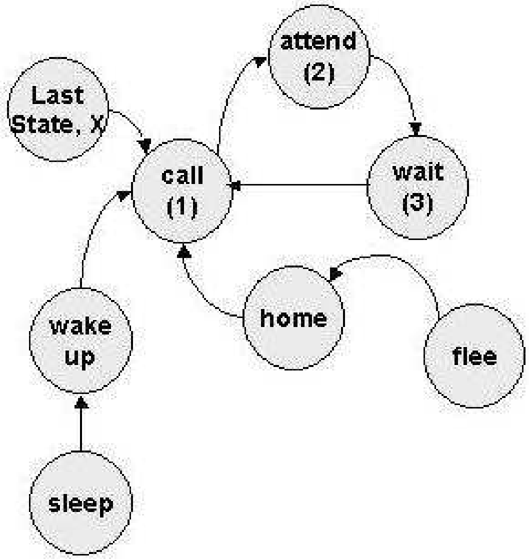


Figure 10-2: The calling motorskill. The states 1, 2, and 3 are described in the text. The remaining states encode knowledge of how to transition from any previously active motor skill state to the `call` state.

the robot's behavior is organized about them. This places the person in a state of action readiness where they are poised to respond to Kismet's gestures.

A classic example of a motor skill is Kismet's *calling* FAP (see figure 10-2). When the current task is to bring a person into a good interaction distance, the motor skill system activates the `calling` FSM. The taxis component of the FAP issues a `hold gaze` request to the oculo-motor system. This serves to maintain the robot's gaze on the person to be hailed. In the first state (1) of the gesture component, Kismet leans its body toward the person (a request to the body posture motor system). This strengthens the person's perception that the robot has taken a particular interest in them. The ears also begin to waggle exuberantly (creating a significant amount of motion and noise) which further attracts the person's attention to the robot. In addition, Kismet vocalizes excitedly which is perceived as an initiation. The FSM transitions to the second state (2) upon the completion of this gesture. In this state, the robot "sits back" and waits for a bit with an expecting expression (ears slightly perked, eyes slightly widened, and brows raised). If the person has not already approached the robot, it is likely to occur during this "anticipation" phase. If the person does not approach within the allotted time period, the FSM transitions to the third state (3) where face relaxes, the robot maintains a neutral posture, and gaze fixation is released. At this point, the robot is able to shift gaze. As long as this FSM is active (determined by the behavior system), the calling cycle repeats. It can be interrupted at any state transition by the activation of another FSM (such as the `greeting` FSM when the person has approached). Chapter 11 presents a table of FAPs that have been implemented on Kismet. A summary of Kismet's FAPs is presented in chapter

11.

10.3 Summary

Kismet's motor behavior is conceptualized, modeled, and implemented on multiple levels. Each level is a layer of abstraction with distinct timing, sensing, and interaction characteristics. Each layer is implemented with a distinct set of mechanisms that addresses these factors. The motor skills system coordinates the primitives of each specialized system for facial animation, body posture, expressive vocalization, and oculo-motor control. We describe each of these specialized motor systems in detail in the following chapters.

Chapter 11

Facial Animation and Expression

The human face is the most complex and versatile of all species (Darwin 1872). For humans, the face is a rich and versatile instrument serving many different functions. It serves as a window to display one's own motivational state. This makes one's behavior more predictable and understandable to others and improves communication (Ekman, Friesen & Ellsworth 1982). The face can be used to *supplement* verbal communication. A quick facial display can reveal the speaker's attitude about the information being conveyed. Alternatively, the face can be used to *complement* verbal communication, such as lifting of the eyebrows to lend additional emphasis to a stressed word (Cassell 1999*b*). Facial gestures can communicate information on their own, such as a facial shrug to express "I don't know" to another's query. The face can serve a regulatory function to modulate the pace of verbal exchange by providing turn-taking cues (Cassell & Thorisson 1999). The face serves biological functions as well – closing one's eyes to protect them from a threatening stimulus, and on a longer time scale to sleep (Redican 1982).

11.1 Design Issues

Kismet doesn't engage in adult-level discourse, but its face serves many of these functions at a simpler, pre-linguistic level. Consequently, the robot's facial behavior is fairly complex. It must balance these many functions in a timely, coherent, and appropriate manner. Below, we outline a set of design issues for the control of Kismet's face.

Real-time Response

Kismet's face must respond at *interactive rates.* It must respond in a timely manner to the person who engages it as well to other events in the environment. This promotes readability of the robot, so the person can reliably connect the facial reaction to the event that elicited it. Real-time response is particularly important for sending expressive cues to regulate social dynamics. Excessive latencies disrupt the flow of the interaction.

Coherence

Kismet has 15 facial actuators, many of which are required for any single emotive expression, behavioral display, or communicative gesture. There must be *coherence* in how these motor ensembles move together, and how they sequence between other motor ensembles. Sometimes Kismet's facial behaviors require moving multiple degrees of freedom to a fixed posture, sometimes the facial behavior is an animated gesture, and sometimes it is a combination of both. If the face loses coherence, the information it contains is lost to the human observer.

Synchrony

The face is one expressive modality that must work in concert with vocal expression and body posture. Requests for these motor modalities can arise from multiple sources in the synthetic nervous system. Hence, *synchrony* is an important issue. This is of particular importance for lip synchronization where the phonemes spoken during a vocal utterance must be matched by the corresponding lip postures.

Expressive Versatility

Kismet's face currently supports four different functions. It reflects the state of the robot's emotion system. We call these *emotive expressions*. It conveys social cues during social interactions with people. We call these *expressive facial displays*. It synchronizes with the robot's speech, and it participates in behavioral responses. The face system must be quite *versatile* as the manner in which these four functions are manifest changes dynamically with motivational state and environmental factors.

Readable

Kismet's face must convey information in a manner as similar to humans as possible. If done sufficiently well, then naive subjects should be able to read Kismet's facial expressions and displays without requiring special training. This fosters *natural and intuitive* interaction between Kismet and the people who interact with it.

Believable

As with much of Kismet's design, there is a delicate *balance between complexity and simplicity*. Enforcing levels of abstraction in the control hierarchy with clean interfaces is important for promoting scalability and real-time response. The design of Kismet's face also strives to maintain a balance. It is quite obviously a caricature of a human face (minus the ears!), and therefore cannot do many of the things that human faces do. However, by taking this approach, we lower people's expectations for realism to a level that is achievable without detracting from the quality of interaction. As argued in chapter 4, a realistic face would set very high expectations for human-level behavior. Trying to achieve this level of realism is a tremendous engineering challenge currently being attempted by others (Hara 1998). However, it is not necessary for our purposes, which is to focus on natural social interaction.

11.2 Levels of Control

The face motor system consists of six subsystems organized into four layers of control. As presented in chapter 10, the face motor system communicates with the motor skill system to coordinate over different motor modalities (voice, body, and eyes). An overview of the face control hierarchy is shown in figure 11-1. Each layer represents a level of abstraction with its own interfaces for communicating with the other levels. The highest layers control ensembles of facial features and are organized by facial function (emotive expression, lip synchronization, facial display). The lowest layer controls the individual degrees of freedom. Enforcing these levels of abstraction keeps the system modular, scalable, and responsive.

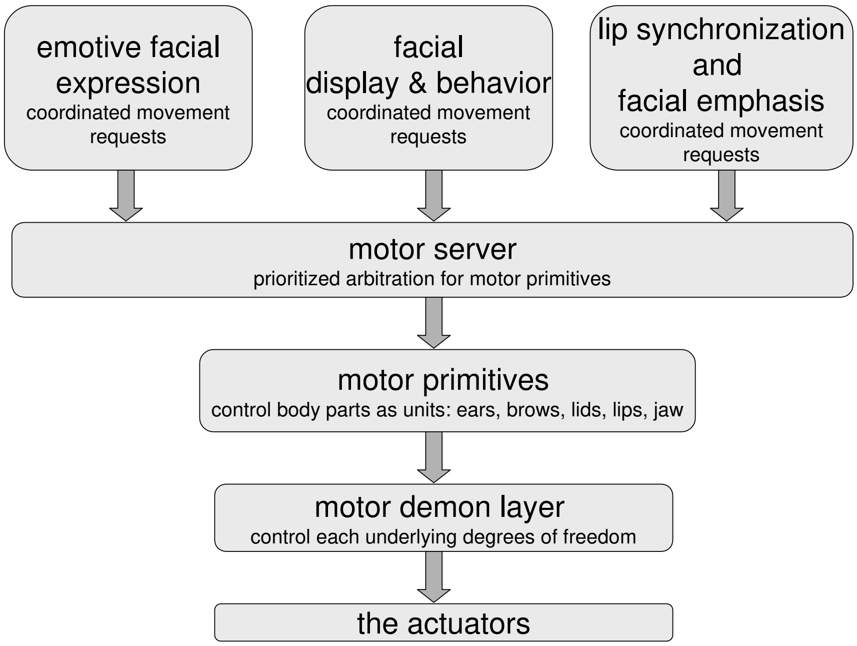


Figure 11-1: Levels of abstraction for facial control.

11.2.1 The Motor Demon Layer

The lowest level is called the *motor demon* layer. It is organized by individual actuators and implements the interface to access the underlying hardware. It initializes the maximum, minimum, and reference positions of each actuator and places safety caps on them. A common reference frame is established for all the degrees of freedom so that values of the same sign command all actuators in a consistent direction. The interface allows other processes to set the position and velocity targets of each actuator. These values are updated in a tight loop 30 times a second. Once these values are updated, the target requests are converted into a pulse-width-modulated control signal. Each is then sent through the *TPU* lines of the 68332 to drive the 14 futaba servo motors. In the case of the jaw these values are scaled and passed on to QNX where the MEI motion controller card servos the jaw.

11.2.2 The Motor Primitives Layer

The next level up is the *motor primitives* layer. Here the interface groups the underlying actuators by facial feature. Each motor primitive controls a separate body part (such as an ear, a brow, an eyelid, the upper lip, the lower lip, or the jaw). Higher-level processes make position and velocity requests of each facial feature in terms of their observed movement (as opposed to their underlying mechanical implementation). For instance, the left ear motor primitive converts requests to control elevation, rotation, and speed to the underlying differentially geared motor ensemble. The interface supports both postural movements (go to a specified position) as well as rhythmic movements (oscillate for a number of repetitions with a given speed, amplitude, and period). The interface implements a second set of primitives for small groups of facial features that often move together (such as wiggling both ears, or knitting both brows, or blinking both lids.) These are simply constructed from those primitives controlling each individual facial feature.

11.2.3 The Motor Server Layer

The motor server layer arbitrates the requests coming from the *facial expression* subsystem, the *facial display* subsystem, or the *lip synchronization* subsystem. Requests originating from these three subsystems involve moving ensembles of facial features in a coordinated manner. These requests are often made concurrently. Hence, this layer is responsible for blending and or sequencing these incoming requests so that the observed behavior is coherent and synchronized with the other motor modalities (voice, eyes, and head).

In some cases, there is blending across orthogonal sets of facial features when subsystems serving different facial functions control different groups of facial features. For instance, when issuing a verbal greeting the lip synchronization process controls the lips and jaw while the facial display subsystem wiggles the ears. However, often there is blending across the same set of facial features by different subsystems. For instance, when vocalizing in a “sad” affective state, the control for lip synchronization with facial emphasis competes for the same facial features needed to convey sadness. Here blending must take place to maintain a consistent expression of affective state.

Figure 11-2 illustrates how the facial feature arbitration is implemented. It is a priority-based scheme, where higher-level subsystems bid for each facial feature that they want to control. The bids are broken down into each observable movement of the facial feature. Hence, instead of bidding for the left ear as a whole, separate bids are made for left ear elevation and left ear rotation. To promote coherency, the bids for each component movement of a facial feature by a given subsystem are generally set to be the same. However, the flexibility is present to have different subsystems control them independently should it be appropriate to do so. The highest bid wins the

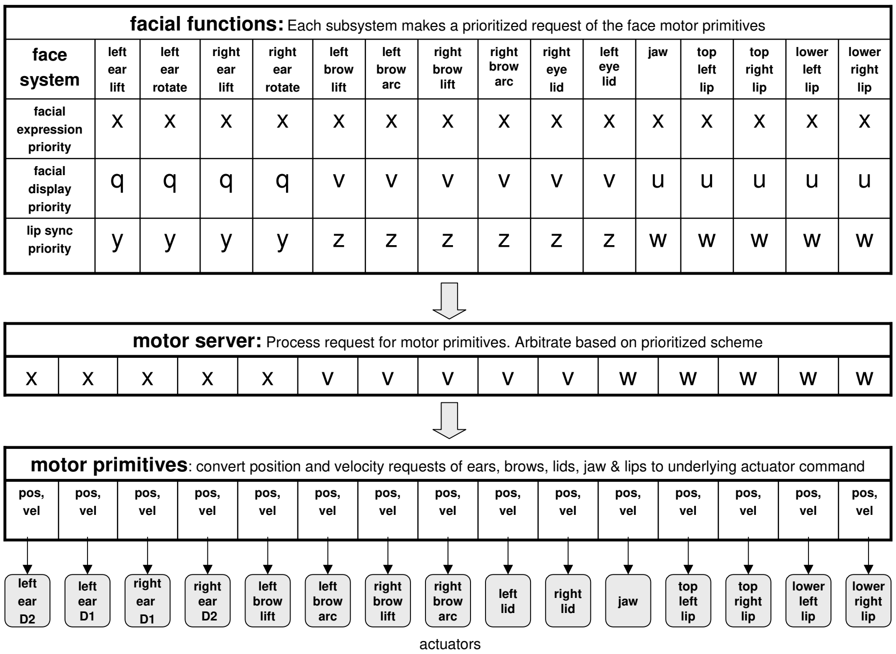


Figure 11-2: Face arbitration is handled through a dynamic priority scheme. In the figure, q, u, v, w, x, y, z are hand-coded priorities. These are updated whenever a new request is made to a face motor subsystem that serves a particular function. The actuators belonging to each type of facial feature are given the same priority so that they serve the same function. Hence, the priorities for the eyebrow motors, the lip motors, and the ear motors are updated together. At the motor server level, the largest priorities get control of those motors. In this example, the ears shall serve the expression function, the eyebrows shall serve the display function, and the lips shall serve the lip synchronization function.

competition and gets to forward its request to the underlying facial feature primitive. The request includes the target position, velocity, and type of movement (postural or rhythmic).

The priorities are defined by hand, although the bid for each facial feature changes dynamically depending on the current motor skill. There are general rules of thumb that are followed. For a low to moderate emotive intensity level, the emotive facial expression subsystem sets the expression baseline and has the lowest priority. It is always active when no other facial function is to be performed. The emotive baseline can be over-ridden by “voluntary” movements (e.g., facial gestures) as well as behavioral responses (such as “sleeping”). However, if an emotional response is evoked (due to a highly active emotion process), the emotive facial expression will be given a higher priority so that it will be expressed. The lip synchronization subsystem has the highest priority over the lips and mouth whenever a request to speak has been made. Thus, whenever the robot says something, the lips and jaw coordinate with the vocal modality. The facial emphasis component of lip synchronization modulates the facial features about the established baseline. In this way, the rest of the face blends with the underlying facial expression. This is critical for having face, voice, and body all convey a similar emotional state.

11.2.4 The Facial Function Layer

The highest level of the face control hierarchy consists of three subsystems: *emotive facial expression*, communicative *facial display and behavior*, and *lip synchronization and facial emphasis*. Each subsystem serves a different facial function. The *emotive facial expression* subsystem is responsible for generating expressions that convey the robot's current motivational state. We cover this system in detail in this chapter. Lip synchronization and facial emphasis is covered in chapter 12. The control of facial displays and behavior are covered here and in chapter 10.

Lip Synchronization and Facial Emphasis Subsystem

The lip synchronization and facial emphasis system is responsible for coordinating lips, jaw, and the rest of the face with speech. The lips are synchronized with the spoken phonemes as the rest of the face lends coordinated emphasis. See chapter 12 for the details of how Kismet's lip synchronization and facial emphasis system is implemented.

Facial Display and Behavior Subsystem

The facial display and behavior subsystem is responsible for postural displays of the face (such as raising the brows at the end of a speaking turn), animated facial gestures (such as exuberantly wiggling the ears in an attention grabbing display), and behavioral responses (such as flinching in response to a threatening stimulus). Taken as a whole, the facial display system encompasses all those facial behaviors not directly generated by the emotional system. Currently, they are modeled as simple routines that are evoked by the motor skills system (as presented in chapter 10) for a specified amount of time and then released (see figure 11-3). The motor skills system handles the coordination of these facial displays with vocal, postural, and gaze/orientation behavior. Ultimately, this subsystem might include learned movements that could be acquired during imitative facial games with the caregiver.

Emotive Facial Expression Subsystem

The emotive facial expression subsystem is responsible for generating a facial expression that mirrors the robot's current affective state. This is an important communication signal for the robot. It lends richness to social interactions with humans and increases their level of engagement. For the remainder of this chapter, we describe the implementation of this system in detail. We also discuss how affective postural shifts complement the facial expressions and lend strength to the overall expression. The

stereotyped display	description
sleep & wake-up display	Associated with the behavioral response of going to "sleep" and "waking up".
grimace & flinch display	Associated the fear response. The eyes close, the ears cover and are lowered, the mouth frowns. Is evoked in conjunction with the *flee* behavioral response.
calling display	Associated with the *calling* behavior. It is a stereotyped movement designed to get a person's attention and to approach the robot. The ears waggle exuberantly (causing significant noise), the lips have a slight smile. It is evoked with a forward postural shift and head/eye orientation to the person. If the eye-detector can find the eyes, the robot makes eye contact with the person. The robot also vocalizes with an aroused affect. The desired impression is for the targeted person to interpret the display as the robot calling to them.
greet display	A stereotyped response involving a smile and small waggling of the ears.
raise brows display	A social cue used to signal the end of the robot's turn in vocal proto-dialog. It is used whenever the robot should look expectant to prompt the human to respond. If the eyes are found, the robot makes eye contact with the person.
perk ears reflex	A social feedback cue whenever the robot hears any sound. It is used as a little acknowledgement that the robot heard the person say something.
blink reflex	A social cue often used when the robot has finished its "speaking" turn. It is often accompanied by a gaze shift away from the listener.
startle reflex	A reflex in response to a looming stimulus. The mouth opens, the lips are rounded, the ears perk, the eyes widen, and the eyebrows elevate.

Figure 11-3: A summary of Kismet's facial displays.

expressions are analyzed and their readability evaluated by subjects with minimal to no prior familiarity with the robot.

11.3 Generation of Facial Expressions

There have been only a few expressive autonomous robots (Velasquez 1998), (Fujita & Kageyama 1997) and a few expressive humanoid faces (Hara 1998), (Takanobu et al. 1998). The majority of these robots are only capable of a limited set of fixed expressions (a single happy expression, a single sad expression, etc.). This hinders both the believability and readability of their behavior. *Believability* refers to how life-like the behavior appears. *Readability* refers to how well the observer can correctly interpret the intended expression. The expressive behavior of many robotic faces is not life-like because of their discrete, mechanical, and reflexive quality – transitioning between expressions like a switch being thrown. This discreteness and discontinuity of transitions limits the readability of the face. It lacks important cues for the intensity of the underlying affective state. It also lacks important cues for the transition dynamics between affective states.

11.3.1 Insights from Animation

Classical and computer animators have a tremendous appreciation for the challenge in creating believable and readable behavior. They also appreciate the role that expressiveness plays in this endeavor. A number of animation guidelines and techniques have been developed for achieving life-like, believable, and compelling animation (Thomas & Johnston 1981), (Parke & Waters 1996). These rules of thumb explicitly consider audience perception. The rules are designed to create behavior that is rich and interesting, yet easily understandable to the human observer. Because Kismet interacts with humans, the robot's expressive behavior must cater to the perceptual needs of the human observer. This improves the quality of social interaction because the observer feels that she understands the robot's behavior. This helps her to better predict the robot's responses to her, and in turn to shape her own responses to the robot.

Of particular importance is timing: how to sequence and how to transition between actions. A cardinal rule of timing is to *do one thing at a time*. This allows the observer to witness and interpret each action. It is also important that each action last for a sufficiently long time span for the observer to read it. Given these two guidelines, Kismet expresses only one emotion at a time, and each expression has a minimum persistence of several seconds before it decays. The time of intense expression can be extended if the corresponding emotion continues to be highly active.

The transitions between expressive behaviors should be smooth. The build up and decay of expressive behavior can occur at different rates, but it should not be discontinuous like throwing a switch. Animators interpolate between target frames for this purpose, while controlling the morphing rate from the initial posture to the final posture. The physics of Kismet's motors does the smoothing for us to some extent, but the velocities and accelerations between postures are important. An aroused robot will exhibit quick movements of larger amplitude. A subdued robot will move more sluggishly. The accelerations and decelerations into these target postures must also be considered. Robots are often controlled for speed and accuracy - to achieve the fastest response time possible with minimal overshoot. Biological systems don't move like this. For this reason, Kismet's target postures as well as the velocities and accelerations that achieve them are carefully considered.

Animators take a lot of care in drawing the audience's attention to the part of the scene where an important action is about to take place. By doing so, the audience's attention is directed to the right place at the right time so that they do not miss out on important information. To enhance the readability and understandability of Kismet's behavior, its direction of gaze and facial expression serve this purpose. People naturally tend look at what Kismet is also looking at. They observe the expression on its face to see how the robot is affectively assessing the stimulus. This is a predictor of the robot's behavior. If the robot looks at a stimulus with an interested expression, the observer predicts that the robot will continue to engage the stimulus. Alternatively, if the robot has a frightened expression, the observer is

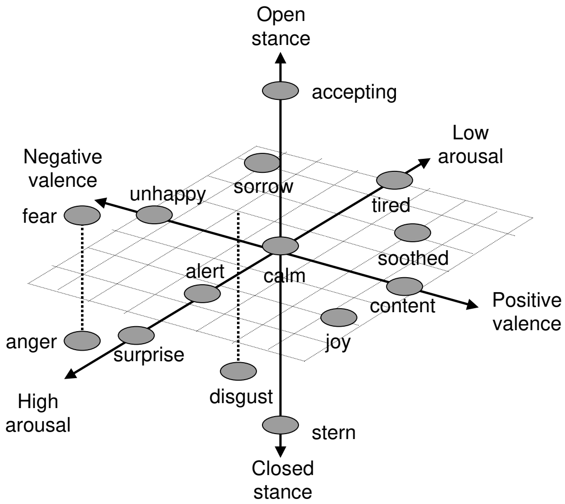


Figure 11-4: The affect space consists of three dimensions. The extremes are: high arousal, low arousal, positive valence, negative valence, open stance, and closed stance. The emotional processes can be mapped to this space.

not surprised to witness a fleeing response soon afterwards. Kismet's expression and gaze precede the behavioral response to make it understandable and predictable to the human who interacts with it.

Expression is not just conveyed through face, but through the entire body. In general, Kismet's expressive shifts in posture may modify the motor commands of more task-based motor skills (such as orienting toward a particular object). Consequently, we must address the issue of expressive blending with neck and eye motors. To accomplish this, the affective state determines the default posture of the robot, and the task based motor commands are treated as offsets from this posture. To add more complexity, the robot's level of arousal sets the velocities and accelerations of the task-based movements. This causes the robot to move sluggishly when arousal is low, and to move in a darting manner when in a high arousal state.

11.3.2 Generating Emotive Expression

Kismet's facial expressions are generated using an interpolation-based technique over a three dimensional space (see figure 11-4). The three dimensions correspond to *arousal*, *valence*, and *stance*. Recall in chapter 8, the same three attributes are used to affectively assess the myriad of environmental and internal factors that contribute to Kismet's affective state. We call the space defined by the $[A, V, S]$ trio the *affect space*. The current affective state occupies a single point in this space at a time. As the robot's affective state changes, this point moves about within this space. Note that this space not only maps to emotional states (i.e., anger, fear, sadness, etc.) but also to the level of arousal as well (i.e., excitement and fatigue). A range of

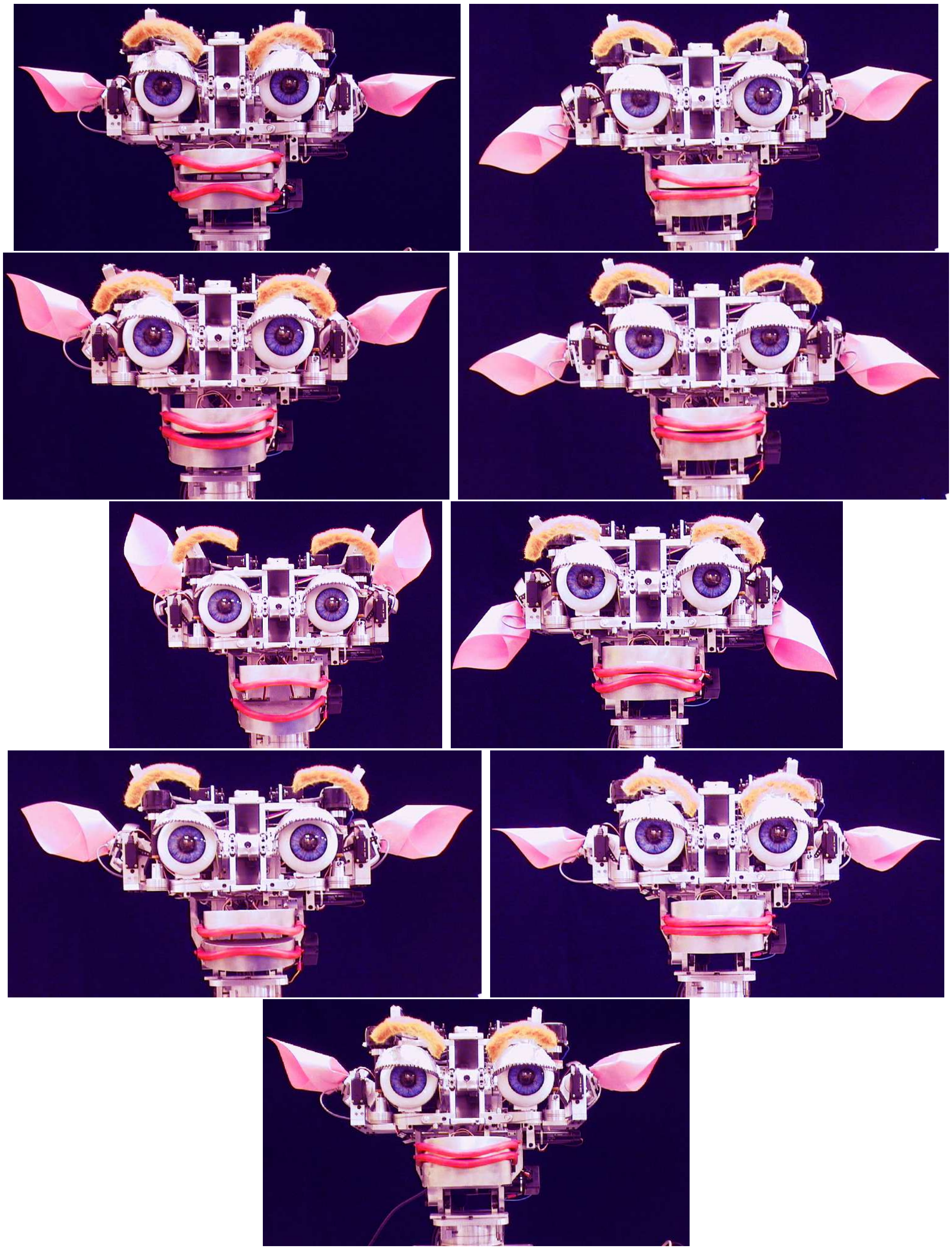


Figure 11-5: Kismet is capable of generating a continuous range of expressions of various intensities by blending the basis facial postures. Facial movements correspond to affect dimensions in a principled way. A sampling is shown here. These correspond to the nine test stimuli of the human sketch comparison experiment.

expressions generated with this technique is shown in figure 11-5. The procedure runs in real-time, which is critical for social interaction.

The affect space can be roughly partitioned into regions that map to each emotion process (see figure 11-4). The mapping is defined to be coarse at first, and the emotion system is initially configured so that only limited regions of the overall space are frequented often. The intention was to support the possibility of emotional and expressive development, where the emotion processes continue to refine as secondary emotions are acquired through experience, and associated with particular regions in affect space with their corresponding facial expressions.

There are nine *basis* (or *prototype*) postures that collectively span this space of emotive expressions. Although some of these postures adjust specific facial features more strongly than the others, each prototype influences most if not all of the facial features to some degree. For instance, the valence prototypes have the strongest influence on lip curvature, but can also adjust the positions of the ears, eyelids, eyebrows, and jaw. The basis set of facial postures has been designed so that a specific location in affect space specifies the relative contributions of the prototype postures in order to produce a net facial expression that faithfully corresponds to the active emotion. With this scheme, Kismet displays expressions that intuitively map to the emotions of anger, disgust, fear, happiness, sorrow, and surprise. Different levels of arousal can be expressed as well from interest, to calm, to weariness.

There are several advantages to generating the robot's facial expression from this affect space. First, this technique allows the robot's facial expression to reflect the nuance of the underlying assessment. Hence, even through there is a discrete number of emotion processes, the expressive behavior spans a continuous space. Second, it lends clarity to the facial expression since the robot can only be in a single affective state at a time (by our choice), and hence can only express a single state at a time. Third, the robot's internal dynamics are designed to promote smooth trajectories through affect space. This gives the observer a lot of information as to how the robot's affective state is changing, which makes the robot's facial behavior more interesting. Furthermore, by having the face mirror this trajectory, the observer has immediate feedback as to how their behavior is influencing the robot's internal state. For instance, if the robot has a distressed expression upon its face, it may prompt the observer to speak in a soothing manner to Kismet. The soothing speech is assimilated into the emotion system where it causes a smooth decrease in the arousal dimension and a push toward slightly positive valence. Thus, as the person speaks in a comforting manner, it is possible to witness a smooth transition to a subdued expression. However, if the face appeared to grow more aroused, then the person may stop trying to comfort the robot verbally and perhaps try to please the robot by showing it a colorful toy.

The *primary* six prototype postures sit at the extremes of each dimension (see figure 11-6). They correspond to high arousal, low arousal, negative valence, positive valence, open (approaching) stance, and closed (withdrawing) stance. The high arousal prototype, P_{high}, maps to the expression for surprise. The low arousal prototype, P_{low}, corresponds to the expression for fatigue (note that sleep is a behavioral response, so it is covered in the facial display subsystem). The positive valence prototype, $P_{positive}$, maps to a content expression. The negative valence prototype, $P_{negative}$, resembles an unhappy expression. The closed stance prototype, P_{closed}, resembles a stern expression, and the open stance prototype, P_{open}, resembles an accepting expression.

The three affect dimensions also map to affective postures. There are six prototype postures defined which span the space. High arousal corresponds to an erect posture with a slight upward chin. Low arousal corresponds to a slouching posture where the neck lean and head tilt are lowered. The posture remains neutral over the valence dimension. An open stance corresponds to a forward lean movement, which suggests strong interest toward the stimuli the robot is leaning towards. A closed stance corresponds to withdraw, reminiscent of shrinking away from whatever the robot is looking at. In contrast to the facial expressions which are continually expressed, the affective postures are only expressed when the corresponding emotion process has sufficiently strong activity. When expressed, the posture is held for a minimum period of time so that the observer can read it, and then it is released. The facial expression, of course, remains active. The posture is presented for strong conveyance of a particular affective state.

The remaining three facial prototypes are used to strongly distinguish the expressions for disgust, anger, and fear. Recall that four of the six primary emotions are characterized by negative valence. Whereas the primary six basis postures (presented above) can generate a range of negative expressions from distress to sadness, the expressions for intense anger (rage), intense fear (terror), and intense disgust have some uniquely distinguishing features. For instance, the prototype for disgust, $P_{disgust}$, is unique in its asymmetry (which is typical of this expression). The prototypes for anger, P_{anger}, and fear, P_{fear}, each have a distinct configuration for the lips (furious lips form a snarl, terrified lips form a grimace).

Each dimension of the affect space is bounded by the minimum and maximum allowable values of $(min, max) = (-1250, 1250)$. The placement of the prototype postures is given in table 11-6. The current net affective assessment from the emotion system defines the $[A, V, S] = (a, v, s)$ point in affect space. The specific (a, v, s) values are used to *weight* the relative *motor* contributions of the basis postures. Using a weighted interpolation scheme, the net emotive expression, P_{net}, is computed. The contributions are computed as follows:

$$P_{net} = C_{arousal} + C_{valence} + C_{stance} \tag{11.1}$$

where:
P_{net} is the emotive expression computed by weighted interpolation,
$C_{arousal}$ is the weighted motor contribution due to the arousal state,
$C_{valence}$ is the weighted motor contribution due to the valence state,
C_{stance} is the weighted motor contribution due to stance state.

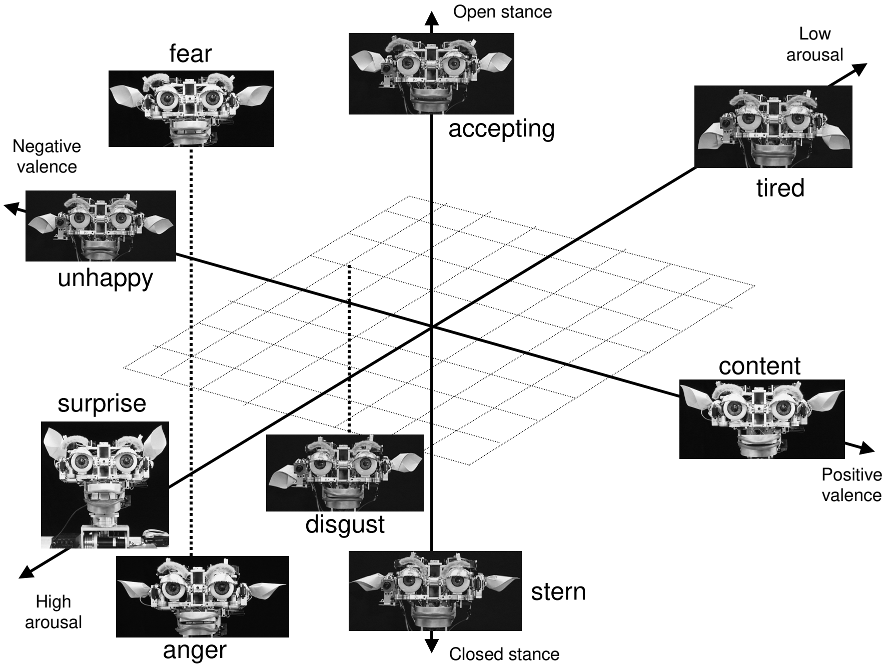


Figure 11-6: This diagram illustrates where the basis postures are located in affect space.

These contributions are specified by the equations:

$C_{arousal} = \alpha P_{high} + (1-\alpha)P_{low}$
$C_{valence} = \beta P_{positive} + (1-\beta)P_{negative}$
$C_{stance} = \gamma P_{open} + (1-\gamma)P_{closed} + F(a, v, s, n, \gamma)$

where $\alpha, 0 \leq \alpha \leq 1$ is the fractional interpolation coefficient for arousal, $\beta, 0 \leq \beta \leq 1$ is the fractional interpolation coefficient for valence, and $\gamma, 0 \leq \gamma \leq 1$ is the fractional interpolation coefficient for stance.

The function:
$F(A, V, S, N, \delta) = f(A, V, S, N, \delta) \cdot P_{anger} + f(A, V, S, N, (1-\delta)) \cdot P_{fear} +$
$f(A, V, S, N, (1-\delta)) \cdot P_{disgust}$

The function $f(A, V, S, N, \delta)$ limits the influence of each specialized prototype posture to remain local to their region of affect space. Recall, there are three specialized postures, P_i for anger, fear, and disgust. Each is located at $(A_{P_i}, V_{P_i}, S_{P_i})$ where A_{P_i} corresponds to the arousal coordinate for posture P_i, V_{P_i} corresponds to the valence coordinate, and S_{P_i} corresponds to the stance coordinate. Given the current net affective state (a, v, s) as computed by the emotion system, one can compute the displacement from (a, v, s) to each $(A_{P_i}, V_{p_i}, S_{p_i})$. As this distance increases from a given posture P_i, that posture contributes less and less to the net emotive expression.

Hence, for each P_i, the function $f(A, V, S, N, \delta)$ is a weighting function that decays linearly with distance from $(A_{P_i}, V_{p_i}, S_{p_i})$. The weight is bounded between $0 \leq f \leq 1$, where the maximum value occurs at $(A_{P_i}, V_{p_i}, S_{p_i})$. The argument N defines the radius of influence (which is kept fairly small).

11.3.3 Comparison to Componential Approaches

It is interesting to note the similarity of this scheme with the affect dimensions viewpoint of emotion (Russell 1997), (Smith & Scott 1997). Instead of viewing emotions in terms of categories (happiness, anger, fear, etc.), this viewpoint conceptualizes the dimensions that could span the relationship between different emotions (arousal and valence, for instance). Instead of taking a production-based approach to facial expression (how do emotions generate facial expressions), Russell (1997) takes a perceptual stance (what information can an observer read from a facial expression). For the purposes of Kismet, this perspective makes a lot of sense given our concern with the issue of readability and understandability.

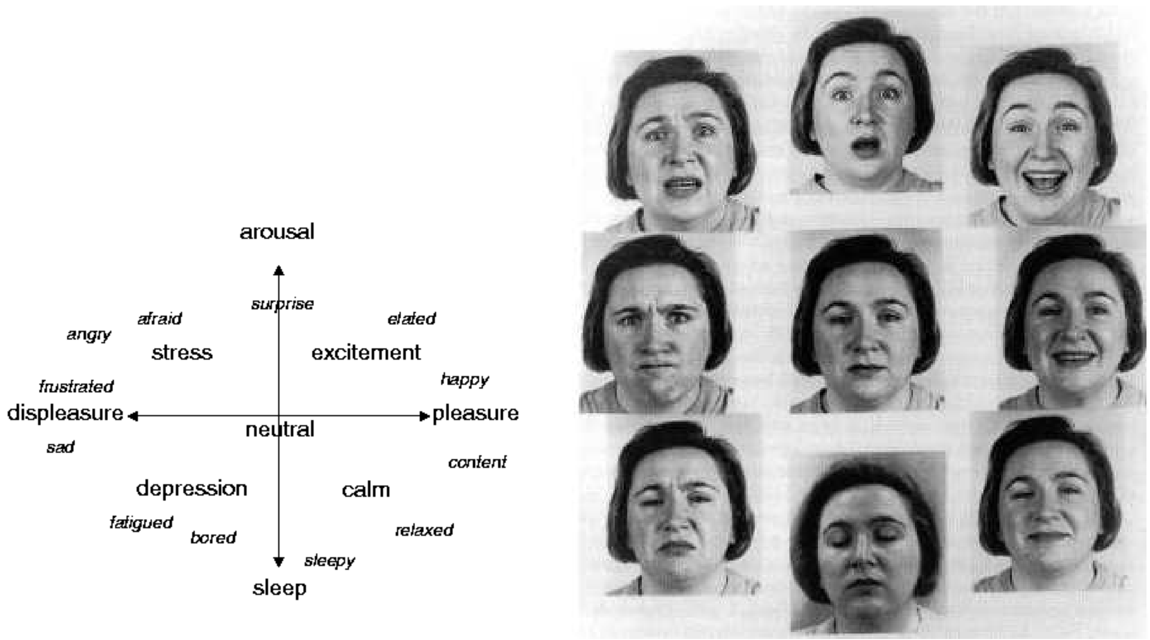


Figure 11-7: Russell's pleasure-arousal space for facial expression.

Psychologists of this view posit that facial expressions have a systematic, coherent, and meaningful structure that can be mapped to affective dimensions (Russell 1997), (Lazarus 1991), (Plutchik 1984), (Smith 1989), (Woodworth 1938). See figure 11-7 for an example. Hence, by considering the individual facial action components that contribute to that structure, it is possible to reveal much about the underlying properties of the emotion being expressed. It follows, that some of the individual features of expression have inherent signal value. This promotes a signaling system that is robust, flexible, and resilient (Smith & Scott 1997). It allows for the mixing of these components to convey a wide range of affective messages, instead of being restricted to a fixed pattern for each emotion. This variation allows for fine-tuning of the expression, as features can be emphasized, de-emphasized, added, or omitted as appropriate. Furthermore, it is well accepted that any emotion can be conveyed equally well by a range of expressions, as long as those expressions share a family resemblance. The resemblance exists because the expressions share common facial action units. It is also known that different expressions for different emotions share some of the same face action components (the raised brows of fear and surprise, for instance). It is hypothesized by Smith and Scott that those features held in common assign a shared affective meaning to each facial expression. The raised brows, for instance, convey attentional activity for both fear and surprise.

Russell argues the human observer perceives two broad affective categories on the face, arousal and pleasantness (Russell 1997). As shown in figure 11-7, Russell maps several emotions and corresponding expressions to these two dimensions. However, we found this scheme fairly limiting for Kismet. First, it is not clear how all the primary emotions are represented with this scheme (disgust is not accounted for). It also does not account for positively valenced yet reserved expressions such as a coy smile or a sly grin (which hints at a behavioral bias to withdraw). More importantly, *anger* and *fear* reside in very close proximity to each other despite their very different behavioral correlates. From an evolutionary perspective, the behavioral correlate of anger is to attack (which is a very strong approaching behavior), and the behavioral correlate for fear is to escape (which is a very strong withdrawing behavior). These are stereotypical responses derived from cross-species studies – obviously human behavior can vary widely. Nonetheless, from a practical engineering perspective of *generating* expression, it is better to separate these two emotional responses by a greater distance to minimize accidental activation of one instead of the other. Adding the stance dimension addressed these issues for Kismet.

Given this three dimensional affect space, our approach resonates well with the work of Smith and Scott. They posit a three dimensional space of *pleasure-displeasure* (maps to our valence), *attentional activity* (maps to our arousal), and *personal agency, control* (roughly maps to our stance). Figure 11-8 summarizes their proposed mapping of facial actions to these dimensions. They posit a fourth dimension that relates to the intensity of the expression. For Kismet, the expressions become more intense as the affect state moves to more extreme values in the affect space. As positive valence increases, Kismet's lip turn upward, the mouth opens, and the eyebrows relax. However, as valence decreases, the brows furrow, the jaw closes, and the lips turn downward. Along the arousal dimension, the ears perk, the eyes widen, and the mouth opens as arousal increases. Along the stance dimension, increasing positive values cause the eyebrows to arc outwards, the mouth to open, the ears to open, and the eyes to widen. These face actions roughly correspond to a decrease in personal agency/control in Smith and Scott's framework. For Kismet, it engenders an expression that looks more eager and accepting (or more uncertain for negative emotions). Although our dimensions do not map exactly to those hypothesized by Smith and Scott, the idea of combining meaningful face action units in a principled

	facial action							
meaning	Eyebrow frown	Raise eyebrows	raise upper eyelid	raise lower eyelid	up turn lip corners	open mouth	tighten mouth	raise chin
pleasantness	↓				↑	↑	↓	↓
goal obstacle/discrepancy	↑							
anticipated effort	↑							
attentional activity		↑	↑					
certainty		↓		↑		↑		
novelty		↑	↑					
personal agency/control		↓	↓			↓		

Figure 11-8: A possible mapping of facial movements to affective dimensions proposed by Smith and Scott. An up arrow indicates that the facial action is hypothesized to increase with increasing levels of the affective meaning dimension. A down arrow indicates that the facial action increases as the affective meaning dimension decreases. For instance, the lip corners turn upwards as "pleasantness" increases, and lower with increasing "unpleasantness".

manner to span the space of facial expressions, and to also relate them in a consistent way to emotion categories, holds strong.

11.4 Analysis of Facial Expressions

Ekman and Freisen (1978) developed a commonly used facial measurement system called *FACS*. The system measures the face itself as opposed to trying to infer the underlying emotion given a particular facial configuration. This is a comprehensive system that distinguishes all possible *visually* distinguishable facial movements. Every such facial movement is the result of muscle action (see figures 11-9, 11-10, 11-15). The earliest work in this area dates back to Duchenne, one of the first anatomists to explore how facial muscles change the appearance of the face (Duchenne 1862). Based on a deep understanding of how muscle contraction changes visible appearance, it is

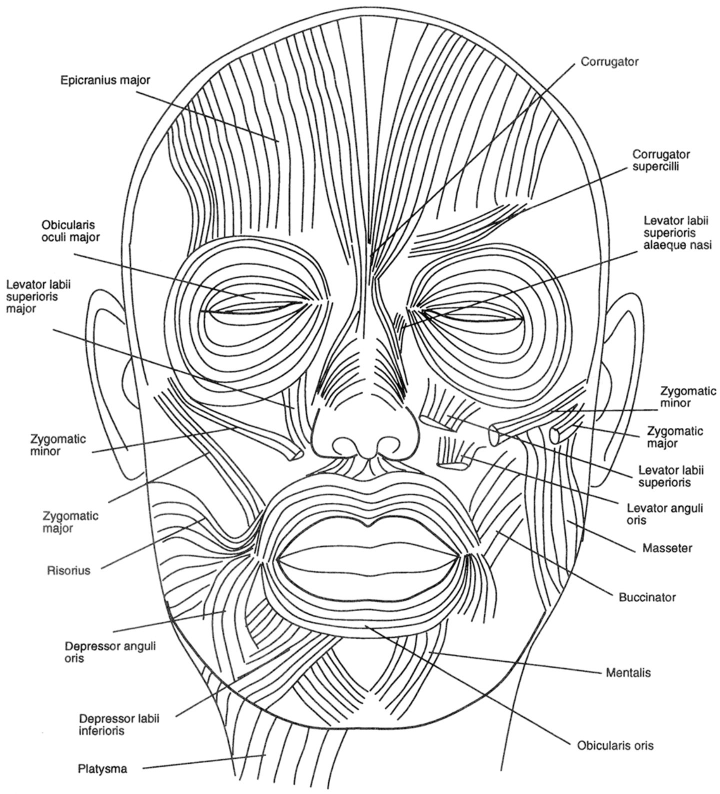


Figure 11-9: A schematic of the muscles of the face. Front view from (Parke & Waters 1996).

possible to decompose any facial movement into anatomically minimal action units. FACS has defined 33 distinct action units for the human face, many of which use a single muscle. However, it is possible for up to two to three muscles to map to a given action unit, since facial muscles often work in concert to adjust the location of facial features, and to gather, pouch, bulge, or wrinkle the skin.

To analyze Kismet's facial expressions, we can use FACS as a guideline. This must obviously be done within reason as Kismet lacks many of the facial features of humans (most notably, skin, teeth, and nose). However, the movements of Kismet's facial mechanisms were designed to roughly mimic those changes that arise in the human face due to the contraction of facial muscles. Kismet's eyebrow movements are shown in figure 11-11, and the eyelid movements in figure 11-12. Kismet's ears are primarily used to convey arousal and stance. Their movements are shown in figure 11-13. The lip and jaw movements are shown in figure 11-14.

Using the FACS system and analyzing the observations of Darwin (1872), Frijda (1969), Scherer (1984), and Smith (1989), Smith and Scott have compiled mappings of FACS action units to the expressions corresponding to anger, fear, happiness,

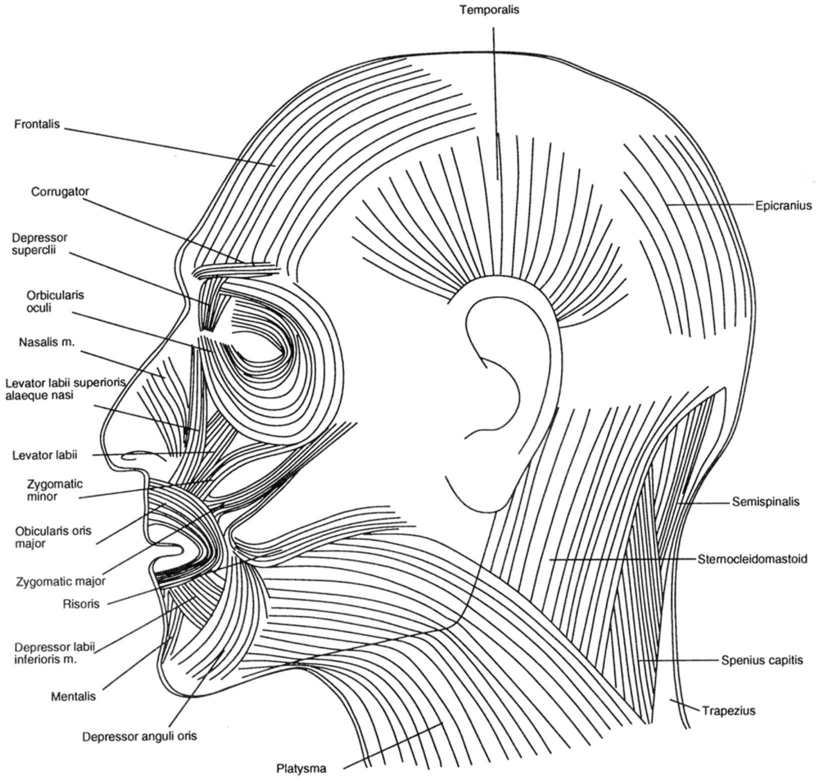


Figure 11-10: A schematic of the muscles of the face. Side view, from (Parke & Waters 1996).

surprise, disgust, and sadness (Smith & Scott 1997). The table, shown in figure 11-15, associates an action unit with an expression if two or more of these sources agreed on the association. The facial muscles employed are also listed. Note that these are not inflexible mappings. Any emotion can be expressed by a family of expressions. Also, the expressions vary in intensity. Nonetheless, this table highlights several key features.

Of the seven action units listed in the table, Kismet lacks only one (the lower eyelid). Of the facial features it does possess, it is capable of all the independent movements listed (given its own idiosyncratic mechanics). Kismet performs some of these movements in a manner that is different, yet roughly analogous, to that of a human. The series of figures, figure 11-11 to 11-14, relates the movement of Kismet's facial features to those of humans. There are two notable discrepancies. First, the use of the eyelids in Kismet's angry expression differs. In conjunction with brow knitting, Kismet lowers its eyelids to simulate a squint that is accomplished by raising both the lower and upper eyelids in humans. The second is the manner of arcing the eyebrows away from the centerline to simulate the brow configuration in sadness and fear. For

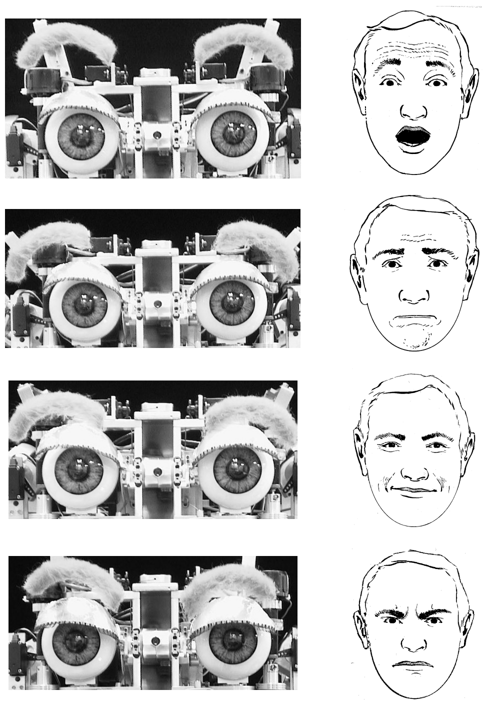


Figure 11-11: Kismet's eyebrow movements for expression. To the right, there is a human sketch displaying the corresponding eyebrow movement. The top figure shows the elevation of the brows characteristic of surprise. The upper-middle figure shows the brow movement that is characteristic of uncertainty or sorrow. The bottom-middle figure shows the eyebrows in their neutral position. The lower figure shows the brows knitted, as in an angry expression. The eyelids are also shown to lower as one moves from the top figure to the bottom figure.

humans, this corresponds to simultaneously knitting and raising the eyebrows. See figure 11-11.

Overall, Kismet does address each of the facial movements specified in the table (save those requiring a lower eyelid) in its own peculiar way. One question is how do people identify Kismet's facial expressions with human expressions? And do they map Kismet's distinctive facial movements to the corresponding human counterparts?

11.4.1 Comparison with Line Drawings of Human Expressions

To explore this question, we asked naive subjects to perform a comparison task where they compared color images of Kismet's expressions with a series of line drawings of human expressions. We felt it was unreasonable to have people compare images of

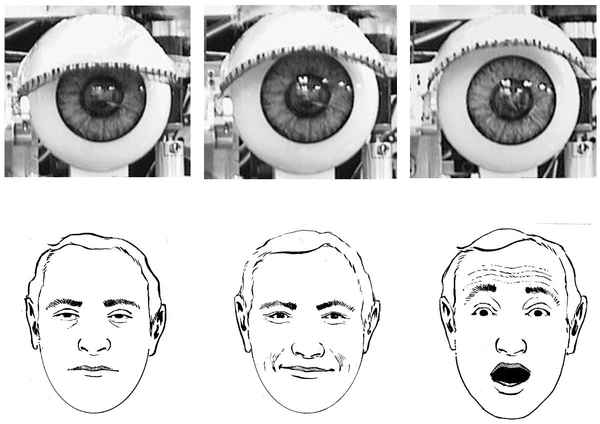


Figure 11-12: Kismet's eyelid movements for expression. Below each image of Kismet's eye, there is a human sketch displaying an analogous eyelid position. Kismet's eyelid rests just above the pupil for low arousal states. It rests just below the iris for neutral arousal states. It rests above the iris for high arousal states.

Kismet with human photos since the robot lacks skin. However, the line drawings provide a nice middle ground. The artist can draw lines that suggest the wrinkling of skin, but for the most part this is minimally done.

Ten subjects filled out the questionnaire. Five of the subjects were children (11 to 12 years old), and five were adults (ranging in age from 18 to 50). The gender split was four females and six males. The adults had never seen the robot before. Some of the children reported having seen a short school magazine article, so had some minimal familiarity.

The questionnaire was nine pages long. On each page was a color image of Kismet in one of nine facial expressions (from top to bottom, left to right they correspond to anger, disgust, happiness, content, surprise, sorrow, fear, stern, and a sly grin). These are shown in figure 11-5. Adjacent to the robot's picture was a set of twelve line drawings labeled *a* though *l*. The drawings are shown in figure 11-17 with our emotive labels. The subject was asked to circle the line drawing that most closely resembled the robot's expression. There was a short sequence of questions to probe

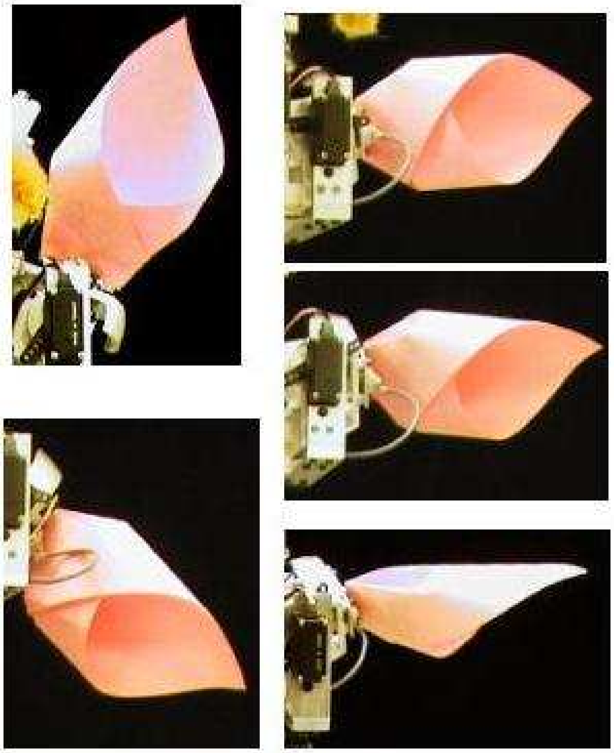


Figure 11-13: Kismet's ear movements for expression. There is no human counterpart, but they move somewhat like that of an animal. They are used to convey arousal by either pointing upwards as shown in the upper left figure, or by pointing downwards as shown in the bottom left figure. The ears also convey approach (the ears rotate forward as shown in the upper right figure) versus withdraw (the ears close as shown in the lower right figure). The middle right figure shows the ear in the neutral position.

the similarity of the robot to the chosen line drawing. One question asked how similar the robot's expression was to the selected line drawing. Another question asked the subject to list the labels of any other drawings they found to resemble the robot's expression and why. Finally, the subject could write any additional comments on the sheet. Table 11-16 presents the compiled results.

The results are substantially above random chance (8%), with the expressions corresponding to the primary emotions giving the strongest performance (70% and above). Subjects could infer the intensity of expression for the robot's expression of happiness (a contented smile verses a big grin). They had decent performance (60%) in matching Kismet's stern expression (produced by zero arousal, zero valence, and strong negative stance). The "sly grin" is a complex blend of positive valence, neutral arousal, and closed stance. This expression gave the subjects the most trouble, but their matching performance is still significantly above chance.

The misclassifications seem to arise from three sources. Certain subjects were

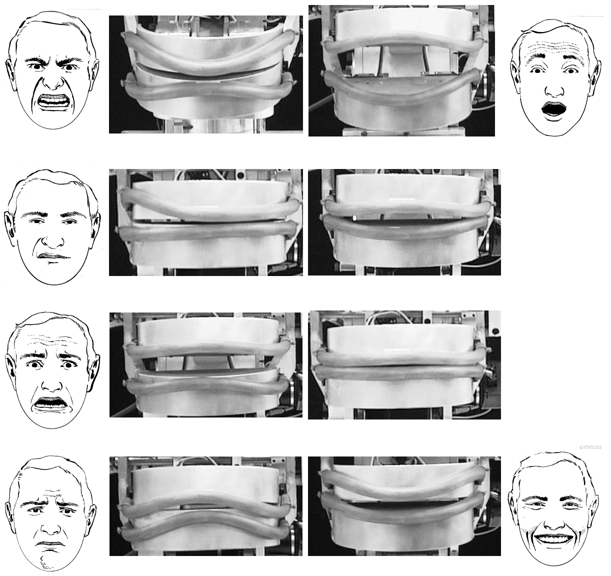


Figure 11-14: Kismet's lip movements for expression. Along side each of Kismet's lip postures is a human sketch displaying an analogous posture. The upper left figure shows Kismet bearing it's "teeth" which is characteristic of anger for animals as well as humans. The next figure down shows a curled lip characteristic of disgust. The next figure down shows a fearful grimace where the lips are parted but curved downwards. The lower left figure shows a frown. To the right, the upper figure shows the lips curved and parted, similar to an expression of surprise. The next figure down is a lower arousal version that indicates interest. The next figure down shows the lips in a neutral posture. The lower right figure shows the lips smiling.

confused by Kismet's lip mechanics. When the lips curve either up or down, there is a slight curvature in the opposite direction at the lever arm insertion point. Most subjects ignored the bit of curvature at the extremes of the lips, but others tried to match it to the lips in the line drawings. Occasionally, Kismet's frightened grimace was matched to a smile, or its smile matched to repulsion. Some misclassifications arose from matching the robot's expression to a line drawing that conveyed the same sentiment to the subject. For instance, Kismet's expression for disgust was matched to the line sketch of the "sly grin" because the subject interpreted both as "sneering" although none of the facial features match. Some associated Kismet's surprise expression with the line drawing of "happiness". There seems to be a positive valence communicated though Kismet's expression for surprise. Misclassifications also arose

facial action								
	Eyebrow frown	Raise eyebrows	raise upper eyelid	raise lower eyelid	up turn lip corners	down turn lip corners	open mouth	raise upper lip
muscular basis	corrugator supercilii	medial frontalis	levator palpebrae superioris	orbicularis oculi	zygomaticus major	depressor anguli oris	orbicularis oris	levator labii superioris
action units	4	1	5	6,7	12	15	26,27	9,10
emotion expressed								
happiness				x	x		x	
surprise		x	x				x	
anger	x		x	x				
disgust	x			x				x
fear	x	x	x				x	
sadness	x	x				x		

Figure 11-15: A summary of how FACS action units and facial muscles map to facial expressions for the primary emotions. Adapted from (Smith & Scott 97).

when subjects only seemed to match a single facial feature to a line drawing instead of multiple features. For instance, one subject matched Kismet's stern expression to the sketch of the "sly grin", noting the similarity in the brows (although the robot is not smiling). Overall, the subjects seem to intuitively match Kismet's facial features to those of the line drawings, and interpreted their shape in a similar manner. It is interesting to note that the robot's ears seem to communicate an intuitive sense of arousal to the subjects as well.

11.5 Evaluation of Expressive Behavior

The line drawing study did not ask the subjects what they thought the robot was expressing. However, this is clearly an important question for our purposes. To explore this issue, a separate questionnaire was devised. Given the wide variation in language that people use to describe expressions and our small number of subjects, a forced choice paradigm was adopted.

Seventeen subjects filled out the questionnaire. Most of the subjects were children twelve years of age (note that Kolb, Wilson & Laughlin (1992) found that the ability

	most similar sketch	data	comments
anger	**anger**	10/10	shape of mouth and eyebrows are strongest reported cues
disgust	**disgust**	8/10	shape of mouth is strongest reported cue
	sly grin	2/10	described as "sneering"
fear	**fear**	7/10	shape of mouth and eyes are strongest reported cues. Mouth open "aghast"
	surprise	1/10	subject associates look of "shock" with sketch of "surprise" over "fear"
	happy	1/10	lip mechanics cause lips to turn up at ends, sometimes confused with a weak smile
joy	**happy**	7/10	report lips and eyes are strongest cues. Ears may provide arousal cue to lend intensity.
	content	1/10	report lips used as strongest cue
	repulsion	1/10	lip mechanics turn lips up at end, causing shape reminiscent of lips in repulsion sketch
	surprise	1/10	perked ears, wide eyes lend high arousal. sometimes associated with a pleasant surprise
sorrow	**sad**	9/10	lips reported as strongest cue. Low ears may lend to low arousal.
	repulsion	1/10	lip mechanics turn lips up and end, causing shape reminiscent of repulsion sketch
surprise	**surprise**	9/10	reported open mouth, raised brows, wide eyes and elevated ears all lend to high arousal
	happy	1/10	subject remarks on similarity of eyes, but not mouth
pleased	**content**	9/10	reported relaxed smile, ears, and eyes lend low arousal and positive valence
	sly grin	1/10	subject reports the robot exhibiting a reserved pleasure. Associated with the "sly grin" sketch
sly grin	**sly grin**	5/10	lips and eyebrows reported as strongest cues
	content	3/10	subjects use robot's grin as the primary cue
	stern	1/10	subject report's the robot looking "serious", which is associated with "sly grin" sketch
	repulsion	1/10	lip mechanics curve lips up at end. Subject sees similarity with lips in "repulsion" sketch
stern	**stern**	6/10	lips and eyebrows are reported as strongest cues
	mad	1/10	subject reports robot looking "slightly cross". Cue on robot's eyebrows and pressed lips.
	tired	2/10	subjects may cue in on robot's pressed lips, low ears, lowered eyelids
	sly grin	1/10	subject reports similarity in brows.

Figure 11-16: Human subject's ability to map Kismet's facial features to those of a human sketch. The human sketches are shown in figure 11-17. Six of Kismet's basic expressions were tested (anger, disgust, fear, happiness, sorrow, and surprise). An intensity difference was explored (content versus happy). The stern expression is characteristic of a strongly closed stance with neutral valence and neutral arousal. An interesting blend of positive valence with closed stance was also tested (the sly grin).

to recognize expressions continues to develop, reaching adult level competence at approximately 14 years of age). There were six girls, six boys, three adult men, and two adult women. Again, none of the adults had seen the robot before. Some of the children reported minimal familiarity through reading a children's magazine article. There were seven pages in the questionnaire. Each page had a large color image of Kismet displaying one of seven expressions (anger, disgust, fear, happiness, sorrow, surprise, and a stern expression). The subjects could choose the best match from ten possible labels (accepting, anger, bored, disgust, fear, joy, interest, sorrow, stern, surprise). In a follow-up question, they could circle any other labels that they thought could also apply. With respect to their best-choice answer, they were asked to specify

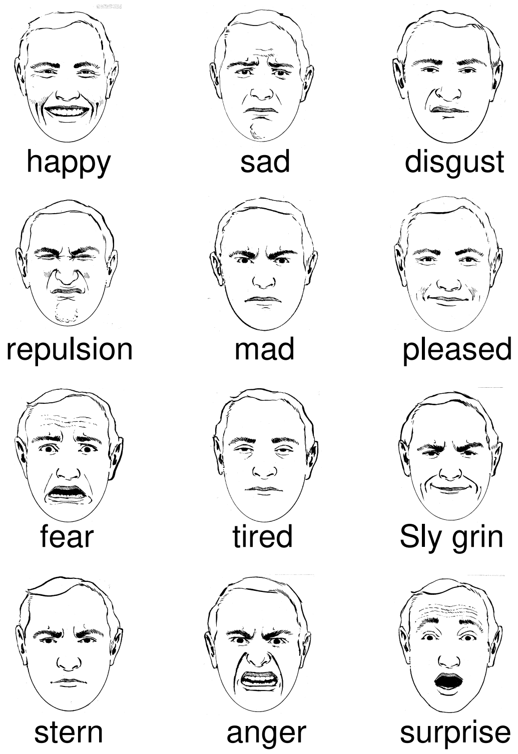


Figure 11-17: The sketches used in the evaluation, adapted from (Faigin 1990). The labels are for our purposes here, in the study they were labeled a through l. The nine Kismet stimuli are shown in figure 11-5.

on a ten-point scale how confident they were of their answer, and how intense they found the expression. The complied results are shown in figure 11-18. The subject's responses were significantly above random choice (10%), ranging from 47% to 83%.

Some of the misclassifications are initially confusing, but made understandable in light of the aforementioned study. Given that Kismet's surprise expression seems to convey positive valence, it is not surprising that some subjects matched it to *joy*. The knitting of the brow in Kismet's stern expression is most likely responsible for the associations with negative emotions such as anger and sorrow. Often, negatively valenced expressions were misclassified with negatively valenced labels. For instance, labeling the sad expression with fear, or the disgust expression with anger or fear. Kismet's expression for "fear" seems to give people the most difficulty. The lip mechanics probably account for the association with joy. The wide eyes, elevated brows, and elevated ears suggest high arousal. This may account for the confusion with

forced choice percentage (random=10%)

	accepting	anger	bored	disgust	fear	joy	interest	sorrow	stern	surprise	% correct
anger	5.9	76.5	0	0	5.9	11.7	0	0	0	0	76.5
disgust	0	17.6	0	70.6	5.9	0	0	0	5.9	0	70.6
fear	5.9	5.9	0	0	47.1	17.6	5.9	0	0	17.6	47.1
joy	11.7	0	5.9	0	0	82.4	0	0	0	0	82.4
sorrow	0	5.9	0	0	11.7	0	0	83.4	0	0	83.4
stern	7.7	15.4	0	7.7	0	0	0	15.4	53.8	0	53.8
surprise	0	0	0	0	0	17.6	0	0	0	82.4	82.4

Figure 11-18: This table summarizes the results of the color image based evaluation. The questionnaire was forced choice where the subject chose the emotive word that best matched the picture. See text.

surprise.

The still image and line drawing studies were useful in understanding how people read Kismet's facial expressions, but it says very little about expressive posturing. Humans and animals not only express with their face, but with their entire body. To explore this issue for Kismet, we showed a small group of subjects a set of video clips.

There were seven people who filled out the questionnaire. Six were children of age 12, four boys and two girls. One was an adult female. In each clip Kismet performs a coordinated expression using face and body posture. There were seven videos in all (anger, disgust, fear, joy, interest, sorrow, and surprise). Using a forced choice paradigm, for each video the subject was asked to select a word that best described the robot's expression (anger, disgust, fear, joy, interest, sorrow, surprise). On a ten-point scale, the subjects were also asked to rate the intensity of the robot's expression and the certainty of their answer. They were also asked to write down any comments they had. The results are compiled in table 11-19. Random chance is 14%.

forced choice percentage (random=14%)

	anger	disgust	fear	joy	interest	sorrow	surprise	% correct
anger	86	0	0	14	0	0	0	86
disgust	0	86	0	0	0	14	0	86
fear	0	0	86	0	0	0	14	86
joy	0	0	0	57	28	0	15	57
interest	0	0	0	0	71	0	29	71
sorrow	14	0	0	0	0	86	0	86
surprise	0	0	29	0	0	0	71	71

Figure 11-19: This table summarizes the results of the video evaluation.

The subjects performed significantly above chance, with overall stronger recognition performance than on the still images alone. The video segments of anger, disgust, fear, and sorrow were correctly classified with a higher percentage than the still images. However, there were substantially fewer subjects who participated in the video evaluation than the still image evaluation. The recognition of joy most likely dipped from the still image counterpart because it was sometimes confused with the expression of interest in the video study. The perked ears, attentive eyes, and smile give the robot a sense of expectation that could be interpreted as interest.

Misclassifications are strongly correlated with expressions having similar facial or postural components. Surprise was sometimes confused for fear; both have a quick withdraw postural shift (the fearful withdraw is more of a cowering movement whereas the surprise posture has more of an erect quality) with wide eyes and elevated ears. Surprise was sometimes confused with interest. Both have an alert and attentive quality, but interest is an approaching movement whereas surprise is more of a startled movement. Sorrow was sometimes confused with disgust; both are negative expressions with a downward component to the posture. The sorrow posture shift is more down and "sagging", whereas the disgust is a slow "shrinking" retreat.

Overall, the data gathered from these small evaluations suggests that people with little to no familiarity with the robot are able to interpret the robot's facial expressions and affective posturing. For our data set, there was no clear distinction in recognition performance between adults verses children, or males verses females. The subjects intuitively correlate Kismet's face with human likenesses (e.g., the line drawings). They map the expressions to corresponding emotion labels with reasonable consistency, and many of the errors can be explained through similarity in facial features or similarity in affective assessment (e.g., shared aspects of arousal or valence).

The data from the video studies suggest that witnessing the movement of the robot's face and body strengthens the recognition of the expression. However, more subjects must be tested to strengthen this claim. Nonetheless, observations from other interaction studies discussed throughout the thesis support this hypothesis. For instance, the postural shifts during the affective intent studies (see chapter 7) beautifully illustrate how subjects read and affectively respond to the robot's expressive posturing and facial expression. This is also illustrated in the social amplification studies of chapter 13. Based on the robot's withdraw and approach posturing, the subjects adapt their behavior to accommodate the robot.

11.6 Limitations and Extensions

More extensive studies need to be performed for us to make any strong claims about how accurately Kismet's expressions mirror those of humans. However, given our small sample size, the data suggests that Kismet's expressions are readable by people with minimal to no prior familiarity with the robot.

The evaluations have provided us with some useful input for how to improve the strength and clarity of Kismet's expressions. A lower eyelid should be added. Several subjects commented on this being a problem for them. We know from the FACS system that the movement of the lower eyelid is a key facial feature in expressing the basic emotions. The eyebrow mechanics could be improved. They should be able to elevate at both corners of the brow, as opposed to the arc of the current implementation. This would allow us to more accurately portray the brows of fear and sorrow. Kismet's mechanics attempts to approximate this, but the movement could be made stronger. The insertion point of the motor lever arm to the lips needs to be improved, or at least masked from plain view. Several subjects confused the additional curve at the ends for other lip shapes.

In this chapter, we have only evaluated the readability of Kismet's facial expressions. The evaluation of Kismet's facial displays will be addressed in chapter 14, when we discuss social interactions between human subjects and Kismet.

As a longer term extension, Kismet should be able to exert "voluntary" control over its facial expressions and be able to learn new facial displays. We have a strong interest in exploring facial imitation in the context of imitative games. Certain forms of facial imitation appear very early in human infants (Meltzoff & Moore 1977). Meltzoff posits that imitation is an important discovery procedure for learning about and understanding people. It may even play a role in the acquisition of a theory of mind. For adult level human social intelligence, the question of how a robot could have a genuine theory of mind will need to be addressed.

11.7 Summary

We have developed a framework to control the facial movements of Kismet. The expressions and displays are generated in real-time and serve four facial functions. The lip synchronization and facial emphasis subsystem is responsible for moving the lips and face to accompany expressive speech. The emotive facial expression system is responsible for computing an appropriate emotive display. The facial display and behavior subsystem produces facial movements that serve communicative functions (such as regulating turn taking) as well as producing the facial component of behavioral responses. With so many facial functions competing for the face actuators, a dynamic prioritizing scheme was developed. This system addresses the issues of blending as well as sequencing the concurrent requests made by each of the face subsystems. The overall face control system produces facial movements that are timely, coherent, intuitive and appropriate. It is organized in a principled manner so that incremental improvements and additions can be made. An intriguing extension is to learn new facial behaviors through imitative games with the caregiver, as well as to learn their social significance.

Chapter 12

Expressive Vocalization System

> *The language-creating process is a social one, a product of the interaction between the child and those with whom his experiences are shared in common... At no stage is language development an individual matter. Meaning and learning to mean are social processes. From birth the child is one among others. Ultimately language has been shaped by the functions it has to serve in the actions and reflections of reality by the child.* Halliday (1979).

Halliday (1975) explores the acquisition of meaningful communication acts from the viewpoint of how children *use* language to serve themselves in the course of daily life. From this perspective, a very young child may already have a linguistic system *long before* he has any words or grammar. Prior to uttering his first words, a baby is capable of expressing a considerable range of meanings which bear little resemblance to adult language, but which can be readily interpreted from a functional perspective, i.e. *what has the baby learned to do by means of language?* At a very young age, he is able to use his voice for doing something; it is a form of action that influences the behavior of the external world (such as the caregiver), and these meaningful vocal acts soon develop their own patterns and are used in their own significant contexts.

From Kismet's inception, the synthetic nervous system has been designed with an eye toward exploring the acquisition of meaningful communication. As Haliday argues, this process is driven internally through motivations and externally through social engagement with caregivers. Much of Kismet's social interaction with its caregivers is based on vocal exchanges when in face-to-face contact. At some point, these exchanges could be ritualized into a variety of vocal games that could ultimately serve as learning episodes for the acquisition of shared meanings. Towards this goal, this chapter focuses on Kismet's vocal production, expression, and delivery.

12.1 Design Issues

Production of Novel Utterances

Given the goal of acquiring a proto-language, Kismet must be able to experiment with its vocalizations to explore their effects on the caregiver's behavior. Hence the vocalization system must support this exploratory process. At the very least the system should support the generation of short strings of phonemes, modulated by pitch, duration, and energy. Human infants play with the same elements (and more) when exploring their own vocalization abilities and the effect these vocalizations have on their social world.

Expressive Speech

Kismet's vocalizations should also convey the affective state of the robot. This provides the caregiver with important information as to how to appropriately engage Kismet. The robot could then use its emotive vocalizations to convey disapproval, frustration, disappointment, attentiveness, or playfulness. As for human infants, this ability is important for meaningful social exchanges with Kismet. It helps the caregiver to correctly read the robot and to treat the robot as an intentional creature. This fosters richer and sustained social interaction, and helps to maintain the person's interest as well as that of the robot.

Lip Synchronization

For a compelling verbal exchange, it is also important for Kismet to accompany its expressive speech with appropriate motor movements of the lips, jaw, and face. The ability to lip synchronize with speech strengthens the perception of Kismet as a social creature that expresses itself vocally. A disembodied voice would be a detriment to the life-like quality of interaction that we have worked so hard to achieve in many different ways. Furthermore, it is well accepted that facial expressions (related to affect) and facial displays (which serve a communication function) are important for verbal communication. Synchronized movements of the face with voice both complement as well as supplement the information transmitted through the verbal channel. For Kismet, the information communicated to the human is grounded in affect. The facial displays are used to help regulate the dynamics of the exchange.

12.2 Emotion in Speech

There has been an increasing amount of work in identifying those acoustic features that vary with the speaker's affective state (Murray & 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 sub-glottal 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. Picard (1997) presents a nice overview of work in this area.

The effect of emotions on the human voice

	fear	anger	sorrow	joy	disgust	surprise
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	breathy blaring	grumbled chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	slurring	normal	normal	

Figure 12-1: Typical effect of emotions on adult human speech. Adapted from (Murray and Arnott 1993). The table has been extended to include some acoustic correlates of the emotion of surprise.

Hence, the effects of emotion in speech tend to alter the pitch, timing, voice quality, and articulation of the speech signal (Cahn 1990). Table 12-1 summarizes these key features. However, several of these features are also modulated by the prosodic effects that the speaker uses to communicate grammatical structure and lexical correlates. These tend to have a more localized influence on the speech signal, such as emphasizing a particular word. For recognition tasks, this makes isolating those feature characteristics modulated by emotion challenging.

Even humans are not perfect at perceiving the intended emotion for those emotional states that have similar acoustic characteristics. For instance, surprise can be perceived or understood as either joyous surprise (happiness) or apprehensive surprise (fear). Disgust is a form of disapproval and can be confused with anger.

There have been a few systems developed to synthesize emotional speech. The *Affect Editor* by Janet Cahn is among the earliest work in this area (Cahn 1990). Her system was based on *DECtalk3*, a commercially available text-to-speech speech synthesizer. Given an English sentence and an emotional quality (one of anger, disgust, fear, joy, sorrow, or surprise), she developed a methodology for mapping the emotional correlates of speech (changes in pitch, timing, voice quality, and articulation) onto the underlying DECtalk synthesizer settings. She took great care to introduce the global prosodic effects of emotion while still preserving the more local influences of grammatical and lexical correlates of speech intonation. In a different approach

Jun Sato (see *www.ee.seikei.ac.jp/user/junsato/research/*)trained a neural network to modulate a neutrally spoken speech signal (in Japanese) to convey one of four emotional states (happiness, anger, sorrow, disgust). The neural network was trained on speech spoken by Japanese actors. This approach has the advantage that the output speech signal sounds more natural than purely synthesized speech. However, it has the disadvantage that the speech input to the system must be pre-recorded.

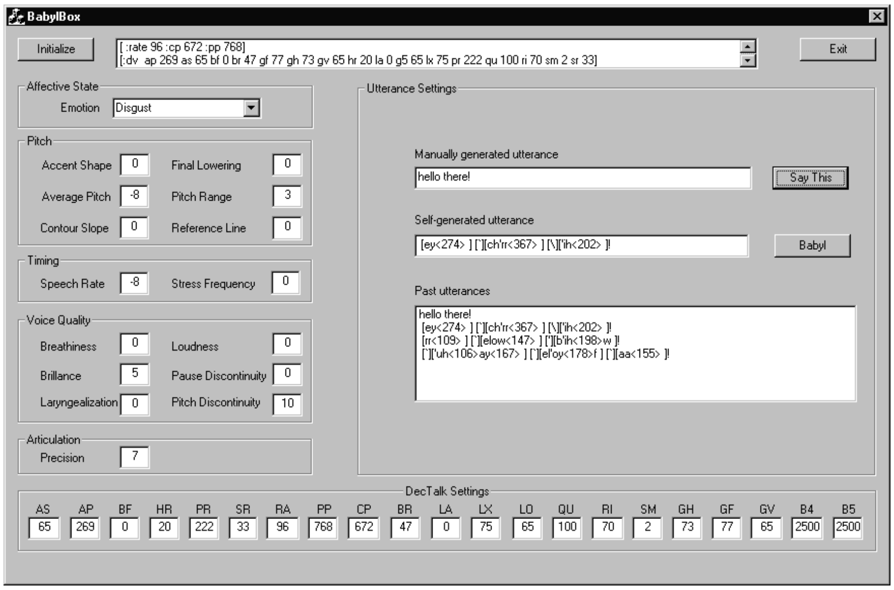


Figure 12-2: Kismet's expressive speech GUI. Listed is a selection of emotive qualities, the vocal affect parameters, and the synthesizer settings. A user can either manually enter an English phrase to be said, or can request an automatically generated "kismet-esque" babble. During run-time, Kismet operates in automatic generation mode.

With respect to giving Kismet the ability to generate emotive vocalizations, Janet Cahn's work (Cahn 1990) is a valuable resource. The DECtalk software gives us the flexibility to have Kismet generate its own utterance by assembling strings of phonemes (with pitch accents). We use Cahn's technique for mapping the emotional correlates of speech (as defined by her vocal affect parameters) to the underlying synthesizer settings. Because Kismet's vocalizations are at the proto-dialog level, there is no grammatical structure. As a result, we are only concerned with producing the purely global emotional influence on the speech signal.

12.3 Expressive Voice Synthesis

Cahn's *vocal affect parameters (VAP)* alter the pitch, timing, voice quality, and articulation aspects of the speech signal. She documented how these parameter settings can be set to convey anger, fear, disgust, gladness, sadness, and surprise in synthetic speech. Emotions have a global impact on speech since they modulate the respiratory system, larynx, vocal tract, muscular system, heart rate, and blood pressure. The pitch related parameters affect the pitch contour of the speech signal, which is the primary contributor for affective information. The pitch related parameters include *accent shape*, *average pitch*, *pitch contour slope*, *final lowering*, *pitch range*, and *pitch reference line*. The timing related parameters modify the prosody of the vocalization, often being reflected in speech rate and stress placement. The timing related parameters include *speech rate*, *pauses*, *exaggeration*, and *stress frequency*. The voice quality parameters include *loudness*, *brilliance*, *breathiness*, *laryngealization*, *pitch discontinuity*, and *pause discontinuity*. The articulation parameter modifies the precision of what is uttered, either being more enunciated or slurred. We describe these parameters in detail in the next section.

12.4 The Vocal Affect Parameters

For Kismet, only some of these parameters are needed since several are inherently tied to sentence structure (the types and placement of pauses, for instance). See figure 12-2. In this section, we briefly describe those VAPs that are incorporated into Kismet's synthesized speech. These vocal affect parameters modify the DECtalk synthesizer settings (summarized in table 12-3) according to the emotional quality to be expressed. The default values and max/min bounds for these settings are given in table 12-4. There is currently a single fixed mapping per emotional quality. Table 12-5 along with the equations presented in section 12.5.1 summarize how the vocal affect parameters are mapped to the DECtalk synthesizer settings. Table 12-6 summarizes how each emotional quality of voice is mapped onto the VAPs. Slight modifications in Cahn's specifications were made for Kismet – this should not be surprising as a different, more child-like voice was used. The discussion below motivates the mappings from VAPs to synthesizer settings as shown in figure 12-5. Cahn (1990) presents a detailed discussion of how these mappings were derived.

12.4.1 Pitch Parameters

The following six parameters influence the pitch contour of the spoken utterance. The pitch contour is the trajectory of the fundamental frequency, f_0 over time.

Accent Shape

Modifies the shape of the pitch contour for any pitch accented word by varying the rate of f_0 change about that word. A high accent shape corresponds to speaker agitation where there is a high peak f_0 and a steep rising and falling pitch contour slope. This parameter has a substantial contribution to DECtalk's `stress rise` setting, which regulates the f_0 magnitude of pitch accented words.

DECtalk Synthesizer Setting	Description
`average pitch` (Hz)	The average pitch of the pitch contour.
`assertiveness` (%)	The degree to which the voice tends to end statements with a conclusive fall.
`baseline fall` (Hz)	The desired fall (in Hz) of the baseline.The reference pitch contour around which all rule governed dynamic swings in pitch are about.
`breathiness` (dB)	Specifies the breathy quality of the voice due to the vibration of the vocal folds
`comma pause` (ms)	Duration of pause due to comma.
`gain of frication`	Gain of frication sound source.
`gain of aspiration`	Gain of aspiration sound source.
`gain of voicing`	Gain of voicing sound source.
`hat rise` (Hz)	Nominal hat rise to the pitch contour plateau upon the first stressed syllable of the phrase. The hat-rise influence lasts throughout the phrase.
`laryngealization` (%)	Creaky voice. Results when the glottal pulse is narrow and the fundamental period is irregular.
`loudness` (dB)	Controls amplitude of speech waveform.
`lax breathiness` (%)	Specifies the amount of breathiness applied to the end of a sentence when going from voiced to voiceless sounds.
`period pause` (msec)	Duration of pause due to period.
`pitch range` (%)	Sets the range about the average pitch that the pitch contour expands and contracts. Specified in terms of % of the nominal pitch range.
`quickness` (%)	Controls the speed of response to sudden requests to change pitch (due to pitch accents). Models the response time of the larynx.
`speech rate` (wpm)	Rate of speech.
`richness` (%)	Controls the spectral change at lower frequencies (enhances the lower frequencies). Rich and brilliant voices are more forceful.
`smoothness` (%)	Controls the amount of high frequency energy. There is less high frequency energy in a smoother voice. Varies inversely with brilliance. Smoother voices sound friendlier.
`stress rise` (Hz)	The nominal height of the pitch rise and fall on each stressed syllable. This has a local influence on the contour about the stressed syllable.

Figure 12-3: A description of the DECtalk synthesizer settings (see the DECtalk Software Reference Guide). Figure 12-11 illustrates the nominal pitch contour for neutral speech, and the net effect of changing these values for different expressive states. (Cahn 1990) presents a detailed description of how each of these settings alters the pitch contour.

Average Pitch

Quantifies how high or low the speaker appears to be speaking relative to their normal speech. It is the average f_0 value of the pitch contour. It varies directly with DECtalk's `average pitch`.

DECtalk Synthesizer Setting	unit	neutral setting	min setting	max setting
average pitch	Hz	306	260	350
assertiveness	%	65	0	100
baseline fall	Hz	0	0	40
breathiness	dB	47	40	55
comma pause	msec	160	-20	800
gain of frication	dB	72	60	80
gain of aspiration	dB	70	0	75
gain of voicing	dB	55	65	68
hat rise	Hz	20	0	80
laryngealization	%	0	0	10
loudness	dB	65	60	70
lax breathiness	%	75	100	0
period pause	msec	640	-275	800
pitch range	%	210	50	250
quickness	%	50	0	100
speech rate	wpm	180	75	300
richness	%	40	0	100
smoothness	%	5	0	100
stress rise	Hz	22	0	80

Figure 12-4: Default DECtalk synthesizer settings for Kismet's voice (see the DECtalk Software Reference Guide). Section 12.5.1 describes the equations for altering these values to produce Kismet's expressive speech.

Contour Slope

Describes the general direction of the pitch contour, which can be characterized as rising, falling, or level. It contributes to two DECtalk settings. It has a small contribution to the `assertiveness` setting, and varies inversely with the `baseline fall` setting.

Final Lowering

Refers to the amount that the pitch contour falls at the end of an utterance. In general, an utterance will sound emphatic with a strong final lowering, and tentative if weak. It can also be used as an auditory cue to regulate turn taking. A strong final lowering can signify the end of a speaking turn, whereas a speaker's intention to continue talking can be conveyed with a slight rise at the end. This parameter strongly contributes to DECtalk's `assertiveness` setting and somewhat to the `baseline fall` setting.

Pitch Range

Measures the bandwidth between the maximum to minimum f_0 of the utterance. The pitch range expands and contracts about the average f_0 of the pitch contour. It varies directly with DECtalk's `pitch range` setting.

Reference Line

Controls the reference pitch f_0 contour from which rule governed swings are about (such expansions and contractions in pitch due to pitch accents). DECtalk's `hat rise` setting approximates this very roughly.

12.4.2 Timing

The vocal affect timing parameters contribute to speech rhythm. Such correlates arise in emotional speech from physiological changes in respiration rate (changes in breathing patterns) and level of arousal.

Speech Rate

Controls the rate of words or syllables uttered per minute. It influences how quickly an individual word or syllable is uttered, the duration of sound to silence within an utterance, and the relative duration of phoneme classes. Speech is faster with higher arousal and slower with lower arousal. This parameter varies directly with DECtalk's `speech rate` setting. It varies inversely with DECtalk's `period pause` and `comma pause` settings as faster speech is accompanied with shorter pauses.

Stress Frequency

Controls the frequency of occurrence of pitch accents and determines the smoothness or abruptness of f_0 transitions. As more words are stressed, the speech sounds more emphatic and the speaker more agitated. It filters other vocal affect parameters such as precision of articulation and accent shape, and thereby contributes to the associated DECtalk settings.

DECtalk Synthesizer Setting	DECtalk Symbol	norm	Controlling Vocal Affect Parameter(s)	Percent of Control
average pitch	ap	.51	average pitch	1
assertiveness	as	.65	final lowering contour direction	.8 .2
baseline fall	bf	0	contour direction final lowering	-.5 .5
breathiness	br	.46	breathiness	1
comma pause	:cp	.238	speech rate	-1
gain of frication	gf	.6	precision of articulation	1
gain of aspiration	gh	.933	precision of articulation	1
gain of voicing	gv	.76	loudness precision of articulation	.6 .4
hat rise	hr	.2	reference line	1
laryngealization	la	0	laryngealization	1
loudness	lo	.5	loudness	1
lax breathiness	lx	.75	breathiness	1
period pause	:pp	.666	speech rate	-1
pitch range	pr	.8	pitch range	1
quickness	qu	.5	pitch discontinuity	1
speech rate	:ra	.2	speech rate	1
richness	ri	.4	brillance	1
smoothness	sm	.05	brillance	-1
stress rise	sr	.220	accent shape pitch discontinuity	.8 .2

Figure 12-5: Percent contributions of vocal affect parameters to DECtalk synthesizer settings. The absolute values of the contributions in the far right column add up to 1 (100%) for each synthesizer setting. See the equations in section 12.5.1 for the mapping. The equations are similar to those used by Cahn.

12.4.3 Voice Quality

Emotion can induce not only changes in pitch and tempo, but in voice quality as well. These phenomena primarily arise from changes in the larynx and articulatory tract.

Breathiness

Controls the aspiration noise in the speech signal. It adds a tentative and weak quality to the voice, when speaker is minimally excited. DECtalk `breathiness` and `lax breathiness` vary directly with this.

Brillance

Controls the perceptual effect of relative energies of the high and low frequencies. When agitated, higher frequencies predominate and the voice is harsh or "brillant". When speaker is relaxed or depressed, lower frequencies dominate and the voice sounds soothing and warm. DECtalk's `richness` setting varies directly as it enhances the lower frequencies. In contrast, DECtalk's `smoothness` setting varies inversely since it attenuates higher frequencies.

Laryngealization

Controls the perceived creaky voice phenomena. It arises from minimal sub-glottal pressure and a small open quotient such that f_0 is low, the glottal pulse is narrow, and the fundamental period is irregular. It varies directly with DECtalk's `laryngealization` setting.

Loudness

Controls the amplitude of the speech waveform. As a speaker becomes aroused, the sub-glottal pressure builds which increases the signal amplitude. As a result, the voice sounds louder. It varies directly with DECtalk's `loudness` setting. It also influences DECtalk's `gain of voicing`.

Pause Discontinuity

Controls the smoothness of f_0 transitions from sound to silence for unfilled pauses. Longer or more abrupt silences correlate with being more emotionally upset. It varies directly with DECtalk's `quickness` setting.

Pitch Discontinuity

Controls smoothness or abruptness of f_0 transitions, and the degree to which the intended targets are reached. With more speaker control, the transitions are smoother. With less control, they transitions are more abrupt. It contributes to DECtalk's `stress rise` and `quickness` settings.

12.4.4 Articulation

The autonomic nervous system modulates articulation by inducing an assortment of physiological changes such as causing dryness of mouth or increased salivation.

Precision

Controls a range of articulation from enunciated to slurring. Slurring has minimal frication noise, whereas greater enunciation for consonants results in this increasing. Stronger enunciation also results in an increase in aspiration noise and of voicing. The precision of articulation varies directly with DECtalk's `gain of frication`, `gain of voicing`, and `gain of aspiration`.

	Anger	disgust	fear	happy	sad	surprise	neutral
accent shape	10	0	10	10	-7	9	0
average pitch	-10	-10	10	3	-7	6	0
contour slope	10	0	10	0	0	10	0
final lowering	10	5	-10	-4	8	-10	0
pitch range	10	5	10	10	-10	10	0
reference line	-10	0	10	-8	-1	-8	0
speech rate	4	-8	10	3	-6	6	0
stress frequency	0	0	10	5	1	0	0
breathiness	-5	0	0	-5	0	-9	0
brillance	10	5	10	-2	-6	9	0
laryngealization	0	0	-10	0	0	0	0
loudness	10	-5	10	8	-5	10	0
pause discontinuity	10	0	10	-10	-8	-10	0
pitch discontinuity	3	10	10	6	0	10	0
precision of articulation	10	7	0	-3	-5	0	0

Figure 12-6: The mapping from each expressive quality of speech to the vocal affect parameters (VAPs). There is a single fixed mapping for each emotional quality.

12.5 Implementation Overview

This section overviews the process for generating an expressive utterance and having Kismet say it with lip synchronization and facial animation (as summarized below). Figure 12-2 shows the controls of the expressive speech synthesizer as it appears on the NT machine. The active "emotion" is established by the emotion system as described in chapter 8. Currently, Kismet can vocalize an utterance with one of seven expressive states (anger, disgust, fear, happy, sad, surprise, and neutral). The decision to engage in vocal behavior is determined by the behavior system (chapter 9). This information is passed to the motor skills system (chapter 10) where the request to speak with a given affective state is issued to the vocalization system. In the remainder of this chapter, we present how the vocalization system processes this request.

The algorithm for generating and performing an expressive kismet-like vocalization is as follows:

1. Update vocal affect parameters based on current emotion.

2. Map from vocal affect parameters to synthesizer settings.

3. Generate the utterance to speak.

4. Assemble the full command and send it to the synthesizer.

5. Extract features from speech signal for lip synchronization.

6. Send the speech signal to the sound card.

7. Execute lip synchronization movements.

12.5.1 Mapping Vocal Affect Parameters to Synthesizer Settings

The vocal affect parameters outlined in section 12.4 are derived from the acoustic correlates of emotion in human speech. To have DECtalk produce these effects in synthesized speech, we need to computationally map these vocal affect parameters to the underlying synthesizer settings. There is a single fixed mapping per emotional quality. With some minor modifications, we adapt Cahn's mapping functions to Kismet's implementation.

The vocal affect parameters can assume integer values within the range of $(-10, 10)$. Negative numbers corresponds to lesser effects, positive numbers correspond to greater effects, and zero is the neutral setting. Hence, for neutral speech, all vocal affect parameters are set to zero. These values are set according to the current specified emotion as shown in table 12-6.

Linear changes in these parameter values result in a non-linear change in synthesizer settings. Furthermore, the mapping between parameters and synthesizer settings is not necessarily one-to-one. Each parameter affects a percent of the final synthesizer setting's value (table 12-5). When a synthesizer setting is modulated by more than one parameter, its final value is the sum of the effects of the controlling parameters. The total of the absolute values of these percentages must be 100%. See table 12-4 for the allowable bounds of synthesizer settings. The computational mapping occurs in three stages.

Consonants				Vowels		Vowels	
b	bet	n	net	aa	bob	oy	boy
ch	chin	nx	sing	ae	bat	rr	bird
d	debt	p	pet	ah	but	uh	book
dh	this	r	red	ao	bought	uw	lute
el	bottle	s	sit	aw	bout	yu	cute
en	button	sh	shin	ax	about	allophones	
f	fin	t	test	ay	bite	dx	rider
g	guess	th	thin	eh	bet	lx	will
hx	head	v	vest	ey	bake	q	we eat
jh	gin	w	wet	ih	bit	rx	oration
k	ken	yx	yet	ix	kisses	tx	Latin
l	let	z	zoo	iy	beat	Silence	
m	met	zh	azure	ow	boat	_(underscore)	

Figure 12-7: Dectalk phonemes for generating utterances.

In the first stage, the percentage of each of the VAPs (VAP_i) to its total range is computed, (PP_i). This is given by the equation:

$$PP_i = \frac{VAP_{value_i} + VAP_{offset}}{VAP_{max} - VAP_{min}}$$

VAP_i is the current VAP under consideration, VAP_{value} is its value specified by the current emotion, $VAP_{offset} = 10$ adjusts these values to be positive, $VAP_{max} = 10$, and $VAP_{min} = -10$.

In the second stage, a weighted contribution ($WC_{j,i}$) of those VAP_i that control each of DECtalk's synthesizer settings (SS_j) is computed. The far right column of table 12-5 specifies each of the corresponding *scale factors* ($SF_{j,i}$). Each scale factor represents a percentage of control that each VAP_i applies to its synthesizer setting SS_j.

For each synthesizer setting, SS_j:
    For each corresponding scale factor, $SF_{j,i}$ of VAP_i:
        If $SF_{j,i} \geq 0$
            $WC_{j,i} = PP_i \times SF_{j,i}$
        If $SF_{j,i} \leq 0$
            $WC_{j,i} = (1 - PP_i) \times (-SF_{j,i})$
    $SS_j = \sum_i WC_{j,i}$

symbol	name	indicates	symbol	name	indicates
[']	apostrophe	primary stress	[,]	comma	clause boundaries
[`]	grave accent	secondary stress	[.]	period	period
["]	quotation mark	emphatic stress	[?]	question mark	question mark
[/]	slash	pitch rise	[!]	exclamation mark	exclamation mark
[\]	backslash	pitch fall	[ ]	space	word boundary
[/ \]	hat	pitch rise and fall			

Figure 12-8: Dectalk accents and end syntax for generating utterances.

At this point, each synthesizer value has a value $0 \leq SS_j \leq 1$. In the final stage, each synthesizer setting SS_j is scaled about 0.5. This produces the final synthesizer value, $SS_{j_{final}}$. The final value is sent to the speech synthesizer. The maximum, minimum and default values of the synthesizer settings are shown in table 12-4.

For each final synthesizer setting, $SS_{j_{final}}$:
Compute $SS_{j_{offset}} = SS_j - norm$
If $SS_{j_{offset}} \geq 0$
$SS_{j_{final}} = SS_{j_{default}} + (2 \times SS_{j_{offset}} \times (SS_{j_{max}} - SS_{j_{min}}))$
If $SS_{j_{offset}} \leq 0$
$SS_{j_{final}} = SS_{j_{default}} + (2 \times SS_{j_{offset}} \times (SS_{j_{default}} - SS_{j_{min}}))$

12.5.2 Generating the Utterance

To engage in proto-dialogs with its human caregiver and to partake in vocal play, Kismet must be able to generate its own utterances. The algorithm outlined below produces a style of speech that is reminiscent of a tonal dialect. As it stands, it is quite distinctive and contributes significantly to Kismet's personality (as it pertains to its manner of vocal expression). However, it is really intended as a place-holder for a more sophisticated utterance generation algorithm to eventually replace it. In time, Kismet will be able to adjust its utterance based on what it hears, but this is the subject of future work.

Based upon DECtalk's phonemic speech mode, the generated string to be synthesized is assembled from pitch accents, phonemes, and end syntax. The end syntax is a requirement of DECtalk and does not serve a grammatical function. However, as with the pitch accents, it does influence the prosody of the utterance and is used in this manner. The DECtalk phonemes are summarized in table 12-7 and the accents are summarized in table 12-8.

Randomly choose number of proto-words, $getUtteranceLength() = length_{utterance}$
For $i = (0, length_{utterance})$, generate a proto-word, $protoWord$
    Generate a $(wordAccent, word)$ pair
    Randomly choose word accent, $getAccent()$
    Randomly choose number of syllables of proto-word, $getWordLength() = length_{word}$
    Choose which syllable receives primary stress, $assignStress()$
    For $j = (0, length_{word})$ generate a syllable
        Randomly choose the type of syllable, $syllableType$
        if $syllableType = vowelOnly$
            if this syllable has primary stress
            then $syllable = getStress() + getVowel() + getDuration()$
            else $syllable = getVowel() + getDuration()$
        if $syllableType = consonantVowel$
            if this syllable has primary stress
            then $syllable = getConsonant() + getStress() + getVowel() + getDuration()$
            else $syllable = getConsonant() + getVowel() + getDuration()$
        if $syllableType = consonantVowelConsonant$
            if this syllable has primary stress
            then $syllable = getConsonant() + getStress() + getVowel() + getDuration() + getConsonant()$
            else $syllable = getConsonant() + getVowel() + getDuration() + getConsonant()$
        if $syllableType = vowelVowel$
            if this syllable has primary stress
            $syllable = getStress() + getVowel() + getDuration() + getvowel() + getDuration()$
            else $syllable = getVowel() + getDuration() + getVowel() + getDuration()$
        $protoWord = append(protoWord, syllable)$
    $protoWord = append(wordAccent, protoWord)$
$utterance = append(utterance, protoWord)$

Where:

- $GetUtteranceLength()$ randomly chooses a number between $(1, 5)$. This specifies the number of proto-words in a given utterance.
- $GetWordLength()$ randomly chooses a number between $(1, 3)$. This specifies the number of syllables in a given proto-word.

- *GetPunctuation*() randomly chooses one of end syntax markers as shown in table 12-8. This is biased by emotional state to influence the end of the pitch contour.

- *GetAccent*() randomly choose one of six accents (including no accent) as shown in table 12-8.

- *assignStress*() selects which syllable receives primary stress.

- *getVowel*() randomly choose one of eighteen vowel phonemes as shown in figure 12-7.

- *getConsonant*() randomly chooses one of twenty-six consonant phonemes as shown in table 12-7.

- *getStress*() gets the primary stress accent.

- *getDuration*() randomly chooses a number between $(100, 500)$ that specifies the vowel duration in msec. This selection is biased by the emotional state where lower arousal vowels tend to have longer duration, and high arousal states have shorter duration.

12.6 Analysis and Evaluation

Given the phonemic string to be spoken and the updated synthesizer settings, Kismet can vocally express itself with different emotional qualities. To evaluate Kismet's speech, we can analyze the produced utterances with respect to the acoustical correlates of emotion. This will reveal if the implementation produces similar acoustical changes to the speech waveform given a specified emotional state. We can also evaluate how the affective modulations of the synthesized speech are perceived by human listeners.

12.6.1 Analysis of Speech

To analyze the performance of the expressive vocalization system, we extracted the dominant acoustic features that are highly correlated with emotive state. The acoustic features and their modulation with emotion are summarized in table 12-1. Specifically, these are average pitch, pitch range, pitch variance, and mean energy. To measure speech rate, we extracted the overall time to speak and the total time of voiced segments.

We extracted these features from three phrases:

- *Look at that picture*

- *Go to the city*

- *It's been moved already*

	nzpmean	nzpvar	pmax	pmin	prange	egmean	length	voiced	unvoiced
anger-city	292.5	6348.7	444.4	166.7	277.7	112.2	81	52	29
anger-moved	269.1	4703.8	444.4	160	284.4	109.8	121	91	30
anger-picture	273.2	6850.3	444.4	153.8	290.6	110.2	112	51	61
anger-average	278.3	5967.6	444.4	160.17	284.2	110.7	104.6	64.6	40
calm-city	316.8	802.9	363.6	250	113.6	102.6	85	58	27
calm-moved	304.5	897.3	363.6	266.7	96.9	103.6	124	94	30
calm-picture	302.2	1395.5	363.6	235.3	128.3	102.4	118	73	45
calm-average	307.9	1031.9	363.6	250.67	112.93	102.9	109	75	34
disgust-city	268.4	2220.0	400	173.9	226.1	102.5	124	83	41
disgust-moved	264.6	1669.2	400	190.5	209.5	101.6	173	123	50
disgust-picture	275.2	3264.1	400	137.9	262.1	102.3	157	82	75
disgust-average	269.4	2384.4	400	167.4	232.5	102.1	151.3	96	55.3
fear-city	417.0	8986.7	500	235.3	264.7	102.8	59	27	32
fear-moved	357.2	7145.5	500	160	340	102.6	89	53	36
fear-picture	388.2	8830.9	500	160	340	103.6	86	41	45
fear-average	387.4	8321.0	500	185.1	314.9	103.0	78	40.3	37.6
happy-city	388.3	5810.6	500	285.7	214.3	106.6	71	54	17
happy-moved	348.2	6188.8	500	173.9	326.1	109.2	109	78	31
happy-picture	357.7	6038.3	500	266.7	233.3	106.0	100	57	43
happy-average	364.7	6012.6	500	242.1	257.9	107.2	93.3	63	30.3
sad-city	279.8	77.9	285.7	266.7	19	98.6	88	62	26
sad-moved	276.9	90.7	285.7	266.7	19	99.1	144	93	51
sad-picture	275.5	127.2	285.7	250	35.7	98.3	138	83	55
sad-average	277.4	98.6	285.7	261.1	24.5	98.7	123.3	79.3	44
surprise-city	394.3	8219.4	500	148.1	351.9	107.5	69	49	20
surprise-moved	360.3	7156.0	500	160	340	107.8	101	84	17
surprise-picture	371.6	8355.7	500	285.7	214.3	106.7	98	54	44
surprise-average	375.4	7910.4	500	197.9	302.0	107.3	89.3	62.3	27

Figure 12-9: Table of acoustic features for the three utterances.

The results are summarized in table 12-9. The values for each feature are displayed for each phrase with each emotive quality (including the neutral state). The averages are also presented in the table and plotted in 12-10. These plots easily illustrate the relationship of how each emotive quality modulates these acoustic features with respect to one another. The pitch contours for each emotive quality are shown in figure 12-11. They correspond to the utterance "It's been moved already".

Relating these plots with table 12-1, we can see that many of the acoustic correlates of emotive speech are preserved in Kismet's speech. We have made several incremental adjustments to the qualities of Kismet's speech according to what we have learned from subject evaluations. Our final implementation differs in some cases from table 12-1 (as noted below), but the results show a dramatic improvement in subject recognition performance from earlier evaluations.

- Fearful speech is very fast with wide pitch contour, large pitch variance, very high mean pitch, and normal intensity. We have added a slightly breathy quality to the voice as people seem to associate it with a sense of trepidation.

- Angry speech is slightly fast with a wide pitch range, high variance, and is loud. We've purposefully implemented a low mean pitch to give the voice a prohibiting

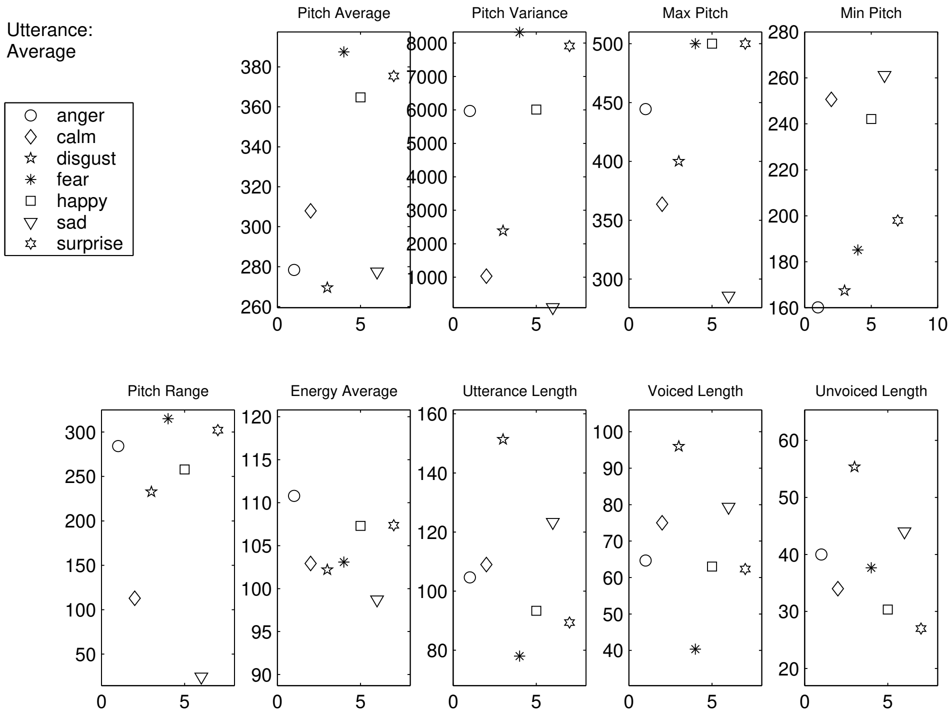


Figure 12-10: Plots of acoustic features of Kismet's speech. Each plot illustrates how each emotion relates to the others for each acoustic feature. The horizontal axis simply maps an integer value to each emotion for ease of viewing (anger=1, calm=2, etc.)

quality. This differs from table 12-1, but a preliminary study demonstrated a dramatic improvement in recognition performance of naive subjects. This makes sense as it gives the voice a threatening quality.

- Sad speech has a slower speech rate, with longer pauses than normal. It has a low mean pitch, a narrow pitch range and low variance. It is softly spoken with a slight breathy quality. This differs from table 12-1, but it gives the voice a tired quality. It has a pitch contour that falls at the end.

- Happy speech is relatively fast, with a high mean pitch, wide pitch range, and wide pitch variance. It is loud with smooth undulating inflections as shown in figure 12-11.

- Disgusted speech is slow with long pauses interspersed. It has a low mean pitch with a slightly wide pitch range. It is fairly quiet with a sort of creaky quality to the voice. The contour has a globally falling downward slope as shown in figure 12-11.

- Surprised speech is fast with a high mean pitch and wide pitch range. It is fairly loud with a steep rising contour on the stressed syllable of the final word.

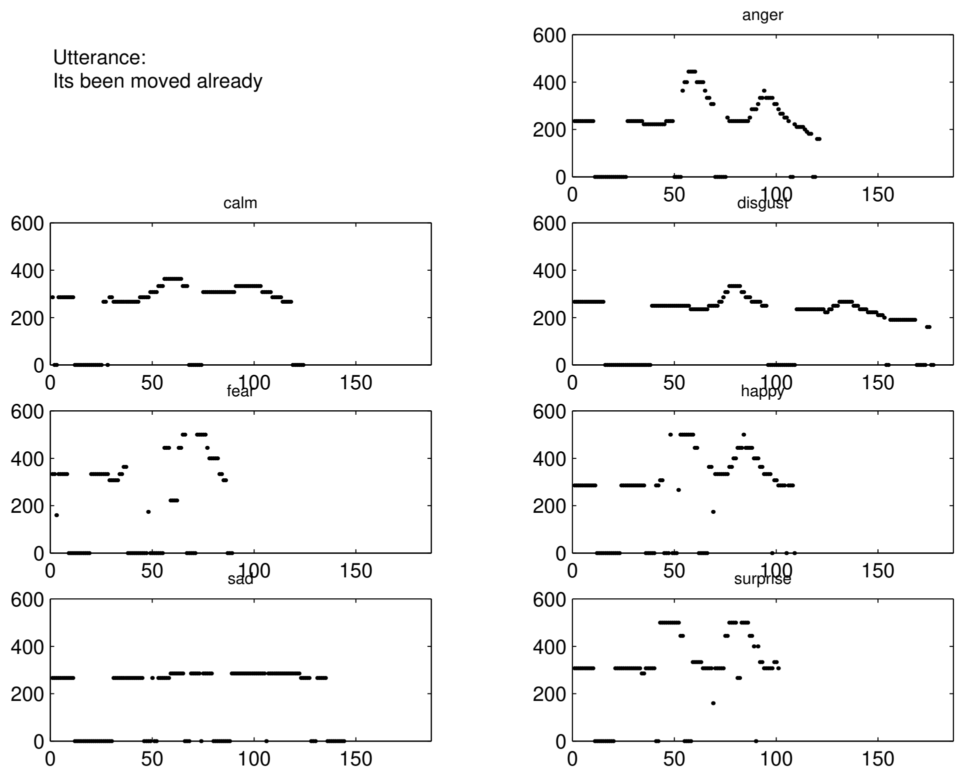


Figure 12-11: Pitch analysis of Kismet's speech for the English phrase "It's been moved already".

12.6.2 Human Listener Experiments

To evaluate Kismet's expressive speech, nine subjects were asked to listen to pre-recorded utterances and to fill out a forced choice questionnaire. Subjects ranged from 23 to 54 years of age, all affiliated with MIT. The subjects had very limited to no familiarity with Kismet's voice.

In this study, each subject first listened to an introduction spoken with Kismet's neutral expression. This was to acquaint the subject with Kismet's synthesized quality of voice and neutral affect. A series of eighteen utterances followed, covering six expressive qualities (anger, fear, disgust, happiness, surprise, and sorrow). Within the experiment, the emotive qualities were distributed randomly. Given the small number of subjects per study, we only used a single presentation order per experiment. Each subject could work at his/her own pace and control the number of presentations of each stimulus.

The three stimulus phrases were: "I'm going to the city", "I saw your name in the paper", and "It's happening tomorrow". The first two test phrases were selected because Cahn had found the word choice to have reasonably neutral affect. In a previous version of the study, subjects reported that it was just as easy to map emotional correlates onto English phrases as to Kismet's randomly generated babbles. Their performance for English phrases and Kismet's babbles supports this. We believed it would be easier to analyze the data to discover ways to improve Kismet's performance

forced choice percentage (random=17%)

	anger	disgust	fear	happy	sad	surprise	% correct
anger	75	15	0	0	0	10	75/100
disgust	21	50	4	0	25	0	50/100
fear	4	0	25	8	0	63	25/100
happy	0	4	4	67	8	17	67/100
sad	8	8	0	0	84	0	84/100
surprise	4	0	25	8	4	59	59/100

Figure 12-12: Naive subjects assessed the emotion conveyed in Kismet's voice in a forced choice evaluation. All emotional qualities were recognized with reasonable performance except for "fear" which was most often confused for "surprise/excitement". Both expressive qualities share high arousal, so the confusion is not unexpected. To improve the recognition of "fear", irregular pausing (as Cahn implemented) should be incorporated. However, this involves grammatical analysis of the sentence. Kismet only babbles for now, so there is no sentence structure to analyze at this time.

if a small set of fixed English phrases were used. However, during run-time Kismet only babbles – it does not vocalize English phrases.

The subjects were simply asked to circle the word which best described the voice quality. The choices were "anger", "disgust", "fear/panic", "happy", "sad", "surprise/excited". From a previous iteration of the study we found that word choice mattered. A given emotion category can have a wide range of vocal affects. For instance, the subject could interpret "fear" to imply "apprehensive", which might be associated with Kismet's whispery vocal expression for sadness. Alternatively, it could be associated with "panic" which is a more aroused interpretation. The results from these evaluations are summarized in table 12-12.

Overall, the subjects exhibited reasonable performance in correctly mapping Kismet's expressive quality with the targeted emotion. However, the expression of "fear" proved problematic. For all other expressive qualities, the performance was significantly above random. Furthermore, misclassifications were highly correlated to similar emotions. For instance, "anger" was sometimes confused with "disgust" (sharing negative valence) or "surprise/excitement" (both sharing high arousal). "Disgust" was confused with other negative emotions. "Fear" was confused with other high arousal emotions (with "surprise/excitement" in particular). The distribution for "happy" was more spread out, but it was most often confused with "surprise/excitement", with which it shares high arousal. Kismet's "sad" speech was confused with other negative emotions. The distribution for "surprise/excitement" was broad, but it was most often confused for "fear".

Since this study, we have adjusted the vocal affect parameter values to improve the distinction between "fear" and "surprise". We have given Kismet's fearful affect a more apprehensive quality by lowering the volume and giving the voice a slightly raspy quality (this was the version that was analyzed in section 12.6.1). In a previous study we found that people often associated the raspy vocal quality with whispering and apprehension. We have also enhanced "surprise" by increasing the amount of stress rise on the stressed syllable of the final word. Cahn analyzed the sentence structure to introduce irregular pauses into her implementation of "fear". This makes a significant contribution to the interpretation of this emotional state. However, in practice Kismet only babbles, so modifying the pausing via analysis of sentence structure is premature as sentences do not exist.

Given the number and homogeneity of our subjects, we cannot make strong claims regarding Kismet's ability to convey emotion through expressive speech. More extensive studies need to be carried out. However, for the purposes of evaluation, the current set of data is promising. Misclassifications are particularly informative. The mistakes are highly correlated with similar emotions, which suggests that arousal and valence are conveyed to people (arousal being more consistently conveyed than valence). We are using the results of this study to improve Kismet's expressive qualities. In addition, Kismet expresses itself through multiple modalities, not just through voice. Kismet's facial expression and body posture should help resolve the ambiguities encountered through voice alone.

12.7 Real-Time Lip Synchronization and Facial Animation

Given Kismet's ability to express itself vocally, it is important that the robot also be able to support this vocal channel with coordinated facial animation. This includes synchronized lip movements to accompany speech along with facial animation to lend additional emphasis to the stressed syllables. These complementary motor modalities greatly enhance the robot's delivery when it speaks, giving the impression that the robot "means" what it says. This makes the interaction more engaging for the human and facilitates proto-dialog.

12.7.1 Guidelines from Animation

The earliest examples of lip synchronization for animated characters dates back to the 1940's in classical animation (Blair 1949), and back to the 1970s for computer animated characters (Parke 1972). In these early works, all of the lip animation was crafted by hand (a very time consuming process). Over time, a set of guidelines evolved that are largely adhered to by animation artists today (Madsen 1969).

According to Madsen, *simplicity is the secret to successful lip animation.* Extreme accuracy for cartoon animation often looks forced or unnatural. Thus, the goal in animation is not to always imitate realistic lip motions, but *to create a visual shorthand that passes unchallenged by the viewer* (Madsen 1969). However, as the realism of the character increases, the accuracy of the lip synchronization follows.

Kismet is a fanciful and cartoon-like character, so the guidelines for cartoon animation apply. In this case, the guidelines suggest that the animator focus on vowel lip motions (especially *o* and *w*) accented with consonant postures (*m, b, p*) for lip closing. Precision of these consonants gives credibility to the generalized patterns of vowels. The transitions between vowels and consonants should be reasonable approximations of lip and jaw movement. Fortunately, more latitude is granted for more fanciful characters. The mechanical response time of Kismet's lip and jaw motors places strict constraints on how fast the lips and jaw can transition from posture to posture. Madsen also stresses that care must be taken in conveying emotion, as the expression of voice and face can change dramatically.

12.7.2 Extracting Lip Synch Info

To implement lip synchronization on Kismet, a variety of information must be computed in real-time from the speech signal. By placing DECtalk in *memory mode* and issuing the command string (utterance with synthesizer settings), the DECtalk software generates the speech waveform and writes it to memory (a 11.025 kHz waveform). In addition, DECtalk extracts time-stamped phoneme information. From the speech waveform, we compute its time-varying energy over a window size of 335 samples. We take care to synchronize the phoneme and energy information, and send $(phoneme(t), energy(t))$ pairs to the QNX machine at 33 Hz to coordinate jaw and lip motor control. A similar technique using DECtalk's phoneme extraction capability is reported by (Waters & Levergood 1993) for real-time lip synchronization for computer generated facial animation.

To control the jaw, the QNX machine receives the phoneme and energy information and updates the commanded jaw position at 10 Hz. The mapping from energy to jaw opening is linear, bounded within a range where the minimum position corresponds to a closed mouth, and the maximum position corresponds to an open mouth characteristic of surprise. Using only energy to control jaw position produces a lively

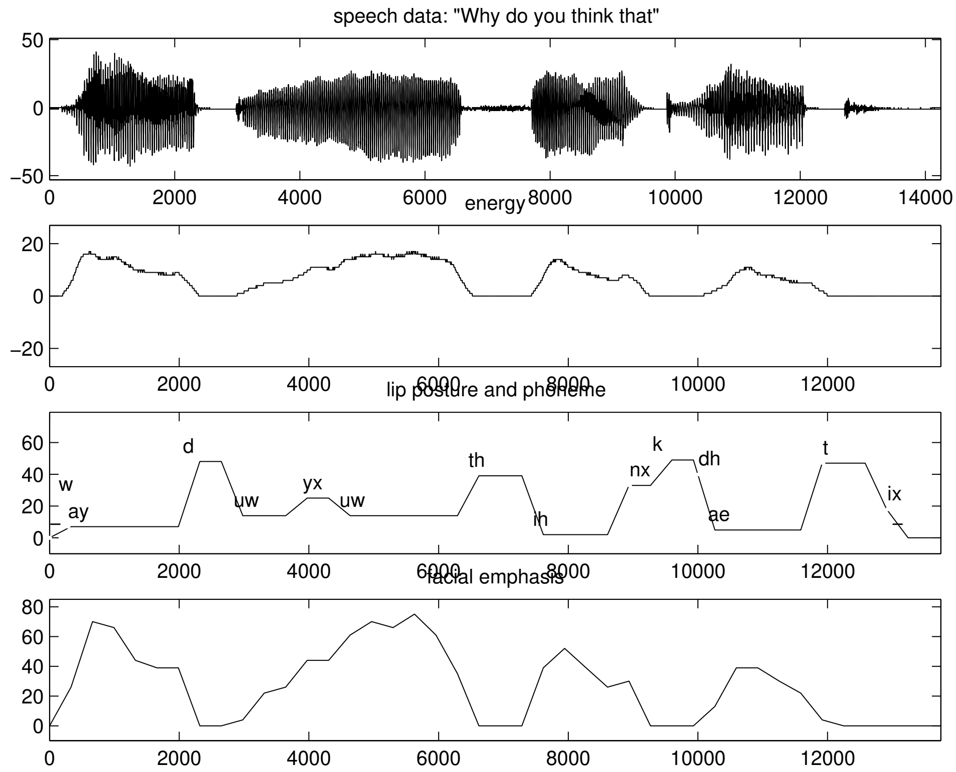


Figure 12-13: Plot of speech signal, energy, phonemes/lip posture, and facial emphasis for the phrase "Why do you think that". Time is in 0.1 ms increments. The total amount of time to vocalize the phrase is 1.4 sec.

effect but has its limitations (Parke & Waters 1996). For Kismet, the phoneme information is used to make sure that the jaw is closed when either a m, p, or b is spoken or there is silence. This may not necessarily be the case if only energy were used.

Upon receiving the phoneme and energy information from the vocalization system, the QNX *vocal communication* process passes this information to the motor skill system via the DPRAM. The motor skill system converts the energy information into a measure of facial emphasis (linearly scaling the energy), which is then passed onto the lip synchronization and facial animation processes of the face control motor system. The motor skill system also maps the phoneme information onto lip postures and passes this information to the *lip synchronization* and *facial animation* processes of the motor system that controls the face (described in chapter 11). Figure 12-13 illustrates the stages of computation from the raw speech signal to lip posture, jaw opening, and facial emphasis.

The computer network involved in lip synchronization is a bit convoluted, but supports real-time performance. Figure 12-14 illustrates the information flow through the system and denotes latencies. Within the NT machine, there is a latency of approximately 250 ms from the time the synthesizer generates the speech signal and extracts phoneme information until that speech signal is sent to the sound card. Immediately following the generation and feature extraction phase, the NT machine

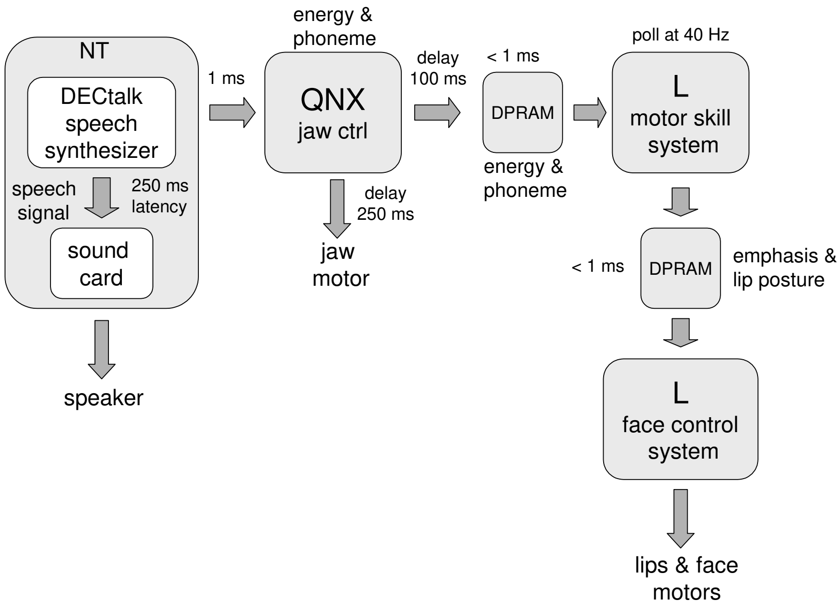


Figure 12-14: Schematic of the flow of information for lip synchronization. This figure illustrates the latencies of the system and the compensatory delays to maintain synchrony.

sends this information to the QNX node that controls the jaw motor. The latency of this stage is less than 1 ms. Within QNX, the energy signal and phoneme information is used to compute the jaw position. To synchronize jaw movement with sound production from the sound card, the jaw command position is delayed by 250 ms. For the same reason, the QNX machine delays the transfer of energy and phoneme information by 100 ms to the L based machines. Dual-ported RAM communication is sub-millisecond. The lip synchronization processes running on *L* polls and updates their energy and phoneme values at 40 Hz, much faster than the phoneme information is changing and much faster than the actuators can respond. Energy is scaled to control the amount of facial emphasis, and the phonemes are mapped to lip postures. The lip synchronization performance is well coordinated with speech output since the delays and latencies are fairly consistent.

Kismet's ability to lip-sync within its limits greatly enhances the perception that it is genuinely talking (instead of some disembodied speech system). It also contributes to the life-like quality and charm of the robot's behavior.

Figure 12-15 shows how the fifty DECtalk phonemes are mapped to Kismet's lip postures. Kismet obviously has a limited repertoire as it cannot make many of the lip movements that humans do. For instance, it cannot protrude its lips (important for *sh* and *ch* sounds), nor does it have a tongue (important for *th* sounds), nor teeth. However, computer animated lip synchronization often maps the 45 distinct English

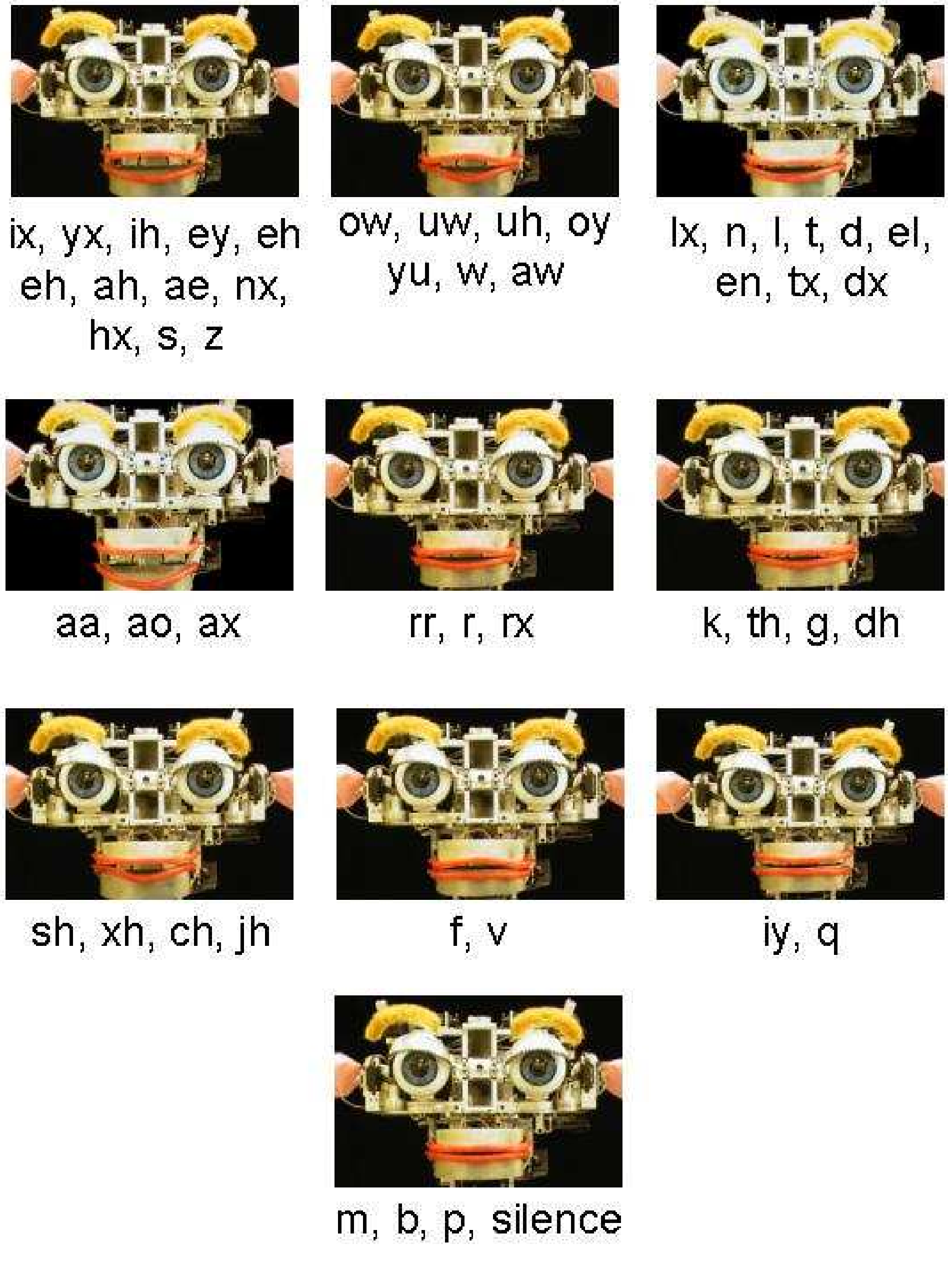


Figure 12-15: Kismet's mapping of lip postures to phonemes.

phonemes onto a much more restricted set of visually distinguishable lip postures; eighteen is preferred (Parke & Waters 1996). The more fanciful the character, the more latitude one is granted in the accuracy of lip animation. For cartoon characters, a subset of ten lip and jaw postures is enough for reasonable artistic conveyance (Fleming & Dobbs 1999). Kismet's ten lip postures tend toward the absolute minimal set specified by (Fleming & Dobbs 1999), but is reasonable given its physical appearance. As the robot speaks, new lip posture targets are specified at 33 Hz. Since the phonemes do not change this quickly, many of the phonemes repeat. There is an inherent limit in how fast Kismet's lip and jaw motors can move to the next commanded, so the challenge of co-articulation is somewhat addressed of by the physics of the motors and mechanism.

Lip synchronization is only part of the equation, however. Faces are not completely still when speaking, but move in synchrony to provide emphasis along with the speech. Using the energy of the speech signal to animate Kismet's face (along with the lips and jaw) greatly enhances the impression that Kismet "means" what it says. For Kismet, the energy ofthe speech signal influences the movement of its eyelids and ears. Larger speech amplitudes result in a proportional widening of the eyes and downward pulse of the ears. This adds a nice degree of facial emphasis to accompany the stress of the vocalization.

Kismet expresses itself through face as well as voice. Since the speech signal influences facial animation, the emotional correlates of facial posture must be blended with the animation arising from speech. How this is accomplished within the face control motor system is described at length in chapter 11. The emotional expression establishes the baseline facial posture about which all facial animation moves. The current emotional state also influences the speed with which the facial actuators move (lower arousal results in slower movements, higher arousal results in quicker movements). In addition, emotions that correspond to higher arousal produce more energetic speech, resulting in bigger amplitude swings about the expression baseline. Similarly, emotions that correspond to lower arousal produce less energetic speech, which results in smaller amplitudes. The end product is a highly expressive and coordinated movement of face with voice. For instance, angry speech is accompanied by large and quick twitchy movements of the ears eyelids. This undeniably conveys agitation and irritation. In contrast, sad speech is accompanied by slow, droopy, listless movements of the ears and eyelids. This conveys a forlorn quality that often evokes sympathy from the human observer.

12.8 Limitations and Extensions

Improving Expressive Speech

Kismet's expressive speech can certainly be improved. In the current implementation we have only included those acoustic correlates that have a global influence on the speech signal and do not require local analysis of the sentence structure. We currently modulate voice quality, speech rate, pitch range, average pitch, intensity, and the global pitch contour. Our data from naive subjects is promising, although we could certainly do more. We have done very little with changes in articulation. We could enhance the precision or imprecision of articulation by substituting voiced for unvoiced phonemes as Cahn describes in her thesis. By analyzing sentence structure, several more influences can be introduced. For instance, carefully selecting the types of stress placed on emphasized and de-emphasized words, as well as introducing different kinds of pausing, can be used to strengthen the perception of negative emotions such as fear, sadness, and disgust. However, given our immediate goal of proto-language, there is no sentence structure to analyze. Nonetheless, to extend Kismet's expressive abilities to English sentences, the grammatical and lexical constraints must be carefully considered.

On a slightly different vein, we could also introduce emotive sounds such as laughter, cries, coos, gurgles, screams, shrieks, yawns, and so forth. DECtalk supports the ability to play pre-recorded sound files. We could modulate an initial set of emotive sounds to add variability.

Extensions to Utterance Generation

Kismet's current manner of speech has wide appeal to those who have interacted with the robot. There is sufficient variability in phoneme, accent, and end syntax choice to permit an engaging proto-dialog. If Kismet's utterance has the intonation of a question, people will treat it as such – often "re-stating" the question as an English sentence and then answering it. If Kismet's intonation has the intonation of a statement, they respond accordingly. They may say something such as "Oh, I see" or perhaps issue another query such as "So then what did you do?" The utterances are complex enough to sound as if the robot is speaking a different language.

Nonetheless, the current utterance generation algorithm is really intended as a place holder for a more sophisticated generation algorithm. There is interest in computationally modeling canonical babbling so that the robot makes vocalizations characteristic of an eight-month-old child (de Boysson-Bardies 1999). This would significantly limit the range of the utterances the robot currently produces, but would facilitate the acquisition of proto-language. Kismet varies many parameters at once, so the learning space is quite large. By modeling canonical babbling, the robot can systematically explore how a limited set of parameters modulates the way its voice sounds. Introducing variations upon a theme during vocal games with the caregiver as well as on its own could simplify the learning process (see chapters 2 and 1). By interfacing what the robot vocally generates with what it hears, the robot could begin to explore its vocal capabilities, how to produce targeted effects, and how these utterances influence the caregiver's behavior.

Improvements to Lip Synchronization

Kismet's lip synchronization and facial animation are compelling and well matched to Kismet's behavior and appearance. However, we can improve upon the current implementation and extend it in a couple of ways. First, we could certainly reduce the latencies throughout the system. This would give us tighter synchronization. We could incorporate higher performance actuators to allow a faster response time. This would also support more precise lip synchronization.

If we wanted more realism, we could add a tongue, teeth, and lips that could move more like those of a human. However this degree of realism is unnecessary for our purposes, and it is tremendously difficult to achieve the required amount of realism. As it stands, Kismet's lip synchronization is a successful shorthand that goes unchallenged by the viewer.

12.9 Summary

We have implemented an expressive vocalization system that supports novel utterances. In doing so, we have addressed issues regarding the expressiveness and richness of Kismet's vocal modality, and how it supports social interaction. We have found that the vocal utterances are rich enough to facilitate interesting proto-dialogs with people. We have found the expressiveness of the voice to be reasonably identifiable. Furthermore, the robot's speech is complemented by real-time animated facial animation which enhances delivery. Instead of trying to achieve realism, we have implemented a system that is well matched with the robot's appearance and capabilities. The end result is a well orchestrated and compelling synthesis of voice, facial animation, and affect that make a significant contribution to the expressiveness and personality of the robot.

Chapter 13

Social Constraints on Animate Vision

The control of animate vision for a social robot poses challenges beyond issues of stability and accuracy, as well as advantages beyond computational efficiency and perceptual robustness (Ballard 1989). We have found that Kismet's human-like eye movements have high communicative value to the people that interact with it. Hence the challenge of interacting with humans constrains how Kismet appears physically, how it moves, how it perceives the world, and how its behaviors are organized. This chapter describes Kismet's integrated visual-motor system that must negotiate between the physical constraints of the robot, the perceptual needs of the robot's behavioral and motivational systems, and the social implications of motor acts. It presents those systems responsible for generating Kismet's compelling visual behavior.

13.1 Human Visual Behavior

From a social perspective, human eye movements have a high communicative value (as illustrated in figure 13-1). For example, gaze direction is a good indicator of the locus of visual attention. We have discussed this at length in chapter 6. Knowing a person's locus of attention reveals what that person currently considers behaviorally relevant, which is in turn a powerful clue to their intent. The dynamic aspects of eye movement, such as staring versus glancing, also convey information. Eye movements are particularly potent during social interactions, such as conversational turn-taking, where making and breaking eye contact plays an important role in regulating the exchange. We model the eye movements of our robots after humans, so that they may have similar communicative value.

From a functional perspective, the human system is so good at providing a stable percept of the world that we have no intuitive appreciation of the physical constraints under which it operates. Fortunately, there is a wealth of data and proposed models for how the human visual system is organized (Kandel et al. 2000). This data provides not only a modular decomposition but also mechanisms for evaluating the performance of the complete system.

Kismet's visual-motor control is modeled after the human oculo-motor system. By doing so, we hope to harness both the computational efficiency and perceptual robustness advantages of an animate vision system, as well as the communicative

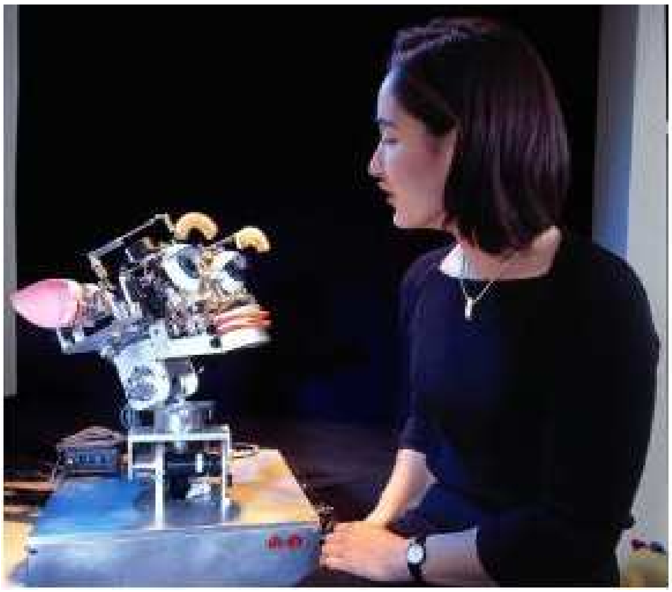


Figure 13-1: Kismet is capable of conveying intentionality through facial expressions and behavior. Here, the robot's physical state expresses attention to and interest in the human beside it. Another person - for example, the photographer - would expect to have to attract the robot's attention before being able to influence its behavior.

power of human eye movements. In this section we briefly survey the key aspects of the human visual system that we used as a guideline to design Kismet's visual apparatus and eye movement primitives.

Foveate Vision

Humans have *foveate* vision. The fovea (the center of the retina) has a much higher density of photoreceptors than the periphery. This means that to see an object clearly, humans must move their eyes such that the image of the object falls on the fovea. The advantage of this receptor layout is that humans enjoy both a wide peripheral field of view as well as high acuity vision. The wide field of view is useful for directing visual attention to interesting features in the environment that may warrant further detailed analysis. This analysis is performed while directing gaze to that target and using foveal vision for detailed processing over a localized region of the visual field.

Vergence Movements

Humans have binocular vision. The visual disparity of the images from each eye give humans one visual cue to perceive depth (humans actually use multiple cues (Kandel et al. 2000)). The eyes normally move in lock-step, making equal, *conjunctive* movements. For a close object, however, the eyes need to turn towards each other somewhat to correctly image the object on the foveae of the two eyes. These disjunctive movements are called vergence, and rely on depth perception (see figure 13-2).

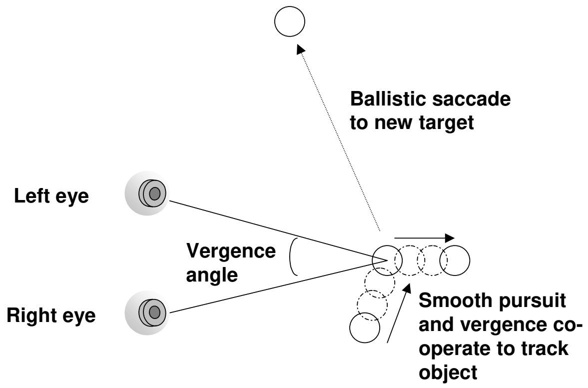


Figure 13-2: Humans exhibit four characteristic types of eye motion. Saccadic movements are high-speed ballistic motions that center a target in the field of view. Smooth pursuit movements are used to track a moving object at low velocities. The vestibulo-ocular reflex (VOR) and opto-kinetic nystagmus (OKN) act to maintain the angle of gaze as the head and body move through the world (not shown). Vergence movements serve to maintain an object in the center of the field of view of both eyes as the object moves in depth.

Saccades

Human eye movement is not smooth. It is composed of many quick jumps, called *saccades*, which rapidly re-orient the eye to project a different part of the visual scene onto the fovea. After a saccade, there is typically a period of fixation, during which the eyes are relatively stable. They are by no means stationary, and continue to engage in corrective micro-saccades and other small movements. Periods of fixation typically end after some hundreds of milliseconds, after which a new saccade will occur.

Smooth Pursuit

However, if the eyes fixate on a moving object, they can follow it with a continuous tracking movement called *smooth pursuit*. This type of eye movement cannot be evoked voluntarily, but only occurs in the presence of a moving object.

Vestibulo-Ocular Reflex and Opto-Kinetic Response

Since eyes also move with respect to the head, they need to compensate for any head movements that occur during fixation. The *vestibulo-ocular reflex* (VOR) uses inertial feedback from the vestibular system to keep the orientation of the eyes stable as the eyes move. This is a very fast response, but is prone to the accumulation of error over time. The *opto-kinetic nystagmus* (OKN) is a slower compensation mechanism that uses a measure of the visual slip of the image across the retina to correct for drift. These two mechanisms work together to give humans stable gaze as the head moves.

We have endowed Kismet with visual perception and visual motor abilities that are human-like in their physical implementation. Our hope is that by following the example of the human visual system, the robot's behavior will be easily understood because it is analogous to the behavior of a human in similar circumstances. For example, when an anthropomorphic robot moves its eyes and neck to orient toward an object, an observer can effortlessly conclude that the robot has become interested in that object (as discussed in chapter 6). These traits lead not only to behavior that is easy to understand, but also allows the robot's behavior to fit into the social norms that the person expects.

Another advantage is robustness. A system that integrates action, perception, attention, and other cognitive capabilities can be more flexible and reliable than a system that focuses on only one of these aspects. Adding additional perceptual capabilities and additional constraints between behavioral and perceptual modules can increase the relevance of behaviors while limiting the computational requirements. For example, in isolation, two difficult problems for a visual tracking system are knowing what to track and knowing when to switch to a new target. These problems can be simplified by combining the tracker with a visual attention system that can identify objects that are behaviorally relevant and worth tracking. In addition, the tracking system benefits the attention system by maintaining the object of interest in the center of the visual field. This simplifies the computation necessary to implement behavioral habituation. These two modules work in concert to compensate for the deficiencies of the other and to limit the required computation in each.

Using the human visual system as a model, we can specify a set of design criteria for Kismet's visual system. These criteria not only address performance issues, but aesthetic issues as well. The importance of functional aesthetics for performance as well as social constraints has been discussed in depth in chapter 4.

13.1.1 Similar Visual Morphology

Special attention has been paid to balancing the functional and aesthetic aspects of Kismet's camera configuration. From a functional perspective, the cameras in Kismet's eyes have high acuity but a narrow field of view. Between the eyes, there are two unobtrusive central cameras fixed with respect to the head, each with a wider field of view but correspondingly lower acuity.

The reason for this mixture of cameras is that typical visual tasks require both high acuity and a wide field of view. High acuity is needed for recognition tasks and for controlling precise visually guided motor movements. A wide field of view is needed for search tasks, for tracking multiple objects, compensating for involuntary ego-motion, etc. As described earlier, a common trade-off found in biological systems is to sample part of the visual field at a high enough resolution to support the first set of tasks, and to sample the rest of the field at an adequate level to support the second set. This is seen in animals with foveal vision, such as humans, where the density of photoreceptors is highest at the center and falls off dramatically towards the periphery. This can be implemented by using specially designed imaging hardware (van der Spiegel, Kreider, Claeys, Debusschere, Sandini, Dario, Fantini, Belluti & Soncini 1989), (Kuniyoshi, Kita, Sugimoto, Nakamura & Suehiro 1995), space-variant image sampling (Bernardino & Santos-Victor 1999), or by using multiple cameras with different fields of view, as we have done.

Aesthetically, Kismet's big blue eyes are no accident. The cosmetic eyeballs envelop the fovea cameras and greatly enhance the readability of Kismet's gaze. The pair of minimally obtrusive wide field of view cameras that move with respect to the head are no accident, either. We did not want their size or movement to distract from Kismet's gaze. By doing so, people's attention is drawn to Kismet's eyes where powerful social cues are conveyed.

13.1.2 Similar Visual Perception

For robots and humans to interact meaningfully, it is important that they understand each other enough to be able to shape each other's behavior. This has several implications. One of the most basic is that robot and human should have at least some overlapping perceptual abilities (see chapter 5). Otherwise, they can have little idea of what the other is sensing and responding to. However, similarity of perception requires more than similarity of sensors. Not all sensed stimuli are equally behaviorally relevant. It is important that both human and robot find the same types of stimuli salient in similar conditions. For this reason, Kismet is designed to have a set of perceptual biases based on the human pre-attentive visual system. We have discussed this issue at length in chapter 6.

13.1.3 Similar Visual Attention

Visual perception requires high bandwidth and is computationally demanding. In the early stages of human vision, the entire visual field is processed in parallel. Later computational steps are applied much more selectively, so that behaviorally relevant parts of the visual field can be processed in greater detail. This mechanism of visual attention is just as important for robots as it is for humans, from the same considerations of resource allocation. The existence of visual attention is also key to satisfying the expectations of humans concerning what can and cannot be perceived visually. We have implemented a context-dependent attention system that goes some way towards this as presented in chapter 6.

13.1.4 Similar Eye Movements

Kismet's visual behaviors address both functional and social issues. From a functional perspective, we have implemented a set of human-like visual behaviors to allow the robot to process the visual scene in a robust and efficient manner. These include saccadic eye movements, smooth pursuit, target tracking, gaze fixation, and ballistic head-eye orientation to target. We have also implemented two visual responses that very roughly approximate the function of the VOR (however, the current implementation does not employ a vestibular system), and the OKN. Due to human sensitivity to gaze, it is absolutely imperative that Kismet's eye movements look natural. Quite frankly, people find it disturbing if they move in a non-human manner.

Kismet's rich visual behavior can be conceptualized on those four levels presented in chapter 10. Namely, the social level, the behavior level, the skills level, and the primitives level. We have already argued how human-like visual behaviors have high communicative value in different social contexts. Higher levels of motor control address these social issues by coordinating the basic visual motor primitives (saccade, smooth pursuit, etc.) in a socially appropriate manner. We describe these levels in detail below, starting at the lowest level (the oculo-motor level), and progressing to the highest level where we discuss the social constraints of animate vision.

13.2 The Oculo-Motor System

Our implementation of an oculo-motor system is an approximation of the human system. The system has been a large scale engineering effort with Brian Scassellati and Paul Fitzpatrick making substantial contributions (Breazeal & Scassellati 1999*a*), (Breazeal et al. 2000). The motor primitives are organized around the needs of higher levels, such as maintaining and breaking mutual regard, performing visual search, etc. Since our motor primitives are tightly bound to visual attention, we will first briefly survey their sensory component.

13.2.1 Low-Level Visual Perception

Recall from chapter 5 and chapter 6, we have implemented a variety of perceptual feature detectors that are particularly relevant to interacting with people and objects. These include low-level feature detectors attuned to quickly moving objects, highly saturated color, and colors representative of skin tones. Looming and threatening objects are also detected pre-attentively, to facilitate a fast reflexive withdrawal (see chapter 6).

13.2.2 Visual Attention

Recall from chapter 6, we have implemented Wolfe's model of human visual search and have supplemented it to operate in conjunction with time-varying goals, with moving cameras, and to address the issue of habituation. This combination of top-down and bottom-up contributions allows the robot to select regions that are visually salient and behaviorally relevant to direct its computational and behavioral resources towards those regions. The attention system runs all the time, even when it is not controlling gaze, since it determines the perceptual input to which the motivational and behavioral systems respond.

13.2.3 Consistency of Attention

In the presence of objects of similar salience, it is useful be able to commit attention to one of the objects for a period of time. This gives time for post-attentive processing to be carried out on the object, and for downstream processes to organize themselves around the object. As soon as a decision is made that the object is not behaviorally relevant (for example, it may lack eyes, which are searched for post-attentively), attention can be withdrawn from it and visual search may continue. Committing to an object is also useful for behaviors that need to be atomically applied to a target (for example, a calling behavior where the robot needs to stay looking at the person it is calling).

To allow such commitment, the attention system is augmented with a tracker. The tracker follows a target in the wide visual field, using simple correlation between successive frames. Usually changes in the tracker target will be reflected in movements of the robot's eyes, unless this is behaviorally inappropriate. If the tracker loses the target, it has a very good chance of being able to reacquire it from the attention system. Figure 13-3 shows the tracker in operation.

13.2.4 Post-Attentive Processing

Once the attention system has selected regions of the visual field that are potentially behaviorally relevant, more intensive computation can be applied to these regions than could be applied across the whole field. Searching for eyes is one such task. Locating eyes is important to us for engaging in eye contact, and as a reference point for interpreting facial movements and expressions. We currently search for eyes after the robot directs its gaze to a locus of attention, so that a relatively high resolution image of the area being searched is available from the foveal cameras (recall chapter 6). Once the target of interest has been selected, we also estimate its proximity to the robot using a stereo match between the two central wide cameras (also discussed in chapter 6). Proximity is important for interaction as things closer to the robot should be of greater interest. It's also useful for interaction at a distance, such as a person standing too far for face to face interaction but close enough to be beckoned closer. Clearly the relevant behavior (calling or playing) is dependent on the proximity of the human to the robot.

13.2.5 Eye Movements

Figure 13-4 shows the organization of Kismet's eye/neck motor control. Kismet's eyes periodically saccade to new targets chosen by an attention system, tracking them smoothly if they move and the robot wishes to engage them. Vergence eye movements

Figure 13-3: Behavior of the tracker. Frames are taken at one-second intervals. The white squares indicate the position of the target. The target is not centered in the images since they were taken from a camera fixed with respect to the head, rather than gaze direction. On the third row, the face slips away from the tracker, but it is immediately reacquired through the attention system. The images are taken from a three-minute session during which the tracker slipped five times. This is typical performance for faces, which tend not to move too rapidly.

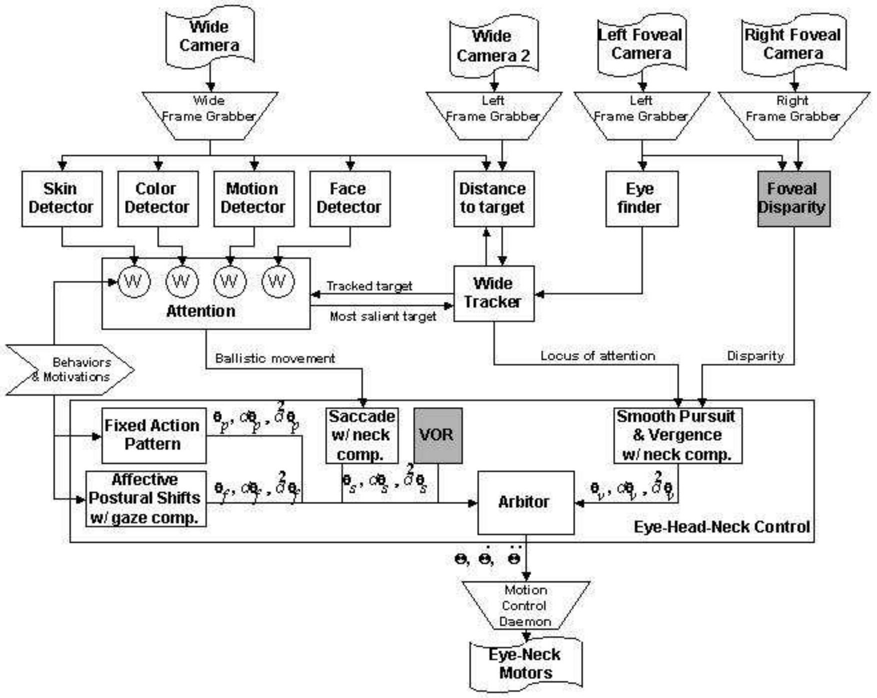


Figure 13-4: Organization of Kismet's eye/neck motor control. Many cross level influences have been omitted. The modules in gray are not active in the results presented in this chapter.

are more challenging to implement in a social setting, since errors in disjunctive eye movements can give the eyes a disturbing appearance of moving independently. Errors in conjunctive movements have a much smaller impact on an observer, since the eyes clearly move in lock-step. A crude approximation of the opto-kinetic reflex is rolled into our implementation of smooth pursuit. Kismet uses an efferent copy mechanism to compensate the eyes for movements of the head.

The attention system operates on the view from the central camera. A transformation is needed to convert pixel coordinates in images from this camera into position set-points for the eye motors. This transformation in general requires the distance to the target to be known, since objects in many locations will project to the same point in a single image (see Figure 13-5). Distance estimates are often noisy, which is problematic if the goal is to center the target exactly in the eyes. In practice, it is usually enough to get the target within the field of view of the foveal cameras in the eyes. Clearly the narrower the field of view of these cameras is, the more accurately the distance to the object needs to be known. Other crucial factors are the distance between the wide and foveal cameras, and the closest distance at which the robot will need to interact with objects. These constraints are determined by the physical distribution of Kismet's cameras and the choice of lenses. The central location of the wide camera places it as close as possible to the foveal cameras. It also has the advantage that moving the head to center a target (as seen in the central camera) will in fact truly orient the head towards that target. For cameras in other locations, accuracy

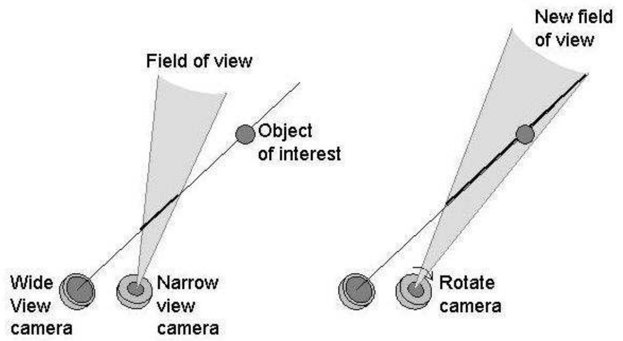


Figure 13-5: Without distance information, knowing the position of a target in the wide camera only identifies a ray along which the object must lie, and does not uniquely identify its location. If the cameras are close to each other (relative to the closest distance the object is expected to be at) the foveal cameras can be rotated to bring the object within their narrow field of view without needing an accurate estimate of its distance. If the cameras are far apart, or the field of view is very narrow, the minimum distance at which the object can be becomes large. We use the former solution in Kismet.

of orientation would be limited by the accuracy of the distance measurement.

Higher-level influences modulate the movement of the neck and eyes in a number of ways. As already discussed, modifications to weights in the attention system translate to changes of the locus of attention about which eye movements are organized. The overall posture of the robot can be controlled in terms of a three-dimensional affective space (chapter 11). The regime used to control the eyes and neck is available as a set of primitives to higher-level modules. Regimes include low-commitment search, high-commitment engagement, avoidance, sustained gaze, and deliberate gaze breaking. The primitive percepts generated by this level include a characterization of the most salient regions of the image in terms of the feature maps, an extended characterization of the tracked region in terms of the results of post-attentive processing (eye detection, distance estimation), and signals related to undesired conditions, such as a looming object, or an object moving at speeds the tracker finds difficult to keep up with.

We now move on to discuss the next level of behavioral organization – motor skills.

13.3 Visual Motor Skills

Recall from chapter 10, given the current task (as dictated by the behavior system), the motor skills level is responsible for figuring out how to move the actuators to carry out the stated goal. Often this requires coordination between multiple motor modalities (speech, body posture, facial display, and gaze control).

The motor skills level interacts with both the behavior level above and the primitives level below. Requests for visual skills (each implemented as a FSM) typically originate from the behavior system. During turn-taking, for instance, the behavior system requests different visual primitives depending upon when the robot is trying to relinquish the floor (tending to make eye contact with the human), or reacquire the floor (tending to avert gaze to break eye contact). Another example is the searching behavior. Here, the FSM for search alternates ballistic orienting movements of the head and eyes to scan the scene, with periods of gaze fixation to lock on the desired salient stimulus. The phases of ballistic orientations with fixations are appropriately timed to allow the perceptual flow of information to reach the behavior releasers and stop the search behavior when the desired stimulus is found. If the timing is too rapid, then the searching behavior would never stop.

We now move up another layer of abstraction, to the behavior level in the hierarchy shown in Figure 10-1.

13.4 Visual Behavior

The behavior level is responsible for establishing the current task for the robot through arbitrating among Kismet's goal-achieving behaviors. By doing so, the observed behavior should be relevant, appropriately persistent, and opportunistic. The details of how this is accomplished are presented in chapter 9. Both the current environmental conditions (as characterized by high-level perceptual releasers, as well as motivational factors (emotion processes and homeostatic regulation) contribute to this decision process.

Interaction of the behavior level with the social level occurs through the world, as determined by the nature of the interaction between Kismet and the human. As the human responds to Kismet, the robot's perceptual conditions change. This can activate a different behavior, whose goal is physically carried out by the underlying motor systems. The human observes the robot's ensuing response and shapes their reply accordingly.

Interaction of the behavior level with the motor skills level also occurs through the world. For instance, if Kismet is looking for a bright toy, then the `seek-toy` behavior is active. This task is passed to the underlying motor skill that carries out the search. The act of scanning the environment brings new perceptions to Kismet's field of view. If a toy is found, then the `seek-toy` behavior is successful and released. At this point, the perceptual conditions for engaging the toy are relevant and the `engage-toy` behaviors become active. Consequently, another set of motor skills become active in order to track and smoothly pursue the toy. This indicates a significantly higher level of interest and engagement.

13.5 Social Level

The social level explicitly deals with issues pertaining to having a human in the interaction loop. As discussed previously, Kismet's eye movements have high communicative value. Its gaze direction indicates the locus of attention. Knowing the robot's locus of attention reveals what the robot currently considers to be behaviorally

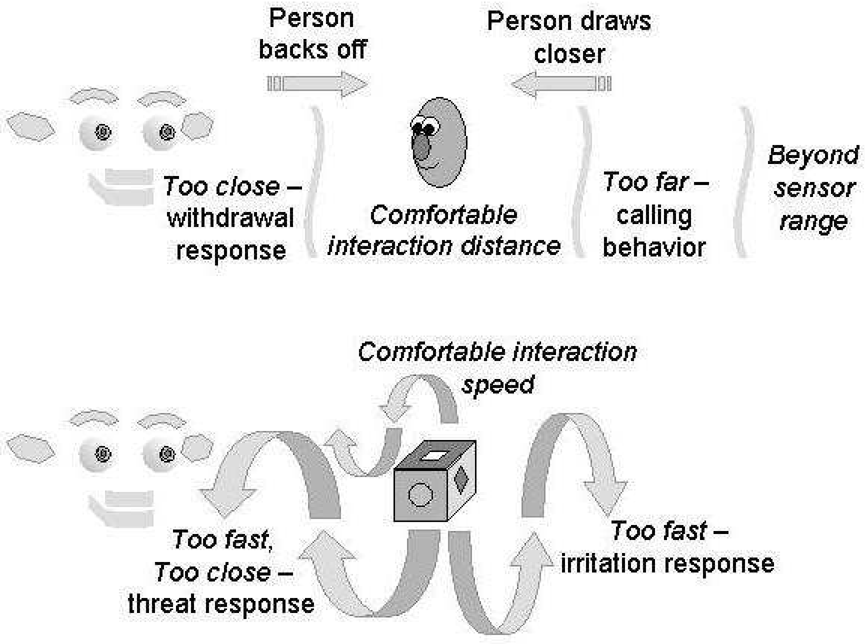


Figure 13-6: Regulating interaction via social amplification. People too distant to be seen clearly are called closer; if they come too close, the robot displays discomfort and withdraws. The withdrawal moves the robot back only a little physically, but is more effective in signaling to the human to back off. Toys or people that move too rapidly cause irritation.

relevant. The robot's degree of engagement can also be conveyed, to communicate how strongly the robot's behavior is organized around what it is currently looking at. If the robot's eyes flick about from place to place without resting, that indicates a low level of engagement, appropriate to a visual search behavior. Prolonged fixation with smooth pursuit and orientation of the head towards the target conveys a much greater level of engagement, suggesting that the robot's behavior is very strongly organized about the locus of attention. Hence, the dynamic aspects of eye movement, such as staring versus glancing, also convey information. Eye movements are particularly potent during social interactions, such as conversational turn-taking, where making and breaking eye contact plays a role in regulating the exchange. As argued previously, we have modeled Kismet's eye movements after humans, so that Kismet's gaze may have similar communicative value.

Eye movements are the most obvious and direct motor actions that support visual perception. But they are by no means the only ones. Postural shifts and fixed action patterns involving the entire robot also have an important role. Kismet has a number of coordinated motor actions designed to deal with various limitations of Kismet's visual perception (see figure 13-6). For example, if a person is visible, but is too distant for their face to be imaged at adequate resolution, Kismet engages in a calling behavior to summon the person closer. People who come too close to the robot also cause difficulties for the cameras with narrow fields of view, since only a small part of a face may be visible. In this circumstance, a withdrawal response is invoked, where Kismet draws back physically from the person. This behavior, by itself, aids the cameras somewhat by increasing the distance between Kismet and the human. But the behavior can have a secondary and greater effect through *social amplification* - for a human close to Kismet, a withdrawal response is a strong social cue to back away, since it is analogous to the human response to invasions of "personal space." Hence, the consequence of Kismet's physical movement aids vision to some extent, but the social interpretation of this movement modulates the person's behavior in a strongly beneficial way for the robot.

Similar kinds of behavior can be used to support the visual perception of objects. If an object is too close, Kismet can lean away from it; if it is too far away, Kismet can crane its neck towards it. Again, in a social context, such actions have power beyond their immediate physical consequences. A human, reading intent into the robot's actions, may amplify those actions. For example, neck-craning towards a toy may be interpreted as interest in that toy, resulting in the human bringing the toy closer to the robot.

Another limitation of the visual system is how quickly it can track moving objects. If objects or people move at excessive speeds, Kismet has difficulty tracking them continuously. To bias people away from excessively boisterous behavior in their own movements or in the movement of objects they manipulate, Kismet shows irritation when its tracker is at the limits of its ability. These limits are either physical (the maximum rate at which the eyes and neck move), or computational (the maximum displacement per frame from the cameras over which a target is searched for).

Such regulatory mechanisms play roles in more complex social interactions, such as conversational turn-taking. Here control of gaze direction is important for regulating conversation rate (Cassell 1999*a*). In general, people are likely to glance aside when they begin their turn, and make eye contact when they are prepared to relinquish their turn and await a response. Blinks occur most frequently at the end of an utterance. These and other cues allow Kismet to influence the flow of conversation to the advantage of its auditory processing. However, Kismet does not perceive these gaze cues when used by others. Here we see the visual-motor system being driven by the requirements of a nominally unrelated sensory modality, just as behaviors that seem completely orthogonal to vision (such as ear-wiggling during the call behavior to attract a person's attention) are nevertheless recruited for the purposes of regulation.

These mechanisms also help protect the robot. Objects that suddenly appear close to the robot trigger a looming reflex, causing the robot to quickly withdraw and appear startled. If the event is repeated, the response quickly habituates and the robot simply appears annoyed, since its best strategy for ending these repetitions is to clearly signal that they are undesirable. Similarly, rapidly moving objects close to the robot are threatening and trigger an escape response.

These mechanisms are all designed to elicit natural and intuitive responses from humans, without any special training. But even without these carefully crafted mechanisms, it is often clear to a human when Kismet's perception is failing, and what corrective action would help. This is because the robot's perception is reflected in behavior in a familiar way. Inferences made based on our human preconceptions are actually likely to work.

13.6 Evidence of Social Amplification

To evaluate the social implications of Kismet's behavior, we invited a few people to interact with the robot in a free-form exchange. There were four subjects in the study, two males (one adult and one child) and two females (both adults). They ranged in age from twelve to twenty-eight. None of the subjects were affiliated with MIT. All had substantial experience with computers. None of the subjects had any prior experience with Kismet. The child had prior experience with a variety of interactive toys. Each subject interacted with the robot for twenty to thirty minutes. All exchanges were video recorded for further analysis.

For the purposes of this chapter, we analyzed the video for evidence of social amplification. Namely, did people read Kismet's cues and did they respond to them in a manner that benefited the robot's perceptual processing or its behavior? We found several classes of interactions where the robot displayed social cues and successfully regulated the exchange.

Establishing a Personal Space

The strongest evidence of social amplification was apparent in cases where people came within very close proximity of Kismet. In numerous instances the subjects would bring their face very close to the robot's face. The robot would withdraw, shrinking backwards, perhaps with an annoyed expression on its face. In some cases the robot would also issue a vocalization with an expression of disgust. In one instance, the subject accidentally came too close and the robot withdrew without exhibiting any signs of annoyance. The subject immediately queried "Am I too close to you? I can back up", and moved back to put a bit more space between himself and the robot. In another instance, a different subject intentionally put his face very close to the robot's face to explore the response. The robot withdrew while displaying full annoyance in both face and voice. The subject immediately pushed backwards, rolling the chair across the floor to put about an additional three feet between himself and the robot, and promptly apologized to the robot.

Overall, across different subjects, the robot successfully established a personal space. As discussed in the previous section, this benefits the robot's visual processing by keeping people at a distance where the visual system can detect eyes more robustly. We added this behavioral response to the robot's repertoire because we had noticed from previous interactions with naive subjects, the robot was not granted any personal space. We attribute this to "baby movements" where people tend to get extremely close to infants, for instance.

Luring People to a Good Interaction Distance

People seem responsive to Kismet's calling behavior. When a person is close enough for the robot to perceive his/her presense, but too far away for face-to-face exchange, the robot issues this social display to bring the person closer (see chapter 11). The most distinguishing features of the display are craning the neck forward in the direction of the person, wiggling the ears with large amplitude, and vocalizing with an excited affect. The function of the display is to lure people into an interaction distance that benefits the vision system. This behavior is not often witnessed as most subjects simply pull up a chair in front of the robot and remain seated at a typical face-to-face interaction distance.

However, the youngest subject took the liberty of exploring different interaction ranges. Over the course of about fifteen minutes he would alternately approach the robot to a normal face-to-face distance, move very close to the robot (invading its personal space), and backing away from the robot. Upon the first appearance of the calling response, the experimenter queried the subject about the robot's behavior. The subject interpreted the display as the robot wanting to play, and he approached the robot. At the end of the subject's investigation, the experimenter queried him about the further interaction distances. The subject responded that when he was further from Kismet, the robot would lean forward. He also noted that the robot had a harder time looking at his face when he was farther back. In general, he interpreted the leaning behavior as the robot's attempt to initiate an exchange with him. We have noticed from earlier interactions (with other people unfamiliar with the robot), that a few people have not immediately understood this display as a "calling" behavior. However, the display is flamboyant enough to arouse their interest and to approach the robot.

Inferring the Level of Engagement

People seem to have a very good sense of when the robot is interested in a particular stimulus or not. By observing the robot's visual behavior, people can infer the robot's level of engagement towards a particular stimulus and generally try to be accommodating. This benefits the robot by bringing it into contact with the desired stimulus. We have already discussed an aspect of this in chapter 6 with respect to directing the robot's attention. Sometimes, however, the robot requires a different stimulus than the one being presented. For instance, the subject may be presenting the robot with a brightly colored toy, but the robot is actively trying to satiate its social drive and searching for something skin-toned. As the subject tries to direct the robot's attention to the toy, the motion is enough to have the robot glance towards it (during the hold-gaze portion of the search behavior). However, not being the desired stimulus, the robot moves its head and eyes to look in another direction. The subject often responds something akin to "You don't want this? Ok, how about this toy", as he/she attempts to get the robot interested in a different toy. Most likely the robot settles its gaze on the person's face fairly quickly. Noticing that the robot is more interested in them than the toy, they will begin to engage the robot vocally.

13.7 Limitations and Extensions

The data from these interactions is encouraging, but more formal studies with a larger number of subjects should be carried out. Whenever introducing a new person to Kismet, there is typically a getting acquainted period of five to ten minutes. During this time, the person gets a sense of the robot's behavioral repertoire and its limitations. As they notice "hiccups" in the interaction, they begin to more closely read the robot's cues and adapt their behavior. We have taken great care in designing these cues so that people intuitively understand the conditions under which they are elicited and what function they serve. We have seen evidence that people readily and willingly read these cues to adapt their behavior in a manner that benefits the robot.

However twenty to thirty minutes is insufficient time to observe all of Kismet's cues, or to observe all the different types of interactions that Kismet has been designed to handle. For each subject, only a subset of these interactions were encountered. Often there is a core set of interactions that most people readily engage in with the robot (such as vocal exchanges and using a toy to play with the robot). The other interactions are more serendipitous (such as exploring the robot's interaction at a distance). People are also constrained by social norms. They rarely do anything that would be threatening or intentionally annoying to the robot. Thus, we have not witnessed how naive subjects interpret the robot's protective responses (such as its fear and escape response).

Extending Oculo-motor primitives

There are a couple of extensions that should be made to the oculo-motor system. The vestibulo-ocular reflex (VOR) is only an approximation of the human counterpart. Largely this is because the robot did not have the equivalent of a vestibular system. However, this issue has been rectified. Kismet now has a three degree of freedom inertial sensor that measures head orientation (as the vestibular system does for people). Our group has already developed VOR code for other robots, so porting the code to Kismet will happen soon. The second extension is to add vergence movements. It is very tricky to implement vergence on a robot like Kismet, because small corrections of each eye give the robot's gaze a chameleon-esque quality that is disturbing for people to look at. Computing a stereo map from the central wide field of view cameras would provide the foveal cameras with a good depth estimate, which could then be used to verge the eyes on the desired target. Since Kismet's eyes are fairly far apart, we will not attempt to exactly center the target with each fovea camera as this gives the robot a cross-eyes appearance even for objects that are near by, but not invading the robot's personal space. Hence, there are many aesthetic issues that must be addressed as we implement these visual capabilities so as to not offend the human who interacts with Kismet.

Improving Social Responsiveness

There are several ways in which Kismet's social responsiveness can be immediately improved. Many of these relate to the robot's limited perceptual abilities. Some of these are issues of robustness, of latency, or both.

Kismet's interaction performance at a distance needs to be improved. When a person is within perceptual range, the robot should make a compelling attempt to bring the person closer. The believability of the robot's behavior is closely tied to how well it can maintain mutual regard with that person. This requires that the robot be more robust in detecting people and their faces at a distance. The difference between having Kismet issue the calling display while looking at a person's face versus looking away from the person is enormous. We find that a person will not interpret the calling display as a request for engagement unless the robot is looking at their face when performing the display. It appears that the robot's gaze direction functions as a sort of social pointer – it says that "I'm directing this request and sending this message to you." For compelling social behavior, it's very important to get gaze direction right.

The perceptual performance can be improved by employing multi-resolution sampling on the camera images. Regions of the wide field of view that indicate the presence of skin-tone could be sampled at a higher resolution to see if that patch corresponds to a person. This requires another stage of processing that is not in the current implementation. If promising, the foveal camera could then be directed to look at that region to see if it can detect a face. Currently the foveal camera only searches for eyes, but at these distances the person's face is too small to reliably detect eyes. A face detector would have to be written for the foveal camera. If the presence of a face has been confirmed, then this target should be passed to the attention system to maintain this region as the target for the duration of the calling behavior. Other improvements to the visual system were discussed in chapter 6. These would also benefit interaction with humans.

13.8 Summary

Motor control for a social robot poses challenges beyond issues of stability and accuracy. Motor actions will be perceived by human observers as semantically rich, regardless of whether the imputed meaning is intended or not. This can be a powerful resource for facilitating natural interactions between robot and human, and places constraints on the robot's physical appearance and movement. It allows the robot to be readable - to make its behavioral intent and motivational state transparent at an intuitive level to those it interacts with. It allows the robot to regulate its interactions to suit its perceptual and motor capabilities, again in an intuitive way with which humans naturally co-operate. And it gives the robot leverage over the world that extends far beyond its physical competence, through social amplification of its perceived intent. If properly designed, the robot's visual behaviors can be matched to human expectations and allow both robot and human to participate in natural and intuitive social interactions.

We have found that different subjects have different personalities and different interaction styles. Some people read Kismet's cues more readily than others. Some people take longer to adapt their behavior to the robot. For our small number of subjects, we have found that people do intuitively and naturally adapt their behavior to the robot. They tune themselves to the robot in a manner that benefits the robot's computational limitations and improves the quality of the exchange. As is evident in the video, they enjoy playing with the robot. They express fondness of Kismet. They tell Kismet about their day, and about personal experiences. They treat Kismet with politeness and consideration (often apologizing if they have irritated the robot). They often ask the robot what it likes, what it wants, or how it feels in an attempt to please it. The interaction takes place on a physical, social, and affective level. In so many ways, they treat Kismet as if it were a socially aware, living creature.

Chapter 14

Summary, Future Directions, and Conclusion

14.1 Summary of Significant Contributions

In the preceding chapters we have given an in-depth presentation of Kismet's physical design and the design of its synthetic nervous system. We have outlined a series of issues that we have found to be important when designing autonomous robots that engage humans in natural, intuitive, and social interaction. Some of these issues pertain to the physical design of the robot: its aesthetics, its sensory configuration, and its degrees of freedom. We have designed Kismet according to these principles.

Other issues pertain to the design of the synthetic nervous system. To address these computational issues, we presented a framework that encompasses the architecture, the mechanisms, the representations, and the levels of control for building a sociable machine. We have emphasized how designing for a human-in-the-loop profoundly impacts how one thinks about the robot control problem, largely because robot's actions have social consequences that extend far beyond the immediate physical act. Hence, one must carefully consider the social constraints imposed on the robot's observable behavior. However, the designer can use this to benefit the quality of interaction between robot and human. We have illustrated this by presenting the numerous ways Kismet pro-actively regulates its interaction with the human so that the interaction is appropriate for both partners. The process of social amplification is a prime example.

In an effort to make the robot's behavior readable, believable, and well matched to the human's social expectations and behavior, we have incorporated several theories, models, and concepts from psychology, social development, ethology, and evolutionary perspectives, into the design of the synthetic nervous system. We highlighted how each system addresses important issues to support natural and intuitive communication with a human. We have paid special attention to designing the infrastructure into the synthetic nervous system to support socially situated learning.

We have integrated these diverse capabilities into a single robot and have situated that robot within a social environment. We have evaluated the performance of the human-robot system with numerous studies with human subjects. Below we summarize our findings as they pertain to the key design issues and evaluation criteria outlined in chapter 3.

14.2 Summary of Key Design Issues

Through our studies with human subjects, we have found that Kismet addresses the key design issues in rich and interesting ways. By going through each design issue, we recap the different ways in which Kismet meets the four evaluation criteria. Recall from chapter 3, these criteria are:

- Can people intuitively read and do they naturally respond to Kismet's social cues?
- Can Kismet perceive and appropriately respond to these naturally offered cues?
- Does the human adapt to the robot, and the robot adapt to the human, in a way that benefits the interaction?
- Does Kismet readily elicit scaffolding interactions from the human that could be used to benefit learning?

Real-time Performance

Kismet successfully maintains interactive rates in all of its systems to dynamically engage a human. We discussed the performance latencies of several systems including visual and auditory perception, visual attention, lip synchronization, and turn-taking behavior during proto-dialog. Although each of these systems does not perform at adult human rates, they operate fast enough to allow a human engage the robot comfortably. The robot provides important expressive feedback to the human that they intuitively use to entrain to the robot's level of performance.

Establish Appropriate Social Expectations

Great care has been taken in designing Kismet's physical appearance, its sensory apparatus, its mechanical specification, and its observable behavior (motor acts and vocal acts) to establish a robot-human relationship that follows the infant-caregiver metaphor. Following the baby-scheme of Eibl-Eiblsfeldt, Kismet's appearance encourages people to treat it as if it were a very young child or infant. Kismet has been given a child-like voice and it babbles in its own characteristic manner. We have observed that our female subjects are willing to use exaggerated prosody when talking to Kismet, characteristic of motherese. Both or male and female subjects tend to sit directly in front of and close to Kismet, facing it the majority of the time. When engaging Kismet in proto-dialog, they tend to slow down, use shorter phrases, and wait longer for Kismet's response. With some subjects, we have observed their use of exaggerated facial expressions. All these behaviors are characteristic of interacting with very young animals (e.g., puppies) or infants.

Self Motivated Interaction

Kismet exhibits self-motivated and pro-active behavior. Kismet is in a never-ending cycle of satiating its drives. As a result, the stimuli it actively seeks out (people-like things vs toy-like things) changes over time. The first level of the behavior system acts to seek out the desired stimulus when it is not present, to engage it when it has been found, and to avoid it if it is behaving in an offensive or threatening manner. The gains of the attention system are dynamically adjusted over time to facilitate this process. Kismet can take the initiative in establishing an interaction. For instance, if Kismet is the process of satiating its `social-drive`, it will call to a person who is present but slightly beyond face-to-face interaction distance.

Regulate Interactions

Kismet is well versed in regulating its interactions with the caregiver. It has several mechanisms for accomplishing this, each for different kinds of interactions. They all serve to slow the human down to an interaction rate that is within the comfortable limits of Kismet's perceptual, mechanical, and behavioral limitations. By doing so, the robot is neither overwhelmed nor under-stimulated by the interaction.

The robot has two regulatory systems that serve to maintain the robot in a state of "well being". These are the emotive responses and the homeostatic regulatory mechanisms. The `drives` establish the desired stimulus and motivate the robot to seek it out and to engage it. The emotions are another set of mechanisms, with greater direct control over behavior and expression, that serve to bring the robot closer to desirable situations (`joy`, `interest`, even `sorrow`), and cause the robot to withdraw from or remove undesirable situations (`fear`, `anger`, or `disgust`). Which emotional response becomes active depends largely on the releasers, but also on the internal state of the robot. The behavioral strategy may involve a social cue to the caregiver (through facial expression and body posture) or a motor skill (such as the escape response). The use of social amplification to define a personal space is a good example of how social cues, that are a product of emotive responses, can be used to regulate the proximity of the human to the robot. It is also used to regulate the movement of toys when playing with the robot.

Kismet's turn taking cues for regulating the rate of proto-dialog is another case. Here, the interaction happens on a more tightly coupled temporal dynamic between human and robot. The mechanism originates from the behavior system instead of the emotion system. It employs communicative facial displays instead of emotive facial expressions. Our studies suggest that subjects read the robot's turn-taking cues to entrain to the robot. As a result, the proto-dialog becomes smoother over time.

Readable Social Cues

Kismet is a very expressive robot. It can communicate emotive state and social cues to a human through face, gaze direction, body posture, and voice. Our results from various forced choice and similarity studies suggest that Kismet's emotive facial expressions and vocal expressions are readable. More importantly, several studies suggest that people readily read and correctly interpret Kismet's expressive cues when actively engaging the robot. We found that several interesting interactions arose between Kismet and female subjects when we combined Kismet's ability to recognize vocal affective intent (for praise, prohibition, etc.) with expressive feedback. The female subjects used Kismet's facial expression and body posture as a social cue to determine when Kismet "understood" their intent. Our video of these interactions suggests evidence of affective feedback where the subject would issue an intent (say, an attentional bid), the robot would respond expressively (perking its ears, leaning forward, and rounding its lips), and then the subject would immediately respond in kind (perhaps by saying "Oh!" or "Ah!"). Several subjects appeared to empathize with the robot after issuing a prohibition - often reporting feeling guilty or bad for scolding the robot and making it "sad". For turn-taking interactions, after a period of entrainment, subjects appear to read the robot's social cues and hold their response until prompted by the robot. This allows for longer runs of clean turns before an interruption or delay occurs in the proto-dialog.

Read Human's Social Cues

We have presented two cases where the robot can read the human's social cues. The first is the ability to recognize praise, prohibition, soothing, and attentional bids from robot directed speech. This could serve as an important teaching cue for reinforcing and shaping the robot's behavior. The second is the ability of humans to direct Kismet's attention using natural cues. This could play an important role in socially situated learning by giving the caregiver a way of showing Kismet what is important for the task, and for establishing a shared reference.

Competent Behavior in a Complex World

Kismet's behavior exhibits robustness, appropriateness, coherency, and flexibility when engaging a human in either physical play with a toy, in vocal exchanges, or affective interactions. It also exhibits appropriate persistence and reasonable opportunism when addressing its time varying goals. These qualities arise from the interaction between the external environment with the internal dynamics of Kismet's synthetic nervous system. The behavior system is designed to address these issues on the task level, but the observable behavior is a product of the behavior system working in concert with the perceptual, attention, motivation, and motor systems. In chapter 10 we conceptualized Kismet's behavior to be the product of interactions within and between four separate levels.

Believable Behavior

Kismet exhibits compelling and life-like behavior. To promote this quality of behavior, we addressed the issues of audience perception, and of biasing the robot's design towards believability, simplicity, and caricature over forced realism. We implemented a set of proto-social responses that are synthetic analogs of those believed to play an important role in launching infants into social exchanges with their caregivers.

From our video recordings of subjects interacting with Kismet, people do appear to treat Kismet as a very young, socially aware creature. They seem to treat the robot's expressive behaviors and vocalizations as meaningful responses to their own attempts at communication. The robot's prosody has enough variability that they answer Kismet's "questions", comment on Kismet's "statements" and react to Kismet's "exclamations". They ask Kismet about its thoughts and feelings, how its day is going, and they share their own personal experiences with the robot. These kinds of interactions are important to foster the social development of human infants. They could also play an important role in Kismet's social development as well.

14.3 Infrastructure for Socially Situated Learning

In the above discussion, we have taken care to relate these issues to socially situated learning. In previous work, we have posed these issues with respect to building humanoid robots that can imitate people (Breazeal & Scassellati 2000). We quickly recap these issues below:

- *Knowing What's Important:* This is largely addressed by the design of the attention system. We have demonstrated that it is easy for people to direct the robot's attention, as well as to confirm when the robot's attention has been successfully manipulated. People can also use their voice to arouse the robot through attentional bids. More work needs to be done, but this is a start.

- *Recognizing Progress:* The robot is designed to have both internal mechanisms as well as external mechanisms for recognizing progress. The change in Kismet's internal state (the satiation of its drives, or the return to a slightly positive affective state) could be used as internal reinforcement signals for the robot. Other systems have used signals of this type for operant as well as classical conditioning of robotic or animated characters (Velasquez 1998), (Blumberg, Todd & Maes 1996), (Yoon et al. 2000). Kismet also has the ability to extract progress measures from the environment, through socially communicated praise, prohibition, and soothing. The underlying mechanism would actually be similar to the previous case, as the human is modulating the robot's affective state by communicating these intents. Eventually, this could be extended to having the robot recognize positive and negative facial expressions.

- *Recognizing Success:* The same mechanisms for recognizing progress could be used to recognize success. The ability for the caregiver to socially manipulate the robot's affective state has interesting implications for teaching the robot novel acts. The robot may not require an explicit representation of the desired goal nor a fully specified evaluation function before embarking upon learning the task. Instead, the caregiver could initially serve as the evaluation function for the robot, issuing praise, prohibition, and encouragement as he/she tries to shape the robot's behavior. It would be interesting if the robot could learn how to associate different affective states to the learning episode. Eventually, the robot may learn to associate the desired goal with positive affect – making that goal an explicitly represented goal within the robot instead of an implicitly represented goal through the social communication of affect. This kind of scenario could play an important part in socially transferring new goals from human to robot. Many details need to be worked out, but the kernel of the idea is intriguing.

- *Structured Learning Scenarios:* Kismet has two strategies for establishing an appropriate learning environment. Both involve regulating the interaction with the human. The first takes place through the motivation system. The robot uses expressive feedback to indicate to the caregiver when it is either overwhelmed or under-stimulated. In time, this mechanism could be embellished so that homeostatic balance of the drives would correspond to a learning environment where the robot is slightly challenged but largely competent. The second form of regulation is turn-taking, which is implemented in the behavior system. Turn-taking is a cornerstone of human-style communication and tutelage. It forms the basis of interactive games and structured learning episodes. Someday, these interaction dynamics could play an important role in socially situated learning for Kismet.

- *Quality Instruction:* Kismet provides the human with a wide assortment of expressive feedback through several different expressive channels. Currently, this is used to help entrain the human to the robot's level of competence, and to help the human maintain Kismet's "well being" by providing the appropriate kinds of interactions at the appropriate times. In time, this could also be used to intuitively help the human provide better quality instruction. Looks of puzzlement, nods or shakes of the head, and other gestures and expressions could be employed to elicit further assistance or clarification from the caregiver.

14.4 Grand Challenges of Building Sociable Machines

In this thesis, we have only begun to explore the question of building a sociable machine. Human beings are a social species of extraordinary breadth and depth. Social interaction has played a critical role in our evolution, our development, our education, and our existence in society. The social dimension of our existence touches upon the most human of qualities: personality, identity, emotions, empathy, loyalty, friendship, and more. If we are to ever achieve a top-down, bottom-up understanding of human intelligence (the mission statement of the MIT AI Lab), then we cannot ignore the social dimension. There are a few researchers already grappling with these difficult questions on autonomous robots such as (Scassellati 2000), (Dautenhahn 1997), (Nehaniv 1999) (this is not a complete list). Through the process of building sociable machines, we hope to come to a deeper understanding and appreciation of our own humanity. Below we list what we view to be a few of the grand challenges in building a socially intelligent machine, a sociable machine:

- Self Identity
- Theory of Mind
- Autobiographical Memory
- Recognition of Self, Other, and Conspecifics
- Social Learning (esp. Imitation)
- Intentionality
- Emotion
- Empathy
- Personality
- Friendship

14.5 Conclusion

We hope that Kismet is a precurser to the socially intelligent machines of the future. Today, Kismet is the only autonomous robot that can engage humans in natural and intuitive interaction that is physical, affective, and social. At times, people interact with Kismet at a level that seems personal – sharing their thoughts, feelings, and experiences with Kismet. They ask Kismet to share the same sorts of things with them.

After a three-year investment, we are in a unique position to study how people interact with sociable autonomous robots. We have some promising results, but many more studies need to be performed to come to a deep understanding of how people interact with these technologies. Also, we are now in the position to study socially situated learning following the infant-caregiver metaphor. From its inception, this form of learning has been the motivation for building Kismet, and for building Kismet in the way we have.

In the near term, we are interested in emulating the process by which infants "learn to mean" (Halliday 1975). Specifically, we are interested in investigating the role social interaction plays in very young children (even african grey parrots, as evidenced by the work of Pepperberg (1990)) learn the meaning their vocalizations have for others, and how to use this knowledge to benefit behavior and communication. There are so many different questions we want to explore in this fascinating area of research. We hope we have succeeded in inspiring others to follow.

In the meantime, kids are growing up with robotic and digital pets such as Aibo, Furby, Tomogotchi, Petz, and others soon to enter the toy market. Their experience with interactive technologies is very different from that of their parents or grandparents. As the technology improves and these children grow up, it will be interesting to see what is natural, intuitive, and even expected of these interactive technologies. Sociable machines and other sociable technologies may become a reality sooner than we think.

Bibliography

Ambrose, R., Aldridge, H. & Askew, S. (1999), NASA's Robonaut system, *in* 'Proceedings of HURO99', Tokoyo, Japan, pp. 131–132.

Ball, G. & Breese, J. (2000), Emotion and personality in a conversational agent, *in* S. P. J Cassell, J. Sullivan & E. Churchill, eds, 'Embodied conversation agents', MIT Press, Cambridge, MA, pp. 189–219.

Ball, G., Ling, D., Kurlander, D., Miller, J., Pugh, D., Skelley, T., Stankosky, A., Thiel, D., Dantzich, M. V. & Wax, T. (1997), Lifelike computer characters: The Persona Project at microsoft research, *in* J. Bradshaw, ed., 'Software Agents', MIT Press.

Ballard, D. (1989), 'Behavioral constraints on animate vision', *Image and Vision Computing* **7**(1), 3–9.

Bates, J. (1994), 'The role of emotion in believable characters', *Communications of the ACM* **37**(7), 122–125.

Bates, J., Loyall, B. & Reilly, S. (1992), An architecture for action, emotion, and social behavior, Technical Report CMU-CS-92-144, CMU, Pittsburgh, PA.

Bateson, M. (1979), The epigenesis of conversational interaction: a personal account of research development, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 63–77.

Bernardino, A. & Santos-Victor, J. (1999), 'Binocular Visual Tracking: Integration of Perception and Control', *IEEE Transactions on Robotics and Automation* **15(6)**, 1937–1958.

Billard, A. & Dautenhahn, K. (1997), Grounding Communication in Situated, Social Robots, Technical Report UMCS-97-9-1, University of Manchester.

Blair, P. (1949), *Animation: Learning how to draw animated cartoons*, Walter T. Foster Art Books, Laguna Beech, CA.

Blumberg, B. (1994), Action Selection in Hamsterdam: Lessons from Ethology, *in* 'Proceedings of SAB94', MIT Press, Cambridge, MA.

Blumberg, B. (1996), Old Tricks, New Dogs: Ethology and Interactive Creatures, PhD thesis, MIT.

Blumberg, B., Todd, P. & Maes, M. (1996), No Bad Dogs: Ethological Lessons for Learning, *in* 'Proceedings of SAB96', MIT Press, Cambridge, MA.

Brazelton, T. (1979), Evidence of Communication in Neonatal Behavior Assessment, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 79–88.

Breazeal, C. & Aryananda, L. (2000), Recognition of affective communicative intent in robot-directed speech, *in* 'Proceedings of Humanoids2000 (submitted)'.

Breazeal, C. & Foerst, A. (1999), Schmoozing with robots: exploring the original wireleses network, *in* 'Proceedings of Cognitive Technology (CT99)', San Francisco, CA, pp. 375–390.

Breazeal, C. & Scassellati, B. (1999*a*), A context-dependent attention system for a social robot, *in* 'Proceedings of IJCAI99', Stockholm, Sweden, pp. 1146–1151.

Breazeal, C. & Scassellati, B. (1999*b*), How to build robots that make friends and influence people, *in* 'Proceedings if IROS99', Kyonju, Korea, pp. 858–863.

Breazeal, C. & Scassellati, B. (2000), *Imitation in Animals and Artifacts*, MIT Press (to appear), chapter Challenges in building robots that imitate people.

Breazeal, C., Fitzpatrick, P., Edsinger, A. & Scassellati, B. (2000), Social constraints on animate vision, *in* 'Proceedings of Humanoids2000 (submitted)'.

Brooks, R. (1986), 'A robust layered control system for a mobile robot', *IEEE Journal of Robotics and Automation* **RA-2**, 253–262.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B. & Williamson, M. M. (1999), The Cog Project: Building a Humanoid Robot, *in* C. L. Nehaniv, ed., 'Computation for Metaphors, Analogy and Agents', Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag.

Bullowa, M. (1979), *Before Speech: The Beginning of Interpersonal Communicaion*, Cambridge University Press, Cambridge, London.

Cahn, J. (1990), Generating Expression in Synthesized Speech, Master's thesis, MIT Media Lab, Cambridge, MA.

Carey, S. & Gelman, R. (1991), *The Epigenesis of Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Carver, C. & Scheier, M. (1998), *On the self-regulation of behavior*, Cambridge University Press, Cambridge, UK.

Cassell, J. (1999*a*), Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialog Systems, *in* Luperfoy, ed., 'Spoken Dialog Systems', MIT Press, Cambridge, MA.

Cassell, J. (1999*b*), Nudge Nudge Wink Wink: Elements of face-to-face conversation for embodied conversational agents, *in* J. Cassell, ed., 'Embodied Conversational Agents', MIT Press, Cambridge, MA.

Cassell, J. & Thorisson, K. (1999), 'The power of a nod and a glance: envelope verses emotional feedback in animated conversational agents', *Applied Artificial Intelligence* **13**, 519–538.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H. & Yan, H. (2000), Human conversation as a system framework: designing embodied conversation agents, *in* S. P. J Cassell, J. Sullivan & E. Churchill, eds, 'Embodied conversation agents', MIT Press, Cambridge, MA, pp. 29–63.

Chen, L. & Huang, T. (1998), 'Multimodal Human Emotion/Expression Recognition', *Proceedings of the second international conference on automatic face and gesture recognition* pp. 366–371.

Cole, J. (1998), *About Face*, MIT Press, Cambridge, MA.

Collis, G. (1979), Describing the Structure of Social Interactin in Infancy, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 111–130.

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

Dario, P. & Susani, G. (1996), Physical and psychological interactions between humans and robots in the home environment, *in* 'Proceedings of the first international symposium on humanoid robots HURO96', Tokoyo, Japan, pp. 5–16.

Darwin, C. (1872), *The expression of the emotions in man and animals*, John Murray, London.

Dautenhahn, K. (1997), 'I could be you – the phenomenological dimension of social understanding', *Cybernetics and Systems Journal* **28**(5), 417–453.

Dautenhahn, K. (1999), Embodiment and interaction in socially intelligent life-like agents, *in* C. L. Nehaniv, ed., 'Computation for Metaphors, Analogy and Agents', Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag, pp. 102–142.

de Boysson-Bardies, B. (1999), *How Language comes to children, from birth to two years*, MIT Press, Cambridge MA.

Dellaert, F., Polzin, F. & A., W. (1996), Recognizing emotion in speech, *in* 'Proceedings of ICSLP96'.

Dennett, D. (1987), *The Intentional Stance*, MIT Press, Cambridge, MA.

Duchenne, B. (1862), *The mechanism of human facial expression or an electro-physiological analysis of the expression of emotions*, Cambridge University Press, New York.

Eckerman, C. & Stein, M. (1987), 'How imitation begets imitation and toddlers' generation of games', *Developmental psychology* **26**, 370–378.

Eibl-Eibesfeldt, I. (1972), Similarities and differences between cultures in expressive movements, *in* R. Hinde, ed., 'Nonverbal communication', Cambridge University Press, Cambridge, pp. 297–311.

Ekerman, C. (1993), Toddlers' achievement of coordinated action with conspecifics: a dynamic systems perspective, *in* L. Smith & E. Thelen, eds, 'A Dynamic systems approach to development: Applications', MIT Press, Cambridge, MA, pp. 333–358.

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

Ekman, P. & Friesen, W. (1982), Measuring facial movement with the Facial Action Coding System, *in* 'Emotion in the human face', Cambridge University Press, Cambridge, pp. 178–211.

Ekman, P. & Oster, H. (1982), Review of research, 1970 to 1980, *in* P. Ekman, ed., 'Emotion in the human face', Cambridge University Press, Cambridge, pp. 147–174.

Ekman, P., Friesen, W. & Ellsworth, P. (1982), What emotion categories or dimensions can observers judge from facial behavior?, *in* P. Ekman, ed., 'Emotion in the human face', Cambridge University Press, Cambridge, pp. 39–55.

Elliot, C. D. (1992), The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System, PhD thesis, Institute for teh Learning Sciences, Northwestern University.

Fernald, A. (1984), The perceptual and affective salience of mothers' speech to infants, *in* C. G. L. Feagans & R. Golinkoff, eds, 'The origins and growth of communication', Ablex publishing, Norwood, NJ, pp. 5–29.

Fernald, A. (1989), 'Intonation and communicative intent in mother's speech to infants: Is the melody the message?', *Child Development* **60**, 1497–1510.

Fernald, A. (1993), 'Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages', *Developmental Psychology* **64**, 657–674.

Ferrier, L. (1985), Intonation in discourse: Talk between 12-month-olds adn their mothers, *in* K. Nelson, ed., 'Children's Language', Erlbaum, Hillsdale, NJ, pp. 35–60.

Fleming, B. & Dobbs, D. (1999), *Animating facial features and expressions*, Charles river media.

Frijda, N. (1969), Recognition of emotion, *in* L. Berkowitz, ed., 'Advances in experimental social psychology', Academic Press, New York, pp. 167–223.

Frijda, N. (1994*a*), Emotions are functional, most of the time, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 112–122.

Frijda, N. (1994*b*), Emotions require cognitions, even if simple ones, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 197–202.

Frijda, N. (1994*c*), Universal Antecedents Exist, and are interesting, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 146–149.

Fujita, M. & Kageyama, K. (1997), An open architecture for robot entertainment, *in* 'Proceedings of Agents97'.

Galef, B. (1988), Imitation in animals: history, definitinos, and interpretation of data from teh psychological laboratory, *in* T. Zentall & G. Galef, eds, 'Social learning: psychological and biological perspectives', Lawrence Erlbaum.

Gallistel, C. (1980), *The organization of action*, MIT Press, Cambridge, MA.

Gallistel, C. (1990), *The organization of learning*, MIT Press, Cambridge, MA.

Garvey, C. (1974), 'Some properties of social play', *Merrill-Palmer Quarterly* **20**, 163–180.

Gould, J. (1982), *Ethology*, Norton.

Grieser, D. & Kuhl, P. (1988), 'Maternal speech to infants in a tonal language: support for universal prosodic features in motherese', *Developmental Psychology* **24**, 14–20.

Halliday, M. (1975), *Learning How to Mean: Explorations in the Development of Language*, Elsevier, New York, NY.

Halliday, M. (1979), One Child's Protolanguage, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 149–170.

Hara, F. (1998), Personality characterization of animate face robot through interactive communication with human, *in* 'Proceedings of IARP98', Tsukuba, Japan, pp. IV–1.

Hauser, M. (1996), *The Evolution of Communication*, MIT Press, Cambridge, MA.

Hayes, G. & Demiris, J. (1994), A Robot Controller Using Learning by Imitation, *in* 'Second international symposium on intelligent robotic systems', Grenoble, France, pp. 198–204.

Heckhausen, J. (1987), 'Balancing for weaknesses and challenging developmental potential: a longitudinal study of mother-infant dyads in appretenticeship interactions', *Developmental psychology* **23**(6), 762–770.

Hendriks-Jansen, H. (1996), *Catching Ourselves in the Act*, MIT Press, Cambridge, MA.

Hirai, K. (1998), Humanoid robot and its applications, *in* 'Proceedings of IARP99', pp. V–1.

Hirsh-Pasek, K., Jusczyk, P., Cassidy, K. W., Druss, B. & Kennedy, C. (1987), 'Clauses are perceptual units of young infants', *Cognition* **26**, 269–286.

Horn, B. (1986), *Robot Vision*, MIT Press, Cambridge, MA.

Irie, R. (1995), Robust Sound Localization: An Application of an Auditory Perception System for a Humanoid Robot, Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Itti, L., Koch, C. & Niebur, E. (1998), 'A Model of Saliency-Based Visual Attention for Rapid Scene Analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **20**(11), 1254–1259.

Izard, C. (1977), *Human Emotions*, Plenum, New York.

Izard, C. (1993), Four Systems for Emotion Activation: Cognitive and Noncognitive Processes, *in* 'Psychological Review', Vol. 100, pp. 68–90.

Izard, C. (1994), Cognition is one of four types of emotion activating systems, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 203–208.

Johnson, M. (1993), *Brain Development and Cognition: A Reader*, Blackwell, Oxford, chapter Constraints on Cortical Plasticity, pp. 703–721.

Johnson, M., Wilson, A., Blumberg, B., Kline, C. & Bobick, A. (1999), Sympathetic interfaces: using a plush toy to direct synthetic characters, *in* 'Proceedings of CHI99'.

Kandel, E., Schwartz, J. & Jessell, T. (2000), *Principles of Neuroscience*, fourth edition edn, McGraw Hill.

Kawamura, K., Wilkes, D., Pack, T., Bishay, M. & Barile, J. (1996), Humanoids: future robots for home and factory, *in* 'Proceedings of the first international symposium on humanoid robots HURO96', Tokoyo, Japan, pp. 53–62.

Kaye, K. (1979), Thickening Thin Data: The Maternal Role in Developing Communication and Language, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 191–206.

Kolb, B., Wilson, B. & Laughlin, T. (1992), 'Developmental changes in the recognition and comprehensino offacial expression: Implications ffor frontal lobe function', *Brain and Cognition* pp. 74–84.

Kuniyoshi, Y., Kita, N., Sugimoto, K., Nakamura, S. & Suehiro, T. (1995), A Foveated wide angle lens for active vision, *in* 'Proceedings of IEEE Intl. conference on robotics and automatino'.

Lazarus, R. (1991), *Emotion and adaptation*, Oxford University Press.

Lazarus, R. (1994), Universal antecedents of the emotions, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 163–171.

Lester, J., Towns, S., Callaway, S., Voerman, J. & FitzGerald, P. (2000), Deictic and emotive communication in animated pedagogical agents, *in* S. P. J Cassell, J. Sullivan & E. Churchill, eds, 'Embodied conversation agents', MIT Press, Cambridge, MA, pp. 123–154.

Levenson, R. (1994), Human emotions: a functional view, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 123–126.

Lorenz, K. (1973), *Foundations of Ethology*, Springer-Verlag, New York, NY.

Madsen, R. (1969), *Animated film: concepts, methods, uses*, Interlnad, New York.

Maes, P. (1990), 'Learning Behavior Networks from Experience', *ECAL90*.

Maes, P., Darrell, T., Blumberg, B. & Pentland, A. (1996), The ALIVE System: wireless, full-body interaction with autonomous agents, *in* 'The ACM special issue on multimedia and multisensory virtual worlds'.

Mataric, M., Williamson, M., Demiris, J. & Mohan, A. (1998), Behavior-based primitives for articulated control, *in* 'Proceedings of SAB98', MIT Press, pp. 165–170.

McFarland, D. & Bosser, T. (1993), *Intelligent Behavior in Animals and Robots*, MIT Press, Cambridge, MA.

McRoberts, G., Fernald, A. & Moses, L. (2000), 'An acoustic study of prosodic form-function relationships infant-directed speech', *Developmental Psychology (in press)*.

Meltzoff, A. & Moore, M. (1977), 'Imitation of facial and manual gestures by human neonates', *Science* **198**, 75–78.

Minsky, M. (1988), *The Society of Mind*, Simon and Schuster, New York.

Mumme, D., Fernald, A. & Herrera, C. (1996), 'Infants' response to facial and vocal emotional signals in a social referencing paradigm', *Child Development* **67**, 3219–3237.

Murray, I. & 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. & Tosa, N. (1999), Emotion Recognition and its application to computer agents with spontaneous interactive capabilities, *in* 'Proceedgins of ICMCS99', Vol. 2, pp. 804–808.

Nehaniv, C. (1999), Story-telling and emotion: cognitive technology considerations in networking temporally and affectively grounded minds, *in* 'Proceedings of Cognitive Technology (CT99)', San Francisco, CA, pp. 313–322.

Newman, R. & Zelinsky, A. (1998), Error analysis of head pose and gaze direction from stereo vision, *in* 'Proceedings if IROS98', pp. 527–532.

Newson, J. (1979), The growth of shared understandings between infant and caregiver, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 207–222.

Niedenthal, P. & Kityama, S. (1994), *The Heart's Eye: Emotional influences in Perception and Attention*, Academic Press, San Diego.

Nothdurft, H. C. (1993), 'The role of features in preattentive vision: Comparison of orientation, motion and color cues', *Vision Research* **33**, 1937–1958.

Ortony, A., Clore, G. & Collins, A. (1988), *The Cognitive Structure of Emotion*, Cambridge University Press, Cambridge, UK.

Papousek, M., Papousek, H. & Bornstein, M. (1985), The naturalistic vocal environment of young infants: on the significance of homogeneity and variability in parental speech, *in* T. Field & N. Fox, eds, 'Social perception in infants', Ablex, Norwood, NJ, pp. 269–297.

Parke, F. (1972), Computer generated animation of faces, Master's thesis, University of Utah, Salt Lake City. UTEC-CSc-72-120.

Parke, F. & Waters, K. (1996), *Computer Facial Animation*, A K Peters, Wellesley, MA.

Pepperberg, I. (1988), 'An interactive modeling technique for acquisition of communiation skills: separation of "labeling" and "requesting" in a psittachine subject', *Applied psycholinguistics* **9**, 59–76.

Pepperberg, I. (1990), 'Referential mapping: a technique for attaching functional significance to the innovative utterances of an african grey parrot', *Applied psycholinguistics* **11**, 23–44.

Picard, R. (1997), *Affective Computation*, MIT Press, Cambridge, MA.

Plutchik, R. (1984), Emotions: A general psychevolutionary theory, *in* K. Scherer & P. Elkman, eds, 'Approaches to Emotion', Lawrence Erlbaum Associates, New Jersey, pp. 197–219.

Plutchik, R. (1991), *The emotions*, University press of america, Lanham, MD.

Redican, W. (1982), An evolutionary perspective on human facial displays, *in* 'Emotion in the human face', Cambridge University Press, Cambridge, pp. 212–280.

Reeves, B. & Nass, C. (1996), *The Media Equation*, CSLI Publications, Stanford, CA.

Reilly, S. (1996), Believable Social and Emotional Agents, PhD thesis, CMU School of Computer Science, Pittsburgh, PA.

Rickel, J. & Johnson, W. L. (2000), Task-oriented collaboration with embodied agents in virtual worlds, *in* S. P. J Cassell, J. Sullivan & E. Churchill, eds, 'Embodied conversation agents', MIT Press, Cambridge, MA, pp. 95–122.

Ross, H. & Lollis, S. (1987), 'Communication within infant social games', *Developmental psychology* **23**(2), 241–248.

Roy, D. & Pentland, A. (1996), 'Automatic Spoken Affect Analysis and Classification', *Proceedings of the 1996 international conference on automatic face and gesture recognition*.

Russell, J. (1997), Reading emotions from and into faces: resurrecting a dimensional-contextual perspective, *in* J. Russell & J. Fernandez-Dols, eds, 'The psychology of facial expression', Cambridge university press, Cambridge, pp. 295–320.

Rutter, D. & Durkin, K. (1987), 'Turn-taking in mother-infant interaction: an examination of volications and gaze', *Developmental psychology* **23**(1), 54–61.

Sanders, G. & Scholtz, J. (2000), Measurement and evaluation of embodied conversational agents, *in* S. P. J Cassell, J. Sullivan & E. Churchill, eds, 'Embodied conversation agents', MIT Press, Cambridge, MA, pp. 346–373.

Scassellati, B. (1998), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of AAAI98'.

Scassellati, B. (1999), Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot, *in* C. L. Nehaniv, ed., 'Computation for Metaphors, Analogy and Agents', Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag.

Scassellati, B. (2000), A Theory of mind for a humanoid robot, *in* 'Proceedings of Humanoids2000 (submitted)'.

Schaal, S. (1999), 'Is imitation learning the route to humanoid robots', *Trends in cognitive science* **3**(6), 233–242.

Schaffer, H. (1977), Early Interactive Development, *in* 'Studies of Mother-Infant Interaction: Proceedings of Loch Lomonds Symposium', Academic Press, New York, pp. 3–18.

Schall, S. (1997), Learning from Demonstration, *in* 'Proceedings of NIPS97', pp. 1040–1046.

Scherer, K. (1984), On the nature and function of emotion: a component process approach, *in* Scherer & Ekman, eds, 'Approaches to emotion', Erlbaum, Hillsdale, NJ, pp. 293–317.

Scherer, K. (1994), Evidence for both universality and cultural specificity of emotion elicitation, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion', Oxford University Press, New York, pp. 172–175.

Siegel, D. (1999), *The developing mind: toward a neurobiology of interpersonal experience*, The Guilford Press.

Sinha, P. (1994), 'Object Recognition via Image Invariants: A Case Study', *Investigative Ophthalmology and Visual Science* **35**, 1735–1740.

Slaney, M. & McRoberts, G. (1998), 'Baby Ears: A Recognition System for Affective Vocalizations', *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing.*

Smith, C. (1989), 'Dimensions of appraisal adn physiological response in emotion', *Journal of personality and social psychology* **56**, 339–353.

Smith, C. & Scott, H. (1997), A componential approach to the meaning of facial expressions, *in* J. Russell & J. Fernandez-Dols, eds, 'The psychology of facial expression', Cambridge university press, Cambridge, pp. 229–254.

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

Stern, D. (1975), 'Infant regulation of maternal play behavior and/or maternal regulation of infant play behavior', *Proceedings of the Society of Research in Child Development.*

Stern, D., Spieker, S. & MacKain, K. (1982), 'Intonation contours as signals in maternal speech to prelinguistic infants', *Developmental Psychology* **18**, 727–735.

Takanobu, H., Takanishi, A., Hirano, S., Kato, I., Sato, K. & Umetsu, T. (1998), Development of humanoid robot heads for natural human-robot communication, *in* 'Proceedings of HURO98'.

Takeuchi, A. & Nagao, K. (1993), Communicative facial displays as a new conversational modality, *in* 'Proceedings of ACM/IFIP inter-CHI93', ACM Press, pp. 187–193.

Thomas, F. & Johnston, O. (1981), *Disney animation: The Illusion of Life*, Abbeville Press, New York, NY.

Thorisson, K. (1998), Real-time Decision Making in Multimodal Face-to-face Communication, *in* 'Second International Conference on Autonomous Agents', ACM SIGART, ACM Press, Minneapolis, MN, pp. 16–23.

Tinbergen, N. (1951), *The Study of Instinct*, Oxford University Press, New York.

Trehub, S. & Trainor, L. (1990), *Rules for listening in infancy*, Elsevier, North Holland, chapter chapter 5.

Trevarthen, C. (1979), Communication and cooperation in early infancy: a description of primary intersubjectivity, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 321–348.

Triesman, A. (1986), 'Features and objects in visual processing', *Scientific American* **225**, 114B–125.

Tronick, E., Als, H. & Adamson, L. (1979), Structure of early Face-to-Face Communicative Interactions, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 349–370.

Tyrrell, T. (1994), 'An Evaluation of Maes's Bottom-Up Mechanism for Behavior Selection', *Adaptive Behavior* **2**(4), 307–348.

van der Spiegel, J., Kreider, G., Claeys, C., Debusschere, I., Sandini, G., Dario, P., Fantini, F., Belluti, P. & Soncini, G. (1989), A foveated retina-like sensor using CCD technology, *in* C. Mead & M. Ismail, eds, 'Analog VLSI implementation of neural systems', Kluwer Academic Publishers, pp. pp. 189–212.

Velasquez, J. (1998), When Robots Weep: A Mechanism for Emotional Memories, *in* 'Proceedings of th 1998 National Conference on Artificial Ingelligence, AAAI98', pp. 70–75.

Vlassis, N. & Likas, A. (1999), 'A Kurtosis-based dynamic approach to gaussian mixture modeling', *IEEE transactions on systems, man, and cybernetics: Part A*.

Waters, K. & Levergood, T. (1993), DECface: an automatic lip synchronization algorithm for synthetic faces, Technical report, DEC Cambridge Research Laboratory, Cambridge, MA. Technical Report CRL 94/4.

Wolfe, J. M. (1994), 'Guided Search 2.0: A revised model of visual search', *Psychonomic Bulletin & Review* **1**(2), 202–238.

Wood, D., Bruner, J. S. & Ross, G. (1976), 'The role of tutoring in problem-solving', *Journal of Child Psychology and Psychiatry* **17**, 89–100.

Woodworth, R. (1938), *Experimental psychology*, Holt, New York.

Yoon, S., Blumberg, B. & Schneider, G. (2000), Motivation Driven Learning for Interactive Synthetic Characters, *in* 'Proceedings of Agents2000 (to appear)'.