11 A Bayesian Heart: Computer Recognition and Simulation of Emotion
Eugene Ball

11.1 Why Does a Computer Need a Heart?
Computers are rapidly becoming a critical and pervasive piece of our societal infrastructure. Within a decade or two they are likely to be the constant companions of most people in the industrialized world. Many of our interactions with them will continue to be like those with simpler machines: We push a button and the machine takes a limited and well-defined action, like zapping our food or washing the dishes. But there will also be many situations wherein spoken conversation will be the preferred means of communicating with a computer: perhaps to ask what the weather is likely to be in Vienna next week, or to select a good book to read on the plane.

There are huge technical challenges that must still be overcome before conversational computers will be competent and reliable enough to use in this fashion, but I have little doubt that we will get there in twenty years (possibly much sooner). A computer with which we engage in casual conversation (even if limited to narrow domains), will inevitably become a significant social presence (at least as noticeable as a human ticket agent with whom we carry out a brief transaction). I suspect that for many people, such a system will eventually become a long-term companion with which (whom?) we share much of our day-to-day activity.

To be useful, conversational interfaces must be competent to provide some desired service; to be usable, they must be efficient and robust communicators; and to be comfortable, they will have to fulfill our deeply ingrained expectations about how human conversations take place. One subtle aspect of those expectations involves the emotional sensitivity of our conversational partners. We would be surprised if an outburst of anger toward someone produced a completely deadpan response, and might even be further angered by their lack of acknowledgment of our own emotional state. If we laugh in response to someone’s joke, we expect them to laugh (or at least smile) along with us—to do otherwise would be disconcerting.
My work (with my colleague Jack Breese) on the computational modeling of emotion and personality is intended as a first step toward an emotionally sensitive conversational interface: one that can recognize the emotional state of the human user and respond in a fashion that adds to the naturalness of the overall interaction.

Communicating with Computers
People frequently find spoken conversation to be the most efficient and comfortable way to conduct interactions with others. Particularly for tasks requiring many back-and-forth steps, written communication (even e-mail) can be tedious and suffers from the need to reacquire working context for each step of the interaction. Graphical computer interfaces have been quite successful as a medium for conducting many well-specified tasks of this sort. For example, common requests to a travel agent (long a favorite target application of the spoken language research community) can be carried out quite efficiently as an interaction with a Web server. However, if a request is unusual, it may be difficult to build a graphical user interface to handle it without unduly complicating the interface for more common cases. “Hidden commands” can provide less common capabilities without complicating simple ones, but they require that the user know both that they exist, and how to find them.

One likely path for computer interface design is to gradually augment graphical user interfaces with linguistic capabilities. An “assistant” would accept flexible descriptions of commands or objects outside of the immediately visible workspace: “I’d like to reserve a group of fifty seats for travel from Minneapolis to Seattle next December” or “Check to see which movies are showing on flights from London to Seattle.” The ability to respond properly to powerful natural language requests (either typed or spoken) would be a welcome addition to current interfaces. While natural language may first be introduced as an “escape” for uncommon requests, its role is likely to steadily expand as it becomes more capable and as speech recognition becomes more reliable.

Natural language requests of the sort suggested above are convenient and powerful, but often result in ambiguities that require clarification: “Do you want round-trip tickets?” Therefore, spoken interactions won’t usually consist of just isolated commands but will become conversational dialogues.
Social Aspects of Computer-Human Interaction

In many fundamental ways, people respond psychologically to interactive computers as if they were human. Reeves and Nass (1995) have demonstrated that strong social responses are evoked even if the computer does not use an explicitly anthropomorphic animated assistant. They suggest that humans have evolved to accord special significance to the movement and sounds produced by other people, and at some deeply ingrained level cannot avoid responding to the communication of modern technology as if it were coming from another person.

As user interface designers, I believe we have a responsibility to try to understand the psychological reality and significance of these effects and to adapt our computer systems to the needs of our users. In addition, we should recognize that this social response is likely to become much stronger when the user is having a spoken conversation with a computer. It is clear that our emotional responses do not disappear while we are interacting with machines of all types: We get annoyed when they do not work properly, we respond with joy when a difficult task is completed smoothly, we even attribute human motives to their inanimate behaviors on occasion. Thus it will not be surprising to see even stronger emotional reactions to computers that talk, including expectations of appropriate emotional responses in the computer itself.

Emotionally Aware Computing

Explicit attention to the emotional aspects of computer interaction will be necessary in order to avoid degrading the user’s experience by generating unnatural and disconcerting behaviors. For example, early text-to-speech systems generated completely monotonic speech, which conveyed a distinctly depressed (and depressing) emotional tone.

Therefore I would argue that the initial goal for emotional interfaces should be to simulate appropriate emotional reactivity by demonstrating an awareness of the emotional content of an interaction.

The type of emotional reactivity that might be appropriately demonstrated by a conversational assistant can be illustrated by considering some imaginary computer responses to different situations. I’ve labeled each example with an emotional term or attitude description that could properly accompany the words,
giving them a more natural feel and a greater communicative potency.

The assistant is reporting the results of an assigned search task:
- I was unable to find anything meeting that description. (sadness)
- This is just what you were looking for. (pride)
- I'm not sure, but one of these might suit your needs. (uncertainty)
- Gee, and it only took me 12 minutes to find it! (embarrassment)

The assistant reacts to difficulties with the communication itself:
- I'm afraid I don't understand what you mean by that. (confusion)
- I believe I just told you that I don't know anything about that topic. (irritation)
- This doesn't seem to be going so well ... could you try again? (embarrassment)
- I'm really sorry, but could you repeat that one more time? (solicitude)

The assistant detects a strong emotional reaction from the user:
- Whoa ... Can we calm down and try again? (calming)
- Gee, that was pretty frustrating, wasn't it? (empathy)
- Great! Glad to be of help. (joy)

The assistant is trying to fulfill a user's request to help modify the user's behavior:
- Shouldn't you get back to work now? (disapproval)
- If you don't leave now, you'll be late. (warning)
- Take a break now, if you want your wrists to get better! (commanding)

Discussion

Picard: “I was unable to find anything meeting that description.” When you assume goodness and sincerity and so forth, sadness could actually be expressed with no tone of voice. But of course we could all hear these said in other ways as well. And that's where, I think, it's interesting to consider not just the emotion of the system that is reading these sentences, but the emotion of the perceiver of these sentences. There are several interesting studies in which you present perceivers with a neutral stimulus, and a perceiver in a good mood will perceive the neutral stimulus as more positive, while the perceiver in a bad mood will perceive it as more negative.
Ball: In human interaction we see that distinction in the perceiver, and we react to that.

Picard: That’s right. E-mail that we send without tone really is more likely to be perceived ambiguously, so we may need to go to even greater efforts to verbally try to be clear, if we are positive, or just hurried, as opposed to angry—things that could easily be confused without the tone.

Bellman: Washington, D.C. had a big controversy over the voice that they were using for closing the doors inside a subway. Did you hear that?

Picard: Actually, there was a similar thing in Atlanta many years ago at the airport.

Bellman: They purposely made the voice a little bit brisk. They wanted to get people to get on: Doors are closing—get on! And there was actually a tremendous backfire against it. People found it was rude, it was a nasty voice. They finally had to get rid of it.

Picard: It’s funny: In Atlanta, they had started with a nice, human sounding voice that sounded very friendly. And people didn’t pay much attention to it. So they went to a more computerized, synthetic-sounding voice, that sounded sort of more high-tech and cold.

The perceiver not just is influenced by their own emotions, but of course by what they think of that entity. They think, what is this computer that’s so stupid? Because so many people harbor these mixed feelings. We are very unusual in how we feel about computers compared to the rest of society.

Ball: And all these examples are imaginary. I think, having the competence to generate the right one of these in the right situation is a huge goal. And getting it wrong is something that people can react very strongly to.

Ortony: There are huge individual differences. I have a friend from New York who has a reputation of being rather brusque. I find this little story illustrative. He finds Chicago intolerable compared to New York. One of the things that irritate him in Chicago is: people get off the bus and they thank the bus driver. He finds this absolutely incomprehensible behavior, like “The goddamned guy is paid to drive the bus. What’s the problem? You get off the bus and you go!”—The point here is that people obviously have different personalities that require different interactional styles. Actually, some people will be upset by one style, while others will be sat-
isfied. I mean, the environment includes the personality of the individual one is interacting with.

Picard: It’s going to be constantly changing in different situations. So, if the computer tries one of these lines on your friend, and your friend trashes the thing, then the computer will better not try any lines similar to that.

In these examples, the character’s linguistic expression is the clearest indicator of its emotional state, but if that expression is to seem natural and believable, it needs to be accompanied by appropriate nonverbal indications as well. Whether generated by pre-authored scripts or from strong AI first principles, such utterances will seem false if the vocal prosody, hand gestures, facial expressions, and posture of the character do not match the emotional state expressed linguistically.

In order to produce responses demonstrating as much emotional sensitivity as these examples suggest, a system must be able to: recognize and/or predict the emotional state of the user, and then synthesize and communicate an appropriate emotional response from the computer.

The next section describes a simple emotional model that can be used to adjust the emotional expression of a talking computer. While the motivation for this work is strongest for conversational systems, its application may be appropriate more generally. As computer use becomes ever more widespread in our culture, it is likely that we will see greatly increased attention to the subjective experiences of computer users, including the aesthetic and emotional impact of computer use. My expectation is that the experience gained from modeling the emotional impact of spoken interfaces will also be used to inform the design (and possibly the dynamic behavior) of conventional graphical interfaces, in order to improve user satisfaction.

11.2 A Bayesian Model of Emotion

Modeling Emotion
The understanding of emotion is the focus of an extensive psychology literature. Much of this work is based upon a deep understanding of an individual’s beliefs about how events will effect
him, and then modeling the way those beliefs lead to an emotional response (Scherer 1984; Ortony, Clore, and Collins 1988).

While a few research efforts are attempting to build agents with sufficiently deep understanding that these models can be applied directly (Bates, Loyall, and Reilly 1994; Martinho and Paiva 1999), we have chosen to utilize a much simpler model of emotion; one that corresponds more directly to the universal responses (including physical responses) that people have to the events that affect them. Although this approach is unable to model many subtle emotional distinctions, it seems like a good match to conversational interfaces that communicate with people (within specific domains) using only a limited understanding of language and the user’s goals.

The term emotion is used in psychology to describe short-term (often lasting only a few seconds) variations in internal mental state, including both physical responses, like fear, and cognitive responses, like jealousy. We focus on two basic dimensions of emotional response (Lang 1995) that can usefully characterize nearly any experience:

- **Valence** represents positive or negative dimension of feeling.
- **Arousal** represents the degree of intensity of the emotional response.

Figure 11.1 shows the emotional space defined by these dimensions, and shows where a few named emotions fit within them. In our model, these two continuous dimensions are further simplified by encoding them as a small number of discrete values. Valence is considered to be either negative, neutral, or positive; similarly, arousal is judged to be excited, neutral, or calm.

Psychologists also recognize that individuals have long-term traits that guide their attitudes and responses to events. The term personality is used to describe permanent (or very slowly changing) patterns of thought, emotion, and behavior associated with an individual. McCrae and Costa (1989) analyzed the five basic dimensions of personality (see Wiggins 1979), which form the basis of commonly used personality tests. They found that this interpersonal circumplex can be usefully characterized within a two-dimensional space. Taking an approach similar to our representation of emotion, we have incorporated into our model a representation of personality based upon the dimensions of:
Dominance, indicating an individual’s relative disposition toward controlling (or being controlled by) others
Friendliness, measuring the tendency to be warm and sympathetic

Dominance is encoded in our model as dominant, neutral, or submissive; friendliness is represented as friendly, neutral, or unfriendly.

Given this quite simple but highly descriptive model of an individual’s internal emotional state and personality type, we wish to relate it to behaviors that help to communicate that state to others. The behaviors to be considered can include any observable variable that could potentially be caused by these internal states. In laboratory settings, some of the most reliable measures of emotional state involve physiological sensing, such as galvanic skin response (GSR) and heart rate. For both emotion and personality, survey questions are often used to elicit accurate measures of internal state (with tests such as the Myers-Briggs Type Indicator; Myers and McCaulley 1985). However, in normal human interaction, we rely primarily on visual and auditory observation to judge the emotion and personality of others.

A computer-based agent might be able to use direct sensors of physiological changes, but if those measures require the attach-
ment of unusual devices, they would be likely to have an adverse effect on the user’s perception of a natural interaction. For that reason, we have been most interested in observing behavior unobtrusively, either through audio and video channels, or possibly by using information (especially timing) that is available from traditional input devices like keyboards and mice, which might be a good indicator of the user’s internal state. More specialized devices like a GSR-sensing mouse or a pressure-sensitive keyboard might be worth investigating as well—although unless they turned out to be extraordinarily helpful, they are unlikely to make it into widespread use.

Bayes Networks
Bayesian networks (Jensen 1996) are a formalism for representing networks of probabilistic causal interactions that have been effectively applied to medical diagnosis (Horvitz and Shwe 1995), troubleshooting tasks (Heckerman, Breese, and Rommelse 1995), and many other domains.

Bayes nets have a number of properties that make them an especially attractive mechanism for modeling emotion. First, they deal explicitly with uncertainty at every stage, which is a necessity for modeling anything as inherently nondeterministic as the connections between emotion and behavior. For example, an approach using explicit rules (if behavior B is observed, then deduce the presence of emotional state E) would have great difficulty accounting for inconsistent reactions to the same events. However, a Bayes network will make predictions about the relative likelihood of different outcomes, which can naturally capture the inherent uncertainty in human emotional responses.

Second, the links in a Bayes net are intuitively meaningful because they directly represent the connections between causes and their effects. For example, a link between emotional arousal and the base pitch of speech can be used to represent the theoretical effect that arousal (and the resulting increased muscular tension) has on the vocal tract. It is quite easy to encode the expectation that with increasing arousal, the base pitch of speech is likely to increase as well. The exact probabilities involved can still be difficult to determine, but if the network is designed so that the parameters represent relatively isolated effects, relevant quantitative information from psychological studies is sometimes
available. Moreover, any model with enough complexity to model even simple emotional responses is likely to have a large number of parameters that have to be determined, and in a Bayesian network these parameters at least have clearly understandable meanings.

Finally, and especially relevant to the twin requirements of emotionally aware computing (recognizing emotion in the user, and simulating emotional response by the computer), Bayesian networks can be used both to calculate the likely consequences of changes to their causal nodes and also to diagnose the likely causes of a collection of observed values at the dependent nodes. This means that a single network (and all of its parameters) can be used for both the recognition and the simulation tasks.

When used to simulate emotionally realistic behavior of the computer, the states of the internal nodes representing dimensions of emotion and personality can be set to the values that we wish the computer to portray. The evaluation of the Bayes net will then predict a probability distribution for each possible category of behavior. This has the extra advantage that by randomly sampling this distribution over time, we can very easily generate a sequence of computer behaviors that are consistent with the desired emotional state, but are not completely deterministic. Because excessively deterministic behavior is a strong indicator of mechanistic origins, observers frequently judge that such behavior appears unnatural. By introducing some random (but consistent) variability, that source of unnaturalness can be avoided.

When the computer observes user behavior (through cameras, microphones, etc.) the observations can be used to set the values of the corresponding leaf (or dependent) nodes in the network. Evaluation of the network then results in estimated values for the internal dimensions of the user’s emotional state. The most probable value can be taken as the user’s state (as perceived by the computer). If multiple values have similar probabilities, the diagnosis can be treated as uncertain.

**Emotion and Behavior**

The Bayesian model that we have built (figure 11.2) contains internal states for emotional valence and arousal, and for the dominance and friendliness aspects of personality. These nodes are
Figure 11.2 A Bayesian network indicating the components of emotion and personality and various types of observable effects.
treated as unobservable variables in the Bayesian formalism, with links connecting them to nodes representing aspects of behavior that are judged to be influenced by that hidden state. The behavior nodes currently represented include linguistic behavior (especially word selection), vocal expression (base pitch and pitch variability, speech speed and energy), posture, and facial expressions. Our Bayesian network therefore integrates information from a variety of observable linguistic and nonlinguistic behaviors.

The static model described above can be extended to a version with temporal dependencies between current and previous values of the internal variables characterizing emotions. In this model, we assume that the values of the observable variables such as speech speed, wording, gesture, and so on are independent of emotions given the current emotional state.

The variables describing emotions evolve over time, and in this model the interval between time slices is posited to be three seconds. Valence, modeled by the variable $E$-Valence$(t)$ in the network, depends on valence in the previous time period, $E$-Valence$(t-1)$, as well as the occurrence of a ValenceEvent$(t)$ in the previous period. A valence event refers to an event in the interaction that affects valence. For example, in a troubleshooting application, a negative valence event might be a failed repair attempt or a misrecognized utterance. We have a similar structure for arousal, where the variable ArousalEvent$(t-1)$ captures external events that may affect arousal in the current period, with discrete states calm, neutral, and exciting. The conditional probability distribution indicating the dynamic transition probabilities is shown in figure 11.3. The distribution does not admit a direct transition from a calm state of arousal to an excited state. (Note: this distribution is illustrative, and not based on a formal study or experiment.)

Because personality, by definition, is a long-term trait, we treat these variables as not being time dependent in the model, hence the lack of a time index in the variable names for personality, $P$-Friendly and $P$-Dominant. Note that the existence of the personality variables in the model induce a dependency among the observables at all times, so the model is not strictly Markovian in the sense that observations are conditionally independent of the past, given the current unknown emotional state. However, this model can be converted to a Markovian representation for inference.
**Figure 11.3** Probability assessment for state of arousal at time $t$ given arousal at time $t - 1$ and the possible occurrence of an arousal event in the application or environment.

<table>
<thead>
<tr>
<th>Arousal Event[$t$]</th>
<th>E-Arousal[$t-1$]</th>
<th>Calm</th>
<th>Neutral</th>
<th>Excited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calming</td>
<td>Calm</td>
<td>.88</td>
<td>.12</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>.66</td>
<td>.34</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Excited</td>
<td>.45</td>
<td>.55</td>
<td>.00</td>
</tr>
<tr>
<td>Neutral</td>
<td>Calm</td>
<td>.25</td>
<td>.73</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>.07</td>
<td>.87</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>Excited</td>
<td>.02</td>
<td>.72</td>
<td>.26</td>
</tr>
<tr>
<td>Exciting</td>
<td>Calm</td>
<td>.00</td>
<td>.39</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>.00</td>
<td>.22</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>Excited</td>
<td>.00</td>
<td>.04</td>
<td>.96</td>
</tr>
</tbody>
</table>
In the next section, we discuss the architecture of the emotional component of a complete interactive system. Following that, we present more detail on the model’s treatment of specific behaviors—particularly linguistic and vocal expression.

11.3 Emotional Interactive Systems

In an emotionally aware interactive system, the recognition and simulation of emotion will play an auxiliary and probably quite subtle role. The goal is to provide an additional channel of communication alongside the spoken or graphical exchanges that carry the main content of the interaction. If the emotional aspects of the system call attention to themselves, the primary motivation of producing natural interactions will have been defeated. In fact, users that get the feeling that the system is monitoring them too closely may begin to feel anxious or resentful (of course, the emotional system, recognizing that fact, could always turn itself off!).

Because recognizing emotional behaviors is likely to require considerable effort and produce only a modest benefit, it will probably require that a single emotional component be shared among many applications in order to be practical. Therefore, another attraction of adopting a simple noncognitive model of emotion is the ability to keep the emotional component independent of most of the domain-aware portions of the system. If we observe and simulate emotional behaviors that are expressed automatically and unconsciously, then the recognition and interpretation of those behaviors can take place in an independent subsystem.

Thus we may well see the creation of just a few competing “emotion chips” that can be incorporated into many applications. These modules will be responsible for receiving sensory input and estimating the current emotional state of the user, selecting the emotional response from the system that will be most appropriate, and then modify the speech and animated behavior of the system in order to express the selected behavior in a natural way.

System Structure

The system architecture that we have experimented with is demonstrated in figure 11.4. In our agent, we maintain two copies of the emotion/personality model. One is used to assess the user’s
emotional state, the other to generate behavior for the agent. The model operates in a cycle, continuously repeating the following steps.

1. **Observation.** First, the available sensory input is analyzed to identify the value of any relevant input nodes. For example, a phrase spoken by the user might be recognized as one possible paraphrase among a group of semantically equivalent, but emotionally distinct, ways of expressing a concept. (The modeling of such alternatives is discussed in the next section.) In parallel, the vision subsystem might report its analysis that the user is currently producing large and fast gestures along with their speech. For each such perception, the corresponding node in the diagnostic copy of the Bayesian network is set to the appropriate value.

2. **Assessment.** Next, we use a standard probabilistic inference (Jensen 1989; Jensen 1996) algorithm to update the emotion and personality nodes in the diagnostic network to reflect the new evidence.

3. **Policy.** The linkage between the models is captured in the policy component. This component makes the judgment of what emotional response from the computer is desirable, given the new

![Figure 11.4](image-url)
estimate of the user's emotional state. Possible approaches to the policy component are discussed in the next section.

4. Simulation. Next, a probabilistic inference algorithm is applied to the second copy of the Bayes network. This time, the consequences of the new states of the emotion and personality nodes are propagated to generate probability distributions over the available behaviors of the agent. These distributions indicate which paraphrases, animations, speech characteristics, and so on would be most consistent with the agent's emotional state and personality (as determined by the policy module).

5. Behavior. Some agent behaviors can be expressed immediately; for example, instructions for changes in posture or facial expression can be transmitted directly to the animation routines, and generate appropriate background movement. Other behavior nodes act as modifiers on application commands to the agent. At a given stage of the dialogue, the application may dictate that the agent should express a particular concept, such as a greeting or an apology. The current distribution for the node corresponding to that concept is then sampled to select a paraphrase to use in the spoken message.

Policy

The policy module has not been explored very thoroughly at this point. In a working system it would likely be quite complex, taking into account the history of the dialogue with the user thus far (or at least its emotional trajectory), and a model of the particular user’s preferences, as well as the estimates from the personality and emotion nodes in the diagnostic network.

The imagined responses shown in the section entitled Emotionally Aware Computing illustrate a few of the difficulties. For example, at what point should a computer agent express irritation toward a user? Conversational systems frequently encounter explicit attempts to “break the demo.” The form of such an attack is sometimes sufficiently predictable that a clever response can be generated in an attempt to deflect it. If the user then persists in generating additional antagonistic input, perhaps an expression of irritation is the appropriate response.

Thus far, we have only considered two very simplistic policies. The empathetic agent tries to match the user’s emotion and personality. There is some evidence that people prefer to deal with a computer agent that is similar to themselves (Reeves and Nass
1995), so this might be a good starting point. Of course, it does lead to a possible positive feedback loop, particularly if the user becomes angry!

We have also experimented briefly with a contrary agent, whose emotions and personality tend to be the exact opposite of the user. While there are particular contexts in which this may produce interesting results—for example, when the user becomes bored or sad—it obviously is too simplistic to be a general policy.

Discussion

Bellman: So, is this policy box like the kind of thing in Eliza and other systems that we know, which would basically decide how friendly or how sympathetic the agent is?

Ball: Yes, it decides what its emotional state is. If the user is angry, then just being deadpan in response to that isn’t, I think, the right choice.

Ortony: The social interaction rules, essentially.

Ball: But there is a complex choice about this, given the history of the interaction and the long-term assessment of this user and all kinds of complex issues.

Picard: A whim to apologize, for example.

Ball: If you decide to apologize, for example, then you want the behavior or expression of the agent’s emotional state to be appropriate.

Bellman: The question I was asking actually was: It seemed to me that the policy you have represents the personality of the agent, and you could have your settings there. But then, you have your definition of the cultural interactions that are allowed, the conversational rules and other kinds of things that would be culturally determined. You wouldn’t sell the same agents and policies in Japan as you would here.

Ball: Sure.

Sloman: But at the moment they are collapsed into a policy box.

Ball: Well, it’s the policy box, and also outside the policy box. So, there’s something that’s controlling the overall interaction, the dialogue, deciding what to say, what choice to take.

Bellman: So, what I am saying is that it should have at least two boxes of that.
Ball: I am saying that this is just a little sideline to some whole other system which is doing the communication and the task and all of that. And so, this is just providing little, subtle modulation on the style of the interaction in order to try to make it feel more believable.

11.4 Recognition and Simulation

**Linguistic Behavior**

A key method of communicating emotional state is by choosing among semantically equivalent, but emotionally diverse paraphrases—for example, the difference between responding to a request with “sure thing,” “yes,” or “if you insist.” Similarly, an individual’s personality type will frequently influence their choice of phrasing—for example, “you should definitely” versus “perhaps you might like to.”

Our approach to differentiating the emotional content of language is based on behavior nodes that represent “concepts,” including a set of alternative expressions or paraphrases. Some examples are shown in table 11.1.

We model the influence of emotion and personality on wording choice in two stages, only the first of which is shown in the network of figure 11.1. Because the choice of a phrase can have a complex relationship with both emotion and personality, the prob-

**Table 11.1** Paraphrases for alternative concepts

<table>
<thead>
<tr>
<th>CONCEPT</th>
<th>PARAPHRASES</th>
</tr>
</thead>
<tbody>
<tr>
<td>greeting</td>
<td>Hello, Hi there, Howdy</td>
</tr>
<tr>
<td>yes</td>
<td>Yes, Yeah, I think so</td>
</tr>
<tr>
<td>suggest</td>
<td>I suggest that you, Perhaps you would like to, Maybe you could</td>
</tr>
<tr>
<td></td>
<td>Greetings, Hey, Absolutely, I guess so, For sure, You should, Let’s</td>
</tr>
</tbody>
</table>
lem of directly assessing probabilities for each alternative depending on all four dimensions rapidly becomes burdensome. However, inspired by Osgood’s work on meaning (Osgood, Suci, and Tannenbaum 1967), in which he identified several dimensions that can be used to characterize the connotations of most concepts, we first capture the relationship emotion and several “expressive styles.” The current model has nodes representing positive, strong, and active styles of expression (similar to Osgood’s evaluative, potent, and active), as well as measures of terseness and formality (see figure 11.5).

These nodes depend upon the emotion and personality nodes and capture the probability that individuals express themselves in a positive (judgmental), strong, active, terse, and/or formal manner. Each of these nodes is binary valued, true or false. Thus this stage captures the degree to which an individual with a given personality and in a particular emotional state will tend to communicate in a particular style.

The second stage captures the degree that each paraphrase actually is positive, strong, active, terse, and formal. This stage says nothing about the individual, but rather reflects a general cultural interpretation of each paraphrase: that is, the degree to which that phrase will be interpreted as positive, active, and so on by a speaker of American English. A node such as “GreetPositive” is also binary valued, and is true if the paraphrase would be interpreted as “positive” and false otherwise.

Finally, a set of nodes evaluates whether the selected paraphrase of a concept actually matches the chosen value of the corresponding expressive style. A node such as “GreetMatchPositive” has value true if and only if the values of “GreetPositive” and “wdsPositive” are the same. The node “GreetMatch” is simply a Boolean that has value true when all of its parents (the match nodes for each expressive style) are true. When using the network, we set “GreetMatch” to have an observed value of true. This causes the Bayesian inference algorithm to force the values of the nodes in the concept and style stages to be consistent. For example, when simulating the behavior of an agent, each style node (like “wdsPositive”) will have a value distribution implied by the agent’s personality and emotional state. The likelihood of alternative phrasings of a concept node (like “Greet”) will then be adjusted in order to produce the best possible match between its attributes and the style nodes. In this fashion, a negative emotional
Figure 11.5  A belief network fragment indicating the relationship of emotion and personality on expressive style, the probability that a modeled concept will be interpreted as a particular style, and whether the interpretation matches the intent for each component and whether they match on all components.
state will greatly increase the chance that the agent will select “Oh, you again” as a greeting.

In developing a version of this Bayes net for a particular application, we need to generate a network fragment, such as shown in figure 11.5, for each conceptual element for which we want emotional expression. These fragments are merged into a global Bayesian network capturing the dependencies between the emotional state, personality, natural language, and other behavioral components of the model.

The various fragments differ only in the assessment of the paraphrase scorings—that is, the probability that each paraphrase will be interpreted as active, strong, and so on. There are five assessments needed for each alternative paraphrase for a concept (the ones mentioned earlier, plus a formality assessment). Note that the size of the belief network representation grows linearly in the number of paraphrases (the number of concepts modeled times the number of paraphrases per concept).

In a previously proposed model structure, we had each of the expressive style nodes pointing directly into the concept node, creating a multistated node with five parents. The assessment burden in this structure was substantial, and a causal independence assumption such as noisy-or is not appropriate (Heckerman 1993). The current structure reduces this assessment burden, and also allows modular addition of new expressive style nodes. If we add a new expressive style node to the network (such as cynical), then the only additional assessments we need to generate are the cynical interpretation nodes of each concept paraphrase. These features of the Bayes network structure make it easy to extend the model for new concepts and dimensions of expressive style.

**Vocal Expression**

As summarized by Murray and Arnott (1993), there is a considerable (but fragmented) literature on the vocal expression of emotion. Research has been complicated by the lack of agreement on the fundamental question of what constitutes emotion, and how it should be measured. Most work is based upon either self-reporting of emotional state or upon an actor’s performance of a named emotion. In both cases, a short list of “basic emotions” is generally used; however, the categories used vary among studies.
A number of early studies demonstrated that vocal expression carries an emotional message independent of its verbal content, using very short fragments of speech, meaningless or constant carrier phrases, or speech modified to make it unintelligible. These studies generally found that listeners can recognize the intended emotional message, although confusions between emotions with a similar arousal level are relatively frequent. Using synthesized speech, in a 1989 MIT masters thesis, Janet Cahn (1989) showed that the acoustic parameters of the vocal tract model in the DECTalk speech synthesizer could be modified to express emotion, and that listeners could correctly identify the intended emotional message in most cases.

Studies done by the Geneva Emotion Research Group (Johnstone, Banse, and Scherer 1995; Banse and Scherer 1996) have looked at some of the emotional states that seem to be most confusable in vocal expression. They suggest, for example, that the communication of disgust may not depend on acoustic parameters of the speech itself, but on short sounds generated between utterances. In more recent work (Johnstone and Scherer 1999), they have collected both vocal and physiological data from computer users expressing authentic emotional responses to interactive tasks.

The body of experimental work on vocal expression indicates that arousal, or emotional intensity, is encoded fairly reliably in the average pitch and energy level of speech. This is consistent with the theoretical expectations of increased muscle tension in high arousal situations. Pitch range and speech rate also show correlations with emotional arousal, but these are less reliable indicators.

The communication of emotional valence through speech is a more complicated matter. While there are some interesting correlations with easily measured acoustic properties (particularly pitch range), complex variations in rhythm seem to play an important role in transmitting positive/negative distinctions. In spite of the widely recognized ability to “hear a smile,” which Tartter (1980) related to formant shifts and speaker-dependent amplitude and duration changes, no reliable acoustic measurements of valence have been found. Roy and Pentland (1996) more recently performed a small study in which a discrimination network trained with samples from three speakers expressing imagined approval or disapproval was able to distinguish those cases with reliability.
comparable to human listeners. Thus recognition of emotional valence from acoustic cues remains a possibility, but supplementary evidence from other modalities (especially observation of facial expression) will probably be necessary to achieve reliable results.

Our preliminary Bayesian subnetwork representing the effects of emotional valence and arousal on vocal expression therefore reflects the trends reported in the literature cited above, as follows:

With increasing levels of emotional arousal, we expect to find:

- Higher average pitch
- Wider pitch range
- Faster speech
- Higher speech energy

As the speaker feels more positive emotional valence, their speech will tend toward:

- Higher average pitch
- A tendency for a wider pitch range
- A bias toward higher speech energy

**Gesture and Posture**

Humans communicate their emotional state constantly through a variety of nonverbal behaviors, ranging from explicit (and sometimes conscious) signals like smiles and frowns, to subtle (and unconscious) variations in speech rhythm or body posture. Moreover, people are correspondingly sensitive to the signals produced by others, and can frequently assess the emotional states of one another accurately even though they may be unaware of the observations that prompted their conclusions.

The range of nonlinguistic behaviors that transmit information about personality and emotion is quite large. We have only begun to consider them carefully and list here just a few of the more obvious examples. Emotional arousal affects a number of (relatively) easily observed behaviors, including speech speed and amplitude, the size and speed of gestures, and some aspects of facial expression and posture. Emotional valence is signaled most clearly by facial expression, but can also be communicated by means of the pitch contour and rhythm of speech. Dominant personalities might be expected to generate characteristic rhythms and amplitude of speech, as well as assertive postures and gestures.
Friendliness will typically be demonstrated through facial expressions, speech prosody, gestures, and posture.

The observation and classification of emotionally communicative behaviors raises many challenges, ranging from simple calibration issues (e.g., speech amplitude) to gaps in psychological understanding (e.g., the relationship between body posture and personality type). However, in many cases the existence of a causal connection is uncontroversial, and given an appropriate sensor (e.g., a gesture size estimator from camera input), the addition of a new source of information to our model will be fairly straightforward.

Within the framework of the Bayesian network of figure 11.1, it is a simple matter to introduce a new source of information to the emotional model. For example, suppose we got a new speech recognition engine that reported the pitch range of the fundamental frequencies in each utterance (normalized for a given speaker). We could add a new network node that represents PitchRange with a few discrete values and then construct causal links from any emotion or personality nodes that we expect to affect this aspect of expression. In this case, a single link from Arousal to PitchRange would capture the significant dependency. Then the model designer would estimate the distribution of pitch ranges for each level of emotional arousal, to capture the expectation that increased arousal leads to generally raised pitch. The augmented model would then be used both to recognize that increased pitch may indicate emotional arousal in the user, as well as adding to the expressiveness of a computer character by enabling it to communicate heightened arousal by adjusting the base pitch of its synthesized speech.

11.5 Concerns

I think it is appropriate at this point to raise two areas of concern for research involving emotional response and computing.

Overhyping Emotional Computing

First, any mention of emotion and computers in the same breath gets an immediate startle reaction from members of the general public (and the media). Even if we were to avoid any discussion of the far-out questions (Will computers ever truly “feel” an emo-
tion?), I think we will be well advised to take exceptional care when explaining our work to others. The good news is that the idea of a computer with emotional sensitivity seems to be getting a lot of serious interest and discussion. Perhaps there is a shared (though unarticulated) appreciation that a device that can be so capable of generating emotional responses should also know how to respond to them!

However, the recent high level of interest may generate an unreasonably high level of expectation for technology that dramatically and reliably interprets emotional behavior. My personal expectation is that the significance of emotionally aware computing will be subtle, and will only reach fruition when spoken interaction with computers becomes commonplace. Moreover, as a technology that works best when it isn’t noticed, emotional computing should probably try to avoid the limelight as much as possible.

**Ethical Considerations**
When considering the ethics of emotional computing, there are many pitfalls. Some people may find the idea of ascribing an attribute as distinctively human as the communication of emotion to a computer as inherently objectionable. But even considering only more prosaic concerns, there are some potential uses of emotionally sensitive computing that clearly cross the boundaries of ethical behavior. Emotion could be a very powerful persuasive tool, if used effectively—especially if coming from the personification of your own computer, with which you may have a long and productive relationship. B. J. Fogg (1999) at Stanford University has begun to seriously consider these issues, under the name of computer aided persuasive technology (CAPTology).

I believe it would be irresponsible of us to pretend that some of our ideas will never be used to unethically persuade people. Con-artists have always been quick to adopt any available technology to defraud unsuspecting people. The age-old defense of “if we don’t do this, someone else will,” while true, doesn’t seem to me a sufficient response. I would be interested in hearing the thoughts of other workshop members on this topic.

My best idea is to try to develop a clear consensus on what exactly would constitute an unethical use of emotionally aware computing. If we could agree on that, we could quickly and vocally
object to unethical behavior (particularly commercial uses) when it occurs, and by announcing that intention in advance, perhaps dissuade at least some people from misusing our ideas.

**Discussion: The Ethics of Emotional Agents**

**Picard:** Let me tell you a short story. We made an experiment. We frustrated the test persons, and we built an agent that tries to make them feel less frustrated by using socially acceptable strategies. And it looks like we succeeded: The people who work with the agent show a behavior that was indicative of significantly less frustration. I was explaining this to some Sloane fellows, one of them from a very large computer company, and he said: No surprise to us. We found that, when we surveyed our customers who had had our product, that those who had had a problem with the product, found it defective, and had gotten this empathy-sympathy-active listening kind of response from customer service people—not from an agent!—were significantly more likely to buy our products again than those who bought a product and had no problems with it. And he said that they seriously confronted this as an ethical issue. Furthermore, all these years I have talked to visitors of our lab. Every one of them lately has raised the issue of how they have to get technology out faster, and they are now aiming not for it to have gone through alpha and beta cycles, and so forth, but they say, “60 per cent ready, and then it goes out there.” They are excited about the fact that getting something out that is defective and handling complaints about it better could actually lead to better sales.

**Ortony:** And the real ethical problem that I guess the guys are focusing on is that this is a sort of unknown property—when do you come in? If one were to explicitly say: “Whenever you encounter a bug in our software, some very nice agents are going to come and calm you down,” then you would not have an ethical problem. Then you think the ethical problem would go away. It is the fact is that this is “unsolicited” behavior from the system that’s problematic, I presume.

**Ball:** I am not sure if it would go away.

**Picard:** I am not so sure if it is unsolicited, either.

**Ortony:** Well, it diminishes, because, after all, you design cars with features to make people comfortable, but they are visible and they
are available for inspection prior to purchase, and so you don’t feel bad that you have made power seats as opposed to manual seats. On the other hand: What does a power seat do? It makes you feel all easy about changing the position of your seat, and all kinds of things that just make you feel better as a user. We don’t have a problem with that, presumably because it’s explicit and openly available and inspectable rather than requiring actual interaction with a product.

Ball: Right. And so, I think there is a deep problem about the emotional component, because it needs to be hidden in order to make sense.

Bellman: And there is a privacy issue somewhere, just in terms of the modeling that you do about the user, and who you pass it to, and what other kinds of reasons it is used for. Let me just mention the example of school children who were being taught how they could get their parents’ financial information back to companies.

Ball: If you have an agent that’s observing your behavior for a long period of time, he is going to know a lot about you. I think it is relatively easy to put a line and say: There is a lot of information, it just never goes out of your machine.

Picard: In an office environment, the company owns your machine and what’s on it. But in the future, you know, you might be much more comfortable with the mediator that you trust, operating between you and the office machine, if that mediator was as comfortable as your earring or your jewelry or something that you owned.

Sloman: Or it’s your personal computer, that you bring into the office and plug into the main one—rather than an earring.

Bellman: What I was pointing out is: Why should this be a secret from the patient or from the person who is being educated? Why can’t they have control over it? And that fits in with your wearable devices. In some sense, yes, you have this wonderful computer-based technology that allows this kind of collection about you, but you have control over it. That’s part of the solution.

The other thing is that a lot of the virtual world work is still highly effective even when it’s transparent to the user. Many of you seem to assume that it takes away from the mythology of the character if somehow people begin to lift the hood. In my experience we have just found the opposite. In really hundreds and hundreds of cases, letting people actually, for example, walk into your office,
and they meet your characters, and they actually look on the lifted
hood, see how it’s set up, see the way in which it works, and they
pick up issues about how responsive it is, or what it does, or what
it collects on them. We have not found that this knowledge would
actually take away the experience. It’s very empowering, and it’s
an interesting way of thinking about these control issues.

Sloman: These points are all about how much the end user has
access to information. And it’s quite important that often they can-
not absorb and evaluate the information if they get it themselves.
You may have to have third parties, like consumer associations
and other people who have the right to investigate these things, to
evaluate them, and then to publicize. If you are not an expert, you
go to someone you trust. That is not necessarily your earring or
your personal computer, but it might be another person or an organ-
ization who has looked at this thing, and you will be in a better
position. So, the information must be available.

Picard: We have to distinguish real-time, run-time algorithms from
store-up-plots of do-it-slowly algorithms. If you allow accumula-
tion, then we are getting to know somebody over a long period of
time, and you can build up their goals, values, and expectations,
all these other things that help predict the emotion, since it’s not
just what you see right now, but it’s also what you know about the
person. Whereas if you don’t keep all that person-specific memory,
you can only have “commonsense about prototypes about people,”
and then take what you observe at face value from this stranger, so
to speak. So, in the latter case, I believe, we can do without any
problems of privacy. We just build up really good models of what
is typical, and they won’t always be as good. But if you really want
the system to get to know you intimately, to know your values,
how you are likely to respond to the situation, there is going to
have to be some memory. We can’t do it all in the run-time. And
that’s an issue of privacy.

But we can go a long way without hitting the privacy. And once
we do with the privacy, there are a whole lot of possible solutions
to it.

References

Ball, G., Ling, D., Kurlander, D., Miller, J., Pugh, D., Sally, T., Stankosky, A., Thiel, D.,


