# Affective Computing

Rosalind W. Picard

The MIT Press

From The MIT Press

MITCogNet

First MIT Press paperback edition, 2000



This book was set in Stone Serif and Stone Sans by Windfall Software using ZzTEX and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Picard, Rosalind W.
Affective computing / Rosalind W. Picard.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-16170-1 (hc.: alk. paper)—978-0-262-66115-7 (pb.: alk. paper)
1. Human–computer interaction. 2. User interfaces (Computer systems). I. Title.
QA76.9.H85P53 1997
004'.01'9—dc21 97-33285
CIP

10 9 8 7

# 6 Recognizing and Expressing Affect

Emotions are like thoughts in that they rely upon words, gesture, music, behavior, and other creative forms of expression for their communciation. Affective communication occurs in the physical world through the senses, whether the message is conveyed through a sound pressure waveform, a visible motion, or via mediating instruments such as physiological sensors.[1] Emotions can be expressed voluntarily or involuntarily, in ways that are easy to control, or in ways of which a person may not be aware. Expressions may be publicly visible, for example a smile, or accessible only to someone in close physical contact, who feels your clammy hand. Emotions may also be communicated by behavior, as through loving actions. In each case, patterns of information are communicated, and these patterns can be represented in a computer.

This chapter casts affect recognition as a pattern recognition problem, and affect expression as pattern synthesis. Taking this approach, a variety of techniques become available for computer communication of emotion. The methodology and most of the tools used in this chapter can be found in textbooks on pattern recognition and modeling.[2] However, very little work has been done to apply these tools to affective patterns. In particular, little is known about which kinds of patterns tend to be the best indicators of a person's emotions, and how these patterns might be learned, recognized, and understood. The goal of this chapter is to lay a foundation for modeling affective patterns so that computers can be given the basic abilities of affect recognition and expression. This is a first step toward enabling them to interact more naturally with people, recognizing our emotions, and expressing emotions when appropriate.

### *Key Issues for Characterizing Affective Patterns*

One of the biggest questions in affect recognition is, "What are the couplings between affective states and their patterns of expression?" Numerous experimenters have proposed relationships, some of which hold across groups of individuals and some of which do not. There have also been debates over the years about whether or not characteristic bodily patterns accompany emotions. In particular, the work of Schachter and Singer in the early 60's argued that autonomic patterning only varies in intensity for different emotions, and that differentiation of emotion is not physical, but cognitive (Schachter, 1964). However, over the years, as technology and signal analysis have progressed, physiological patterns characteristic of emotions have been discovered. Cacioppo and Tassinary (1990) describe many cases where the specifics of data collection and analysis have made a big difference in the reliability of finding physiological patterns that differentiate emotions. This is not to say that the problem is easy to solve; it is not. Some signals are better at communicating emotions than others, and which is best can depend on the emotion, the person expressing it, and the conditions under which the emotion is elicited. One thing that is widely agreed upon is that no single signal is a trusted indicator of emotional response. Instead, patterns of signals are needed.

It is important to mention what kind of success we can expect. For example, it is not appropriate to expect computers to perfectly recognize all of your feelings. Most people have difficulty recognizing their own feelings and articulating them. Furthermore, the computer is an outside observer with limited access to your body and mind; it will not have the same information as you. It sees from the third-person viewpoint while you see from the first-person. It does not know everything you know about yourself. Its ability to recognize your emotions should at best be compared to the ability of another person to recognize your emotions. A reasonable criterion of success is to get a computer to recognize affect as well as another person, i.e., better than chance, but below 100% accuracy.

In some cases we can expect computers to perform better than people. In particular, with wearable computers and "smart clothing," the computer can continuously attend to physiological patterns, especially biosignals such as heart rate or muscular tension. Computers have superior abilities for processing patterns, although humans remain superior at interpreting meaning in patterns. The best results are likely to come from a combination of human and computer abilities. In particular, a person with a wearable affective computer will find an opportunity to learn things about himself or herself that might not be learned otherwise. I will address wearables in a later chapter.

Although I write of "recognizing emotions," I am not proposing that computers could recognize or measure affective states directly. Because affective states are internal and involve cognitive thoughts as well as physical changes, they cannot be fully recognized by anyone but the person having the affective state. Outsiders only have access to observable functions of the affective state—expressions, behaviors, and so forth. Given reliable observations of these functions, then the underlying states may be inferred. Hence, the expression "recognizing emotions" should be interpreted as "inferring an emotional state from observations of emotional expressions and behavior, and through reasoning about an emotion-generating situation." In particular, the pattern recognition tools in this chapter focus on modeling patterns of expression and behavior. The next chapter addresses models that can be used for reasoning about situations.

Despite its immense difficulty, emotion recognition is easier than thought recognition. Consider the party game of *charades*, where a player tries to get his team to guess a word or phrase—typically a person, place or thing—without providing any spoken or written clues. The fun and challenge involve trying to act out situations so that the team can guess the correct word or phrase quickly. Now, imagine if the game was limited to guessing emotions. In that case the player would no longer need the elaborate gestural syntax that the game has evolved ("3 syllables, name of a book, sounds like," etc.) and for most emotions the game would cease to be a challenge. Recognition of emotion is easier than recognition of thoughts largely because there are not as many emotions as thoughts. In pattern recognition, the difficulty of the problem almost always increases dramatically with the number of possibilities. The number of possible thoughts you could have right now is virtually limitless, nor are thoughts easily categorized into a small set of possibilities. Thought recognition, even with increasingly sophisticated brain imaging techniques, is arguably the largest recognition problem in the world. In contrast, a relatively small number of categories for emotions have been commonly proposed. The smaller set of categories permits a smaller language, making emotion recognition easier than thought recognition.

### *Basic Emotions and Discrete Categories*

Theorists have long discussed a small set of categories for describing emotional states. In 1962 Tomkins suggested that there are eight basic emotions: fear, anger, anguish, joy, disgust, surprise, interest, and shame (Tomkins, 1962). Plutchik later distinguished among a different eight basic emotions: fear, anger, sorrow, joy, disgust, surprise, acceptance, and anticipation (Plutchik, 1980). More recently, Ortony, Clore, and Collins have collected a summary of lists of basic emotions (Ortony, Clore and Collins, 1988). From these

lists, the most common four emotions (combining near synonyms, like joy and happiness) are fear, anger, sadness, and joy. The next most common two are disgust and surprise and, after these six, the lists diverge. Over the years, various researchers have proposed that there are from two to twenty "basic" emotions.

"Basic emotions" may be defined in many ways. Perhaps the most thorough definition has been given by Paul Ekman, who has linked basic emotions to those which have distinctive universal facial expressions associated with them, as well as eight other properties (Ekman, 1992, 1992a). By these criteria, Ekman identified six basic emotions: fear, anger, sadness, happiness, disgust, and surprise. Basic emotions can also be deduced by analyzing words for emotion, an approach taken by Johnson-Laird and Oatley on 590 English terms describing emotions, which concluded that the words could be based on one or more of five basic emotions: fear, anger, sadness, happiness, and disgust (Johnson-Laird and Oatley, 1989).

Whether or not basic emotional states exist is disputed by some authors, and is a topic of long-standing debate in the emotion theory literature (Ortony and Turner, 1990; Stein and Oatley, 1992, Panksepp, 1992). Some emotions show up universally, and others seem to involve cultural specifics. Universality poses only a slight problem to computers trying to recognize emotions, which I will address briefly below. Affective computing, fortunately, does not hinge on the resolution of whether or not there are basic emotions. Rather, the topic concerns us primarily as a problem of representation: should emotions be represented as discrete categories, or otherwise?

### *Emotion Spaces and Continuous Dimensions*

Some authors have been less concerned with the existence of eight or so basic emotions and instead refer to continuous dimensions of emotion (Schlosberg, 1954). Three dimensions show up most commonly, although only the names of the first two are widely agreed on. The two most common dimensions are "arousal" (calm/excited), and "valence" (negative/positive). These were the axes illustrated earlier in Fig. 3.2, together with titles of pictures classified in this continuous space, according to the work of Peter Lang. Lang has assembled an international archive of imagery rated by arousal and valence (Lang, 1995).

Numerous researchers have worked with dimensions of emotion instead of with basic emotions or discrete emotion categories. Lang writes that self-reports across subjects are more reliable with respect to dimensions than with respect to discrete categories such as anger, fear, etc. (Lang, 1984). A number of researchers have also proposed various mappings between continuous dimensions of emotions and basic emotion categories. In the next chapter

we will see several "cognitive appraisal" models that effectively do this, stating criteria that partition a continuous space into ten or more discrete outcomes. In general, two dimensions cannot be used to distinguish all the basic emotions; for example, intense fear and anger lie in the same region of high arousal negative valence. However, these two dimensions do account for the most common descriptions of mood.

The lack of a definition of emotion, and the lack of agreement on whether there are basic emotions or continuous spaces of emotions are obstacles to the goals of computer-based recognition and synthesis. However, these obstacles are not insurmountable. Similar hindrances occur in fields such as image content analysis where, despite the difficulties, pattern analysis and learning tools have proved helpful. Hence, it is reasonable to expect similar success in modeling affective patterns. Moreover, the question of whether to try to represent emotions with discrete categories or continuous dimensions can be considered a choice, as each representation has advantages in different applications. The choice of discrete or continuous states is, in one sense, like the choice of particles or waves in describing light: the best choice depends on what you are trying to explain.

If desired, discrete categories of emotions can be treated as regions in a continuous space. Categories may be "fuzzy" in the sense that an element can belong in more than one category at once. For example, a feeling of sadness can occur in both "grief" and "melancholy." Researchers such as Paul Ekman define affective phenomena that are not basic emotions, such as "grief," to be not emotions, but "emotional plots." The plot of grief specifies two actors with a prior relationship of attachment, a deceased and a survivor, and an event of separation, followed by emotions in the survivor such as distress, sadness, and perhaps fear or anger. Alternatively, one might consider grief a cognitively-generated emotion, or perhaps a mixture of more basic emotions. In the game of charades, basic emotions are easiest to portray, and emotions like grief involve more effort.

The piano teacher application described in the opening chapter of this book used discrete emotions, recognizing states of interest, frustration, and joy, to allow the teacher to give more personal feedback. In contrast, an application involving television news broadcasts is naturally suited to description with the arousal and valence axes. A high-arousal story captures attention: many people rush to the television to see the emotional gold-medal Olympic victory. Extremely negative content has a powerful influence on memory, perhaps in part because it is almost always also high arousal: many people have a keen memory of the shock they felt upon hearing of John F. Kennedy's death, or upon hearing of the space shuttle Challenger exploding. High-arousal stories attract viewers; however, people tend to not want too much

negativity, so it becomes important for broadcasters to try to find positive-valence high-arousal content. Valence and arousal are critical dimensions in entertainment, as well as in many other applications.

However, just because this representation is useful in this application does not imply that all emotions are continuously valenced. Neither does successful representation with a small set of discrete emotions imply that emotions are discrete, or that there is only a small set of them. Both representations have uses and limitations. Fortunately, we do not need consensus about one representation being "right" to carry out the ideas presented below.

In summary, the recognition and modeling problems are simplified by either the assumption of a small set of discrete "basic" emotions, or by the assumption of a small number of dimensions. The fact that both yield a concise representation is an advantage. Even if these are later found to be an oversimplification, they at least form a good point to begin the modeling effort. A small repertoire of emotions is characteristic in developing humans—the younger baby has a smaller repertoire of emotions than does a child, and the child a smaller repertoire than an adult. One can expect the first affective computers to start with only a small number of categories or dimensions.

### *Universal vs. Person-Specific*

Much of emotion theory has been stymied on the issue of universality. If there are emotions that occur with similar physiological responses in all humans, then what is this set of emotions and how can they be recognized, regardless of race, gender, culture, etc.? Like many questions in emotion theory, the study of this one is complicated by factors such as how emotion is defined, elicited, expressed, and communicated. Different languages do not necessarily use the same words for describing emotive phenomena, which further complicates attempts to demonstrate universality.

One of the potential benefits of affective computing lies in its ability to make measurements and analyze patterns of affective signals, conditioned on individuals and on circumstances affecting them. Given similar conditions, measurements, and patterns of responses, conclusions can begin to be made about the universality of various kinds of affective expressions. Hence, the solution proposed by affective computing is, first, person-specific—measuring data for individuals of all kinds, and, second, universal—examining the individual data to see what common patterns are present.

Common patterns are expected for universal emotions, and may differ slightly for emotions that are variations on these. For example, over 60 expressions of anger have been found, but all members of the anger family include two features: the brows are lowered and drawn together, the upper

eyelid is raised and the muscle in the lips is tightened. Variations on this basic anger expression are hypothesized to reflect whether the anger is controlled, spontaneous, simulated, and so forth (Ekman, 1992). Because these variations tend to occur with different frequencies in different individuals, and because they may invoke various other individual responses, perhaps reinforced by a person's local environment, they can take on additional flavors, much like a language evolves into dialects. Individual factors such as temperament affect thresholds of expression, as well as other physiological characteristics. Just as speakers of the same dialect have individual variations, we can expect temperamental variations in emotional expression.

### *Pure vs. Mixed*

After Uta Pippig won the 100th Boston Marathon, she described feeling tremendously happy for winning the race, surprised because she believed she would not win, somewhat sad that the race was over, and a bit fearful because during the race she had acute abdominal pain. We say she had "mixed emotions." However, emotion theorists do not agree on what it means for emotions to mix. Do they mix together like paints, like chemical compounds, or perhaps according to some mathematical function?

Here are two metaphors for how emotions might be "mixed": first, a microwave oven, and second, a tub of water. Microwaves usually have two pure states: "on" and "off." When you set the oven to cook at high then the oven is on constantly. When you set the oven to cook at medium, then the oven cycles between "on" and "off" to produce a slower heating effect. The state "medium" is created by juxtaposing pure states "high" and "off" in time—mixing in time—even though at any instant of time the oven is in only one state. A different case of mixing is illustrated by a tub of water. If you enjoy a warm bath, then you do not do so by jumping in time back and forth between a tub of cold water and a tub of hot water, but you mix the cold and hot water in the same tub. This kind of mixture allows the states to mingle and form a solution that has a new state—warm.

When examined over a long time scale, both the microwave and the tub result in a mixture state, "warm." However, in the microwave, one can argue that the purity of the states is preserved—you just have to look (or sample the data) quickly enough to detect them. In contrast, in the tub the purity of the cold and hot states is replaced by a warm state. For emotion mixing, both metaphors are useful. For example, Clynes has found in sentograph measurements of finger pressure that an expression of melancholy begins with a form that looks like love and ends with a form that looks like sadness (Clynes, 1977). In other words, the mixture emotion of melancholy is described as a juxtaposition of two forms in time, as in the microwave metaphor. On the

other hand, most theorists have proposed scenarios that are closer to the tub metaphor, with examples such as feeling "wary," which is hypothesized to be a mixture of interest and fear.

"Love–hate" relationships are an example where feelings of love and hate cycle in time. The result is not a simple sum of the two emotions, or a feeling that is in-between love and hate, but a rapid switching between the two in time. In fact, for certain pairs of emotions such as love and hate, or perhaps sadness and joy, it may not be possible for them to truly co-occur at the same time. Instead, it may be that their polarity limits their mixing to be like that of the microwave, one on and the other off, with mixing only in time.

All mixed emotions need not mix in the same way. In fact, this is a logical prediction based on the way emotions coincide with different patterns of bodily responses, and arise with different mechanisms. To the extent that two emotions have non-overlapping generative mechanisms, and their bodily patterns can mingle, then they can coexist in time. But if they require the same generative mechanisms, then only one of them can be generated at a given instant. Alternatively, two emotions generated by the same mechanism may have different lengths of decay. If the second is initiated before the first decays, this can give a different kind of overlap in time. However, given that emotions are short events, this overlap should not be significant. With this reasoning one can predict that a primary emotion like fear, generated initially in the amgydala, could coexist with a cognitively-generated state like anticipation, although extreme fear is likely to temporarily override any cognitive emotions.

Cognitive events can interfere with the purity of emotions. If you are deeply involved in playing a mournful piece of music, you may attain and express a pure state of sadness. However, if your mind wanders to a happy event that you are looking forward to after the concert, then the mournfulness of your playing will not be as pure. This kind of mixing, like the microwave cycling off and on, dilutes the expression of an emotion.

Emotions and cognitions can inhibit other emotions. An intriguing experiment on lying and emotional expression illustrates this inhibition. Thirty-one subjects were asked to express anger or love, using a sentograph. The device recorded two significantly different kinds of essentic forms corresponding to the two emotions. Next the subjects participated in several trials that required them to lie at various points about cards they were holding in their hand. When lying while expressing anger, no significant changes were found either in the subjects' self reports of anger, or in their recorded expressive waveforms of anger. However, when they were asked to lie while expressing love, not only were their self reports of love significantly lower, but their essentic waveforms for love were significantly altered (Clynes, Jurisevic, and Rynn,

1990). This suggests that certain cognitive events such as lying can inhibit certain emotional expressions (love), and not others (anger).

How does sentic modulation change as a person suppresses one strongly-felt state and tries to feel another? Could measurements of affective patterns help people identify an emotion they are masking, such as when anger is expressed to hide fear? Questions such as these can be addressed by the tools presented here. Using a computer with the ability to record and analyze observations that correlate with affective states should aid investigators in understanding these connections between emotions and their expression.

Imagine an actor who feels angry the night of a show, but has to play the role of a joyful character. In order to deliberately express joy he suppresses his anger, or overrides it with joy. If he is successful onstage in communicating joy, has he merely "forgotten" his anger, so that it will return after he has finished his time on stage? Or is there a therapeutic effect that takes place? Measurements of his emotion before, during, and after the performance could be studied both for understanding purity of emotions and for understanding their therapeutic effects. The measurements could be combined with reports from both the actor and audience, to gather their subjective (cognitive/perceptual) evaluations for synchronization with the bodily measurements.

If the actor has merely "forgotten" his anger, then this suggests a cognitive act, which has to occur both consciously and subconsciously so that the bodily response disappears. Otherwise, the audience will still see conflict in the actor, instead of joy, and think him to be a bad actor. The actor who is angry and tense in his body cannot merely think "smile" and appear carefree and light. The will does not have a monopoly on memory; the body also provides a short term memory. The muscles store tension; the posture can remain uptight. The intensity of affective communication is not only a function of thoughts, but a function of bodily modulation—voice, face, posture, and more. As the actor deliberately brings all these modes into a consistent expression, not only is his communication more effective, but he moves himself closer to a pure state of emotion. The purer the emotional state, the more powerful will be the ability of the actor to move the audience to a similar state. Theories that examine the purity of emotions through their power to be expressed bodily become empirically testable with an affective computer that can model emotional states for synthesis and recognition.

### *Modeling Affective Patterns*

Below I will give examples of computational models for the representation of affective patterns, especially for facial expressions, vocal intonation, and physiological signals that vary with affective states. Most of the models were

developed for the purpose of recognizing affective expressions, although some can be used for synthesis of expressions as well. All the models work with present technology and would typically be implemented in software. The models below tend to range from "low-level" to "mid-level," mapping emotions to signal patterns (expression generation or synthesis) and vice-versa (expression recognition). Some of the models assume discrete emotion categories, while others assume continuous dimensions of emotion. None of them are "high-level" in that none consider the semantics of the situation which might generate an emotional response in the first place (as necessary in cognitive emotion generation). The models that exist currently for such high-level processing are rule-based and connectionist models, which will be presented in the next chapter.

A caveat is in order before proceeding: sometimes the term "model" refers to a formula that is capable of fully explaining a phenomenon, both analyzing it and synthesizing it. In the richest sense, a pattern model can both recognize and synthesize the pattern. The use of the word "model" in this book is less narrow. Most of the models below cannot both synthesize and analyze the affective patterns without further development. Some consist of sets of features which discriminate expressions, but which cannot reliably synthesize them. Others can synthesize certain affective expressions, but do not provide parameters for recognition. Nonetheless, the term "model" is used to describe a set of parameters and procedures that are useful for pattern analysis, synthesis, or both.

There is a common misunderstanding that there is one right model of something, and that if there is more than one model, then they cannot all be right. On the contrary, experience has shown that the best choice of model depends on the application, and that there can be many right models just as there are many applications. Each model has its strengths and weaknesses, and sometimes a skillful combination of models gives better results than any single model. These principles have been found to be true in pattern modeling for video and image (Picard, 1996) and can also be expected to hold for pattern modeling of affective information. In other words, which computational model is "best" depends on the specifics of the computer's affective task, and when these change, so does the model. Therefore, equipping a computer with multiple models may be the way to get the best performance. Choosing which models are best for an application is easier after seeing different examples of each model's performance. I will therefore discuss many models below.

### *Recognizing and Synthesizing Facial Expressions*

One of the postulates of affective computing is that computers can be given the ability to recognize emotions as well as a third-person human observer. Let us consider the special case of facial expressions. Recognizing a facial

expression is not always the same as recognizing the emotion that generated it; facial expressions are the most easily controlled of all the expressions. However, because they are also the most visible, they are very important, and it is wise to observe them to assess what a person is trying to communicate. Some of the examples below involve models of facial expression which are not restricted to recognition, but which may also be used for synthesis of facial expressions.

Models for recognizing facial expressions have traditionally operated on a digitized facial image or a short digital video sequence of the facial expression being made, such as neutral, then smile, then neutral. In general, recognition from video is more accurate than recognition from still images. Video captures facial movements that deviate from a neutral expression. Therefore, the models below are based on recognition from video, although there has also been work on recognition of facial expressions using still images.[3]

Facial expression recognition models to date have treated emotions as discrete in the sense that they try to classify facial expressions into a small number of categories such as "happy" or "angry." The underlying theory that links the expressions to these categories was developed by Paul Ekman and his colleagues, and is called the Facial Action Coding system (FACS). The FACS system describes basic emotions and their corresponding sets of *action units*, which are muscular movements used to generate that expression.[4]

Facial expression recognition from video involves capturing spatiotemporal patterns of both local and global changes on the human face, and relating these patterns to a category of emotion. In the recognition examples that follow, two main assumptions are made: (1) there are a small number of discrete categories of emotional expressions; (2) data in the experiments is "pure" in the sense that a user willingly or naturally tried to express exactly one emotion. The first assumption makes this a supervised pattern recognition problem, with *a priori* specified categories of what can be recognized. The second assumption is perhaps the most problematic, as it cannot be verified. There is no guarantee that the facial expression recognized as "sad" corresponds to any genuine affective state of sadness.

None of the methods I describe claim to recognize the underlying emotion, but only the expression on the user's face. In other words, they would recognize your smile even if it is a forced smile when you are not feeling happy. They are currently not good enough to tell a false smile from a genuine smile although, to my knowledge, people have not tried very hard to get a computer to discriminate these cases, and a computer could be capable of this discrimination. Vision-based facial expression recognizers would also fail to recognize a state of joy if the joyful person suppressed all facial expressions. However, the models here have made strides in recognizing facial expressions,

which is a significant step toward giving computers the ability to recognize emotions.[5]

Irfan Essa of the Georgia Institute of Technology and Alex Pentland of the MIT Media Laboratory, have augmented Ekman's FACS system to address two of its limits: (1) action units are purely local spatial patterns; in contrast, real facial motion patterns are almost never completely localized and can include coarticulation effects, and (2) most facial actions occur in three phases: application, release, and relaxation, while FACS does not include such time components. In extending it to non-local spatial patterns and to include temporal information, they have enabled computers to recognize facial expressions from video (Essa and Pentland, 1997). The representation they use is based on representing facial motion dynamics during expression. It can also be used to synthesize facial expressions (Essa, 1995). The model contains both geometric information about facial shape and physical knowledge of facial muscles. It begins by fitting a representation of finite elements to the facial geometry, which then interacts with facial muscles to allow expressions to be synthesized according to the muscles that they involve. To synthesize an expression, values of parameters of the finite element representation for the desired expression need to be determined. These parameters can be calculated by analysis of a video of an expression. The parameters derived from the video sequence correspond to a pattern of peak muscle activations, which are mapped to an emotion category. The facial recognizer typically takes five minutes to process a facial expression (on an SGI Indy R4400) and has a demonstrated accuracy of 98% in recognizing six facial expression categories (anger, disgust, happiness, surprise, eyebrow raise, and neutral) for a group of eight people who deliberately made those expressions.

If faster recognition is needed, then a second, non-physically based model can be used, forming templates of facial motion energy (Essa and Pentland, 1995). The templates are default patterns characterizing the movement at each point between pairs of frames in the video while an expression is made. For the categories of anger, disgust, happiness, surprise, and neutral, recognition rates are as high as 98% in a test involving eight people. Studies are underway to determine how the recognition rate changes when there are more people. The recognition does not work in real time yet; it takes a few seconds to recognize each expression. However, with advances in hardware and pattern recognition, the recognition should become fast enough for an interactive response in the near future.

A different model for facial expression recognition, developed by Yaser Yacoob and Larry Davis of the University of Maryland, also relies on templates of motion energy, but uses a combination of templates and smaller subtemplates (e.g., of just the mouth area) and combines them with rules to

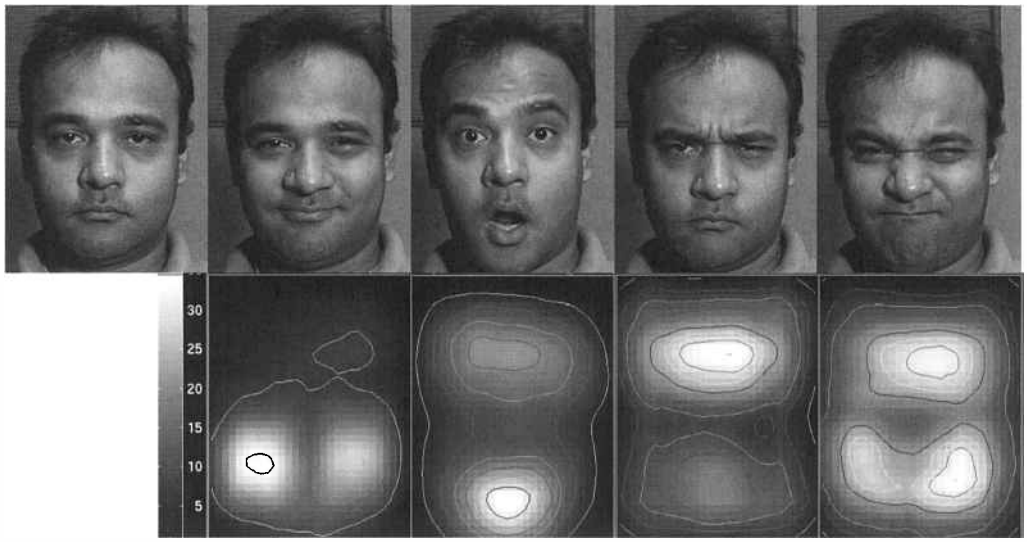


**Figure 6.1**
Facial expression recognition and motion energy maps. Top row: snapshots of the neutral and four other expressions: happiness, surprise, anger, disgust. Bottom row: templates of energy of the facial movement, as different from the neutral expression. The brighter regions correspond to higher energy. (Photographs courtesy of Irfan Essa, Georgia Institute of Technology, copyright 1997.)

formulate expressions (Yacoob and Davis, 1996). For example, templates are extracted of the eye and mouth area, and anger is characterized by inward lowering motion of the eyebrows coupled with compaction of the mouth. This method has been tested on expressions of fear, anger, sadness, happiness, disgust, surprise, and eye blinking, as made by 32 people, for a total of 116 expressions and 106 blinkings in the test set. The recognition accuracy over this database was approximately 65% for blinking, and approximately 80% for the affective expressions. This method is also not real-time, because computing the motion flow is slow. Yacoob, working with Michael Black of Xerox PARC, has developed a similar method that additionally uses camera motion tracking to help recognize expressions in videos of television talk shows, news, and movies (Black and Yacoob, 1995).

The above models use pattern recognition and image analysis, and inherit the current weaknesses of these tools. Most of the methods are sensitive to scene lighting, requiring it to be relatively uniform. All require the person's head to be easily found in the video sequence. Finally, continuous expression recognition, such as a sequence of "smile, frown, surprise," is not handled well; instead, the expressions must either be manually separated, or interleaved with some reliably detectable cue such as a neutral expression, which

has essentially zero motion energy. Continuous expression recognition is difficult in the same sense that continuous speech recognition is difficult—finding the word boundaries, or in this case the expression boundaries, needs to happen simultaneously with identification of the expressions.

When the computer synthesizes a smiling face, the computer may or may not also activate an internal affective state. The way this activation works in people is unknown, but it is true that expression of emotion plays a role in activation and regulation of emotional feeling. A facial expression can elicit an emotion in the person making the expression (Izard, 1990; Ekman, 1993), as well as in the recipient of the expression. When the computer recognizes a smiling face, it is possible to have this recognition influence the generation of an internal affective state—a sort of "emotion contagion" that could be given to computers.

The relation between temperament and facial expression has not been addressed by any of the facial expression recognition models. Facial expressiveness, like other forms of sentic modulation, is influenced by a person's innate physiology, which is related to temperament. In studies of inhibited versus uninhibited children, the inhibited ones have lower overall facial expressiveness—presumably a consequence of their tendency toward greater muscle tension (Kagan, Snidman, Arcus, and Reznick, 1994). Hence, their "baseline" facial dynamics operate over a smaller range. For optimal performance, computer systems that recognize facial expression would first have to calibrate the subject's expressive range, a form of "getting to know" them, before these systems could become adept at recognizing their expressions.

### *Synthesizing and Recognizing Affective Vocal Intonation*

Traditional efforts in computer-based speech recognition have focused on recognition of *what* is said. More recently, efforts have also been made to teach computers to recognize *who* is speaking. Usually the subtle qualities of *how* something is said have been treated as noise for the first two problems. In contrast, humans learn to identify who is talking and how something is said long before they can recognize what is said.

The vocal intonation of *how* something is said breaks down into two components: cues emphasizing which content in the message is most important, and cues arising from the speaker's affective state. Affective cues can convey the most important aspect of what is said, such as whether the speaker liked something or not. Vocal inflection adds flavor to our speech and content to its message. Even in telling a joke, everyone knows it's *how* you tell it that greatly determines its success.

Characterizing affect in speech may be harder than characterizing affect on faces. Facial signals communicate personal identity and expression, but

**Table 6.1**
Summary of human vocal effects most commonly associated with the emotions indicated. Descriptions are given relative to neutral speech. (Adapted with permission from Murray and Arnott (1993), Table 1. Copyright 1993 Acoustical Society of America.)

| | Fear | Anger | Sadness | Happiness | Disgust |
|---|---|---|---|---|---|
| Speech rate | much faster | slightly faster | slightly slower | faster or slower | very much slower |
| Pitch average | very much higher | very much higher | slightly lower | much higher | very much lower |
| Pitch range | much wider | much wider | slightly narrower | much wider | slightly wider |
| Intensity | normal | higher | lower | higher | lower |
| Voice quality | irregular voicing | breathy chest tone | resonant | breathy blaring | grumbled chest tone |
| Pitch changes | normal | abrupt on stressed syllables | downward inflections | smooth upward inflections | wide downward terminal inflections |
| Articulation | precise | tense | slurring | normal | normal |

do not generally communicate a linguistic message. On the other hand, the speech signal contains a mixture of information, including cues to speaker identity, affect, and lexical and grammatical emphasis for the spoken message. Isolating affective information is complicated. Nonetheless, computers are slowly achieving progress in synthesizing and recognizing affect in speech. Examples illustrating progress are provided in this section, although for further information the reader may refer to the overviews of the principal findings on human vocal emotion (Murray and Arnott, 1993; van Bezooyen, 1984). Table 6.1 summarizes the vocal effects most commonly associated with five basic emotions.

The basic problem that needs to be solved is: what is a good computational mapping between emotions and speech patterns? Specifically, we need to find features that a computer can extract, and models it can use to recognize and synthesize affective inflection. These features are generally derived from observing how voices change with emotions. When a speaker is in a state of fear, anger or joy, then his speech is typically faster, louder, and enunciated, with strong high-frequency energy. This is primarily due to arousal of the sympathetic nervous system, increasing heart rate, blood pressure, mouth dryness, and certain muscle activation. When the speaker is bored or sad, then his speech is typically slower and lower-pitched, with very little high-frequency

energy. This is primarily due to arousal of the parasympathetic nervous system, decreasing heart rate and blood pressure, and increasing salivation. In other words, the effects of emotion on speech show up primarily in its frequency and timing, with secondary effects in its loudness and enunciation. The effects of emotion therefore tend to show up in features such as average pitch, pitch range, pitch changes, intensity contour, speaking rate, voice quality, and articulation. However, these effects are complicated by prosodic effects that speakers use to communicate grammatical structure and lexical emphasis; both effects influence several of the same features.

Speech, like other forms of sentic modulation, is influenced by factors such as temperament and cognition. In studies with inhibited versus uninhibited children, those who were inhibited spoke with less pitch period variation in their voices, most likely because of their tendency toward increased muscle tension, as was also correlated with lower facial expressiveness (Kagan, Snidman, Arcus, and Reznick, 1994). For optimal performance, computer systems that recognize affect in speech would first have to learn the subject's vocal range, and then analyze with respect to this range. People are also capable of controlling their speech inflection willfully, although vocal expressions are harder to control than facial expressions. For example, the ability to mask nervousness in public speaking is important—many great speakers admit to being nervous, but they are able to learn to relax their voice in such a way that the nervousness is not heard. The models described below do not incorporate the influences of variables such as temperament, cognitive suppression of emotion, or linguistic content; however, they are pioneering in their attempts to begin to learn mappings between acoustic features and affective states.

The first model was constructed to address the question: Can recognizable affect be generated in computer-synthesized speech? To answer this, Janet Cahn, at the MIT Media Lab, built the "Affect Editor," a computer program that takes an acoustic and linguistic description of an utterance and generates synthesizer instructions for a DECtalk3 synthesizer to produce speech with a desired affect (Cahn, 1990). She identified values of seventeen parameters: six pitch parameters, four timing parameters, six voice quality parameters, and one articulation parameter, which produced speech that sounded scared, angry, sad, glad, disgusted, and surprised. The seventeen parameters were used to control a wide variety of affects—not just for strongly distinguishable emotions, but also for subtle differences, with variations for individuality. To synthesize speech, Cahn's model cooperates with models of the other components of speech to drive a synthesizer. This involves not just the seventeen parameters above, but also an analysis of the syntactic and semantic clauses of the utterance in an effort to identify good locations (e.g., pitch accent and pause locations) for applying both lexical and non-lexical effects.

To test this model of affective speech synthesis, the parameters were used to synthesize five different neutral sentences, such as "I saw your name in the paper." Each sentence was synthesized with six different categories of emotional expression. Listeners were asked to choose whether the speech sounded scared, angry, sad, glad, disgusted, or surprised. In listener studies, the emotion of sadness was correctly recognized 91% of the time. The other emotions were correctly recognized approximately 50% of the time, and mistaken for similar emotions 20% of the time (e.g., disgust was mistaken as anger; scared was mistaken as surprised). The 50% performance was significantly better than the 17% level of chance. Also, the sentences had no explicit context attached, so their content should not have aided the listeners in recognizing the emotion.

Despite the promising results that have been obtained, many research questions remain. For example, the seventeen affect parameters discussed above need more investigation as to how they should co-vary instead of being set independently. Also, their reliability and generalizability are not known beyond the scope of small studies. In particular, the mappings between emotions and vocal features in humans will vary depending on the context. Sometimes an angry person will raise her voice, and sometimes she will lower it. Determining all the possibilities is an open research problem.

As mentioned earlier, people like Stephen Hawking who rely on speech synthesizers could benefit not only from computer voices that can express emotion, but also from computers that could recognize their emotion. Such systems could automatically set the intonation parameters for the synthetic voice. To date, there is no system that takes what a speaking-impaired person is feeling, and has the feelings automatically generate the right settings for their speech synthesizer. Instead, the speaker has to adjust the affect parameters by hand. Nonetheless, the development of affect control knobs for speech synthesis is a step toward this goal.

The Affect Editor can take an input sentence in typed form and synthesize it with a specified affect in acoustic form. But what about the inverse problem, analyzing the affect in a spoken sentence? In Clarke's science fiction novel *2001*, we read that the computer HAL could discern the astronaut Dave's emotions by analyzing his voice harmonics. Will computers be able to do this any time soon? The task is very difficult, especially given that humans are not reliable at recognizing affect in voices. Humans, on average, can recognize affect with about 60% reliability (Scherer, 1981) when tested on neutral speech or on speech where the meaning has been obscured. In the neutral speech studies, people can usually distinguish arousal in the voice (e.g., angry vs. sad) but they frequently confuse valence (e.g., angry vs. enthusiastic). In ordinary conversation, however, a sentence and situation are rarely neutral; the context provides powerful cues to disambiguate the

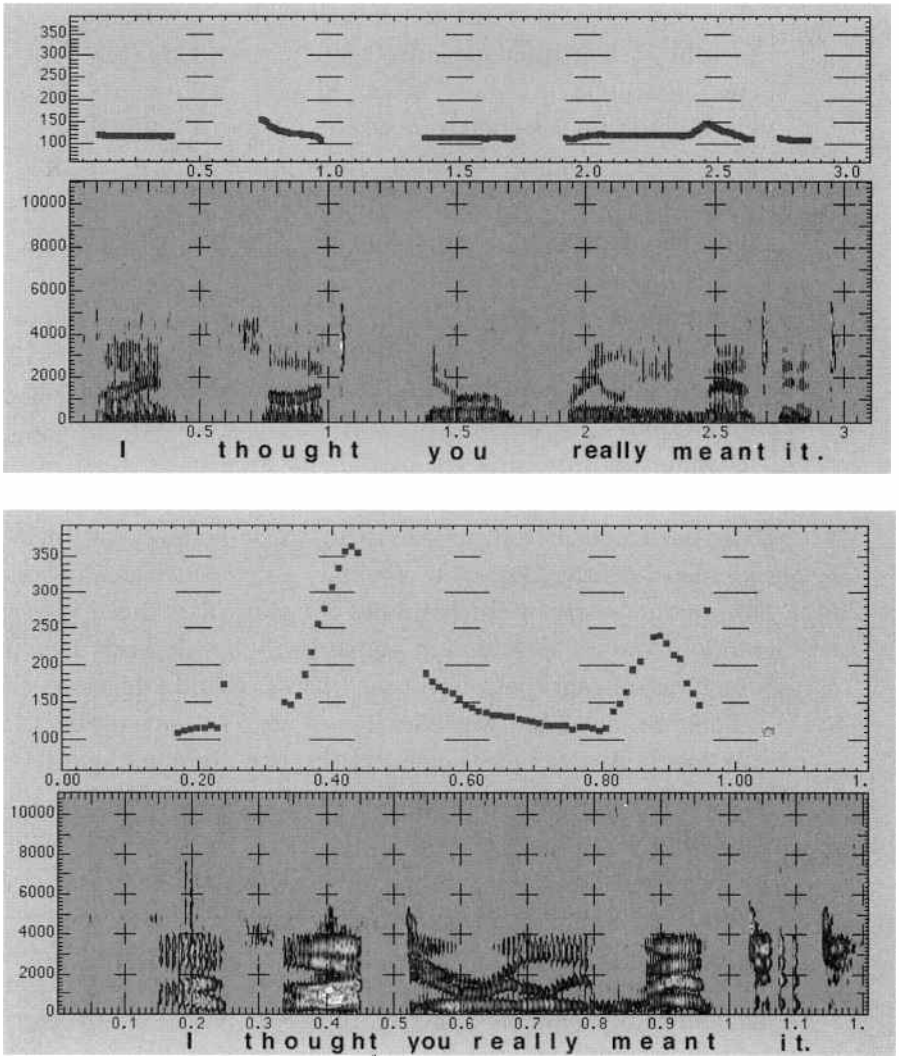


**Figure 6.2**
Voice inflection synthesis. The same sentence, "I thought you really meant it," synthesized for two emotions: sad and annoyed. For each emotion the pitch track (top) and spectrogram (bottom) are shown. Notice the bigger pitch range for annoyed, as opposed to the relatively compressed range for sad. The spectrograms also show differences in speed, pause locations, and enunciation of the two cases. (Spectrograms courtesy of Janet Cahn, MIT Media Lab.)

valence of a spoken message. In other words, the affective cues most readily communicate arousal; the communication of valence is believed to be by more subtle cues, intertwined with the content of the speech.

In efforts to give computers the ability to recognize affect in speech, a variety of features have been proposed. Early studies found that the arousal

dimension of emotion is communicated by pitch and loudness while valence is communicated by subtler and more complex patterns of inflection and rhythm (Davitz, 1964). Some of the earliest research in this area analyzed the voice signals of pilots in stressful situations talking to the control tower (Williams and Stevens, 1969) and actors expressing emotions (Williams and Stevens, 1972), where acoustic features such as the fundamental frequency contour, average speech spectrum, precision of articulation, and other temporal characteristics were used for discriminating certain affective states, especially fear, anger, and sorrow. More recent research has shown a correlation between rising arousal levels, from sorrow to anger or from severe depression to recovery, and a rise in spectral energy in higher frequencies (up to 4kHz); this research also links frequency ranges of long-term voice spectra to the three dimensions of arousal, valence, and control (Pittam, Gallois, and Callanite, 1990). In native Korean actors and French actors speaking neutral sentences with the emotions anger, sorrow, joy, tenderness, and neutral, it was found that arousal was easiest to recognize using the features of pitch range, speech rate, and intensity, and that the duration of the last syllable of a sentence showed promise for valence recognition. This syllable was found to be short in anger and long in joy and tenderness (Chung, 1995). Similarly, a measure of voice quality helped with valence—joyful and tender voices are more resonant than angry or sorrowful voices, which are more aspirated. Linear predictive coding parameters of speech together with speech power and pitch information have also been used in conjunction with a neural net to recognize eight categories—fear, anger, sadness, joy, disgust, surprise, teasing, and neutral—in people interacting with an animated character (Tosa and Nakatsu, 1996).

For training a personal software agent, one of the more useful recognition tasks would be to have the computer recognize whether you liked something or not. However, to date there are no reliable computational measurements of acoustic features of valence. Deb Roy and Alex Pentland, at the MIT Media Lab, have made a preliminary effort to enable computers to classify sentences as approving or disapproving (Roy and Pentland, 1996). This effort used six features—mean and variance of the fundamental frequency, variance and derivative of energy, ratio of amplitude of first to second harmonic, and ratio of first harmonic to third formant—to describe the two classes of approval and disapproval with Gaussian models, and decided which class was present based on Bayesian decision making, a standard method in pattern analysis.[6] The resulting recognition accuracy was 65% - 88% for speaker-dependent, text-independent classification of approving versus disapproving sentences. The same sentences were also judged by people as approving versus disapproving, with similar classification accuracy. The reliability differed from

speaker to speaker; the computational model successfully recognized the approval/disapproval of subject A more easily than B more easily than C, and this pattern of success was duplicated for humans trying to recognize the approval/disapproval of subjects A,B,C. Although this study is very limited, its focus is noteworthy, as indications of approval/disapproval are clearly important to young children, especially pre-verbal infants, and play an important role in learning of right and wrong. If a computer is trying to learn to adapt its behavior to its user, then an ability to sense approval or disapproval from that user would aid in this process.

Studies of affect recognition are complicated by many issues. One complication is how to mask the content of the speech: Play it backwards? Filter it to obscure what is said? Most studies try to get around this problem by choosing sentences with neutral content (e.g., "What time are you leaving?") but there is no guarantee that the content will be received as neutral by the subject. Researchers who work on this should be aware of the pitfalls of various methods for masking sentence content (Scherer, Ladd, and Silverman, 1984). Another potential complication, which apparently none of the studies have considered, is that the mood of the subject assessing the speech may influence the results. As described earlier, studies show that human perception is biased toward positive or negative depending on a subject's mood. In particular, subjects resolve lexical ambiguity in homophones in a mood congruent fashion (Halberstadt, Niedenthal and Kushner 1995), and subjects who look at ambiguous facial expressions judge them as having more rejection/sadness when the subject is depressed, and less invitation/happiness (Bouhuys, Bloem, and Groothuis, 1995). Hence, we can expect that choosing a sentence (or other stimulus) with neutral content and ambiguous affect will tend to be perceived with negative affect by a person in a negative mood, and vice-versa for a person in a positive mood. In other words, the mood of the subjects should be taken into account during recognition experiments.

### *Combinations of Face and Voice*

The above sections gave examples of models for synthesis and recognition of affect both in facial expressions and in voice. The reported results are all preliminary in the sense that they need independent confirmation and would benefit from larger numbers of subjects and expressions, both vocal and facial. Nonetheless, initial results are promising, as all the studies have shown better than random recognition rates and have not revealed any fundamental reasons why affective expression cannot be recognized or synthesized by computers.

A promising area of research is that of combining facial expression and vocal expression to improve recognition results in both domains. The combi-

nation of the two is complimentary, given that arousal is more easily discriminated in speech, and valence is more easily discriminated in facial expressions. Studies on facial expression recognition have mostly been performed only on faces that are not also talking, because the mouth moves differently when someone is simultaneously expressing a facial emotion while speaking. The combination remains a challenge for researchers.

Humans have access to both visual and auditory channels in natural unmediated communication; consequently, it is no surprise that these channels might specialize in different aspects of expression. For example, in the famous McGurk effect, listening to an acoustic "ba" and visually lip-reading a "ga" yields an overall percept of "da" (McGurk and MacDonald, 1976). Neither the visual nor the acoustic signal alone is adequate. The fact that we rely upon both, simultaneously, suggests that it is especially important that face and voice channels be well synchronized in a videoteleconferencing system. When the synchronization is right, then videoteleconferencing is a much richer form of communication than a phone call. Part of the increased value of a ticket close to the stage at a concert or theatre production is the advantage of being able to simultaneously hear the performers and see their facial expressions. In people, the combination of visual and auditory abilities provides richer and more accurate communication; it should also lead to improved performance for computers trying to recognize human affect.

### *Physiological Pattern Recognition*

Patterns of features extracted from physiological signals can be used by a computer to recognize affective information. The idea is to have the computer observe multiple signals gathered while a person is experiencing an emotion, like the ones shown in Fig. 5.7 for grief and for anger, and learn which patterns of physiological signals are most indicative of which affective state. Later, when the system is given only raw signals from a person, then it can use what it has learned previously to try to recognize which affective state most likely gave rise to the signals. Research on this kind of recognition is nascent, but let me illustrate one example of its use with some experiments conducted by Elias Vyzas, working with me at the MIT Media Lab.

In this example, we are given four raw physiological signals—EMG, BVP, GSR, and Respiration—from an actress expressing eight emotions each day. Each emotion was expressed repeatedly over several minutes, with the aid of a sentograph. From each signal, only 100 seconds of the data are used in the experiments below, and this data was taken from the middle of the period of expression. The eight emotions she expressed were: no emotion, anger, hate, grief, platonic love, joy, romantic love, and reverence. These 32 signals were gathered every day for twenty days. Step 1 in analyzing the signals is to

normalize a signal by subtracting its mean and dividing it by its standard deviation, so that every emotion signal on every day has zero mean and unit variance. Step 2 involves computing features of the raw and normalized signals. The decision of which features to compute is mostly an art, since there are an unlimited number of possibilities and much more research is needed to determine which features are best for affect recognition. For this data, we extracted six features: the mean, standard deviation, mean of the absolute value of the first difference, and mean of the absolute value of the second difference, all computed from the raw signals, and the latter two features again, this time computed from the normalized signals. This results in six features for each of four signals per emotion per day. In other words, each emotion on each day is represented by 24 features, or by a point in a 24-dimensional space. Collecting data over 20 days, we obtained 20 such points to characterize each emotion.

It is often useful to look at subsets of the data to try to determine which features give the best discrimination. After trying all possible triplets of emotions and pairs of features, the system finds that the best classification results for this data are obtained when trying to discriminate within the triplet anger, grief, and reverence, or within the triplet anger, joy, and reverence.[7] In both cases, one feature from the EMG signal—the mean of the raw signal—was one of the two best features for classification. However, the best choice for the second feature varied. For the triplet of anger, grief, and reverence the mean of the absolute value of the first difference of the normalized respiration signal gave the best result. For the triplet of anger, joy, and reverence, the same feature but computed from the EMG gave the best result. Figure 6.3 (top) illustrates the 20 points for each of three emotions, where each point is plotted according to the two best features. For the anger, grief, reverence triplet, the recognition accuracy is 72%, and for the anger, joy, reverence triplet the accuracy is 70%. Both are significantly higher than the score of 33%, which would be expected with random guessing.

Using a classic tool of pattern recognition, the Fisher Projection, applied to a subset of the twenty-four original features, we obtained even better results, with 83% classification accuracy for both triplets. The better separation of classes provided by this method can be seen in Fig. 6.3 (bottom). Ideally, the features used to represent each emotion will result in clearly separated clusters for each emotion, although these clusters may need more than two dimensions, and may therefore be much harder to visualize than the examples shown here. In Fig. 6.3, the $\times$, $\circ$, and + signals can be seen to be separated better by the Fisher method (bottom) than by using the two best features (top), although the Fisher method still leaves significant overlap between the reverence and joy classes.

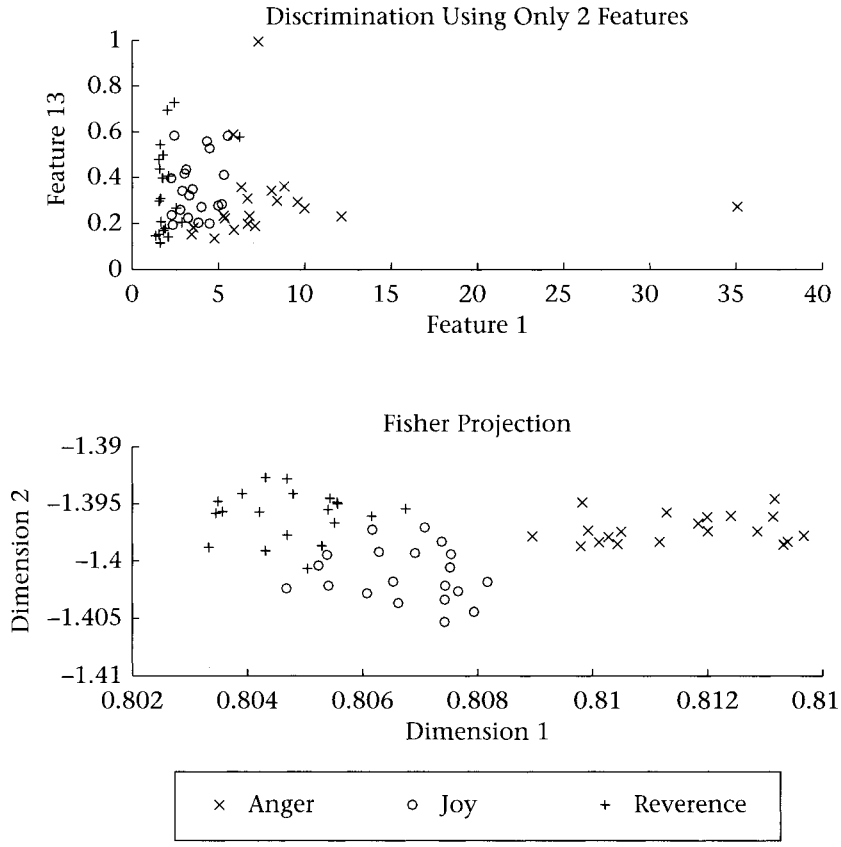


**Figure 6.3**
Each point represents the physiological signals from an actress expressing a state of anger, joy, or reverence. Top: The signals are shown represented by only the two features that were found to best discriminate these three affective states. Bottom: A Fisher projection was used to calculate two dimensions that discriminate these three states.

The six features extracted above were chosen somewhat arbitrarily, to capture variations in the signals that tend to be useful regardless of what the signals represent. In different applications, however, these features may change. Salient features to use for recognizing a person's relative stress and relaxation levels may be different from the six features computed here. In pattern recognition research on images, features representing texture, color, shape, and motion tend to be some of the most useful. A difficult challenge for affective computing research is to determine which features of the physiological signals are most important—to find what is the equivalent of color, texture, shape, and motion in affect.

The four kinds of signals used in this example—EMG, BVP, GSR, and Respiration—communicate different information, and it is an open research

question to determine which combination of these, and other signals, provides the best indicator of affective state changes. For example, various experiments have shown that certain patterns in a person's electroencephalogram (EEG) signals relate to approach vs. withdrawal, which might be used to distinguish affects such as like vs. dislike (Davidson, 1994). However, wearing EEG sensors is not yet as easy as the sensors used in this example.

This example illustrates how features of physiological signals can be combined with pattern recognition tools to provide cues about a person's affective state. In particular, combining information sensed from a user in this way, with both expressive and contextual information from cameras and microphones, provides a rich opportunity for a computer to understand more about its user's affective responses. However, much more research is needed to determine which physiological signals, and which features of these signals, provide the most useful information for the states of interest in the human-computer interaction.

### *Models for Affective Behavior*

The discussion so far has focused on the use of pattern modeling tools for recognizing, classifying, and generating affective patterns, especially facial expressions, vocal intonation, and physiological signals. The models in each case have been used to map patterns and signals to emotion categories, a low-to-medium level transformation. In this section, the emphasis is on mid-level models for representing discrete emotional states. The assumption is that these internal states are "hidden" and that what is not hidden are the observations of sentic modulation, such as a facial expression, which tend to be produced when a person is in these states. Models need to be capable of recognizing that you might express an emotion through a combination of modalities; you might sometimes frown when you are sad, but sadness might also show up in your posture or voice. The model should learn probabilities that given certain observations, a person is in a particular affective state.

Figure 6.4 shows an example of one possible model that meets these requirements, the Hidden Markov Model (HMM). This figure shows only three states, for ease of illustration, but it is straightforward to include more states. For example, a fourth circle could be added for a baseline or neutral state of "no emotion." The premise is that you will be in one state at any instant, and can transition between states with certain probabilities. In the example of the computer tutor, we would expect the probability of the pupil moving from an interest state to a joy state to be higher than the probability of moving from a distress state to a joy state.

The HMM learns probabilities by training on observations, which could be any measurements of sentic modulation varying with the underlying states,

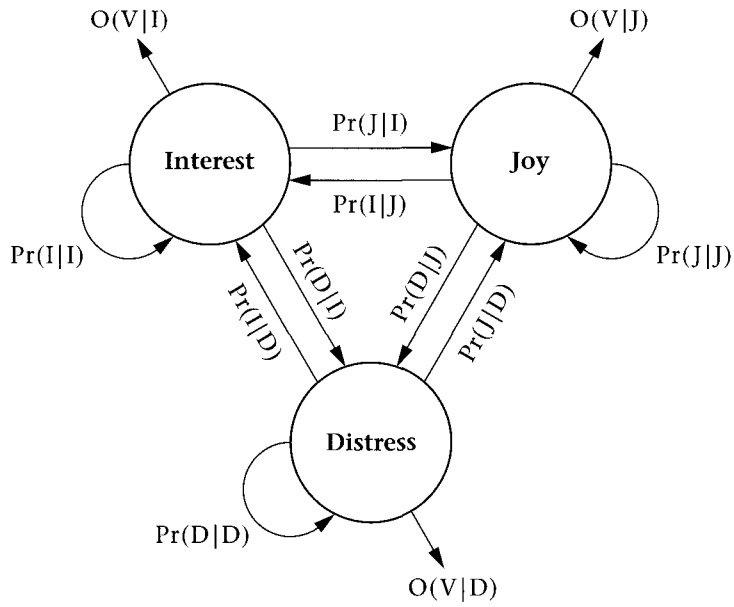


**Figure 6.4**
The Hidden Markov Model shown here characterizes probabilities of transitions among three "hidden" states: interest (I), distress (D), and joy (J). It also characterizes the likelihood of certain observations given these states, such as how features of voice inflection, V, will change with each state. The affective state of a person cannot be observed directly; only observations that depend on a state can be made. Given a series of observations over time, the computer tries to determine which sequence of states best explains the observations.

such as changes in voice inflection, facial expression, or autonomic changes such as heart rate. The input at any time is these observations; the output can be either the state that the person is most likely in, or it can be identification of an entire HMM configuration, thereby recognizing a larger pattern of emotional behavior. In the latter case, there would need to be a family of HMM configurations, one corresponding to each emotional behavior, or each person's characteristics for a given behavior. For example, the computer tutor might recognize different patterns for different pupils, which might help it to tailor its feedback more effectively.[8]

HMM's work in multiple contexts. Different HMM's can be trained as functions of environmental, cultural or social context. Your sentic modulation patterns may differ if you're driving a car in the country on a Sunday versus in the city at rush hour, going out with an old friend versus meeting a blind date. The probabilities of certain expressions vary given different conditioning events. For example, the probability of showing facial expressions at the office is smaller than the probability of showing them at home. Context can

also include temporal events. Different HMM's may be learned as a function of timing relative to a hormone cycle or to exam season. Hence, the probabilities, states, and structure of the model vary depending on a variety of factors, ultimately determined by the intended use of the model.

In any of these cases, the HMM states can correspond to pure emotional states as illustrated in Fig. 6.4, or they can correspond to more fundamental building blocks, perhaps identified by the computer as it works to fit the data. The states do not have to have recognizable names of emotions; they might instead correspond to regions of a dimensioned space where the person's sentic modulation measurements cluster. For example, one HMM state might be made by noticing clusters of physiological variables that occur in particular situations, and assigning each cluster to its own state. Alternatively, a complex pattern of clusters might be represented by its cluster-based probability model (Popat and Picard, 1993). In either case, the model is customized to an individual, and can learn to represent unnamed feelings that happen reliably in certain situations. Furthermore, the model can capture the dynamic aspects of an emotion—associating a whole HMM to one emotion. The model is free to adapt to new theories of emotional building blocks—whether at the granularity of the basic emotions of anger, sadness, etc., or at a smaller granularity from which dynamic emotions may be constructed.

HMM's are also suitable for representing emotion mixtures, following either the bathtub or microwave metaphors used earlier. In the case of the former, a state can be established as a mixed emotion; it can be constructed out of several simultaneous components, as melancholy might be constructed out of the components of love and sadness. In the case of the latter, pure states can be alternately visited in rapid succession in time. An HMM for a "love-hate" relationship would cycle between two or more states of love and hate, perhaps occasionally pausing in a neutral state.

A model such as the HMM can be used not only to recognize certain affective patterns, but also to predict what state a person is most likely to be in next, given the state they are in now. The prediction process is one of partial recognition: First, fit the model to both previous and present observations. Second, use these results to synthesize the most probable state to occur next. The synthesized state acts as the prediction. Like a human observer, such a model-based prediction can give a likely outcome, but can never say with 100% certainty what will happen. When these models synthesize or predict they do so only in a probabilistic way, not taking into consideration high-level reasoning or logic. Consequently, they are not as well suited to predicting emotions based on cognitive appraisals as some of the models I will describe in the next chapter. Nevertheless, they are well-suited to describing patterns of affective state transitions, and inferring hidden states given these patterns.

### *Additional Models and Learning*

Numerous other models may prove to be useful in modeling affective information. An artificial neural net is one general purpose tool which has already been applied to emotion expression recognition, and which will be applied to emotion's influence on memory and performance in the next chapter. As an aside, it is interesting to note that the most popular method used for training artificial neural nets, backpropagation, was originally inspired by the idea of emotional energy being attached to associations. Paul Werbos writes that he came up with the idea of backpropagation while trying to mathematically translate an idea from Freud, who proposed that human behavior is governed by emotions, and that people attach cathexis (emotional energy) to things Freud called "objects." Quoting from Werbos (1994):

> According to his [Freud's] theory, people first of all learn cause-and-effect associations; for example, they may learn that "object" A is associated with "object" B at a later time. And his theory was that there is a *backwards* flow of emotional energy. If A causes B, and B has emotional energy, then some of this energy flows back to A. If A causes B to an extent W, then the backwards flow of emotional energy from B back to A will be proportional to the forwards rate. That really is backpropagation. . . . If A causes B, then you have to find a way to credit A for B, directly. . . . If you want to build a powerful system, you need a backwards flow.

The use of some form of backwards flow is a significant part of most computer learning methods today. It can be implemented without having to give the computer an emotional system. Nonetheless, the mechanism is apparently similar to the role that emotions play in human learning.

There are a host of other possible models that can be employed for analyzing and synthesizing emotional expressions. Camras (1992) has proposed that dynamical systems theory be considered for explaining some of the variable physiological responses observed during basic emotions, but has not suggested any models. Emotion system dynamics might be captured by nonlinear models such as the M-Lattice (Sherstinsky and Picard, 1994) , a model that generalizes certain kinds of neural nets. Grossberg and Gutowski (1987) have proposed that emotional processing can be accomplished with an opponent processing neural network called a gated dipole. Freeman has modeled olfaction with dynamical systems and argues the relevance of this approach for modeling limbic influences on intention and motivation in his book *Societies of Brains* (Freeman, 1995). There are, no doubt, many more possibilities; the field of research is wide open for exploring which models are best suited to capturing the most useful features of emotions.

Note that no one model—discrete, continuous, implicit, emergent, linear, nonlinear, or otherwise—is likely to perfectly recognize an underlying

emotional state. For example, tears may be recognized from a video image of a face, but they don't necessarily correspond to sadness—they could be tears of happiness. The most successful recognition can be expected to occur when a computer learns a personalized combination of low-level perceptual cues, such as pattern recognition of visual, vocal, and other biosignals, and high-level cognitive cues, such as reasoning that the viewed event satisfied a long-term goal of the user, and might make her extremely happy. Additionally, these cues will work best when considering the context; for example, is it a poker game where bluffing is the norm, or a marriage proposal where sincerity is expected? The important influence of reasoning, especially cognitive appraisal of a situation, and the synthesis of so-called "cognitive emotions," is the subject of the next chapter.

### *Summary*

This chapter has described models that can be used to start giving computers the abilities necessary to recognize and express emotions. In particular, tools from pattern recognition and analysis have been suggested for recognizing and synthesizing facial expressions, recognizing and synthesizing vocal inflection, recognizing physiological patterns corresponding to affective states, and modeling emotional behavior. Research in this area is very new, but results on small sets of emotions and small sets of people already indicates that computers can achieve useful performance in recognizing and expressing affect.