Review

Beyond emotion archetypes: Databases for emotion modelling using neural networks

Roddy Cowie*, Ellen Douglas-Cowie, Cate Cox

School of Psychology, Queen’s University Belfast

Received 20 March 2005; accepted 23 March 2005

Abstract

There has been rapid development in conceptions of the kind of database that is needed for emotion research. Familiar archetypes are still influential, but the state of the art has moved beyond them. There is concern to capture emotion as it occurs in action and interaction ('pervasive emotion') as well as in short episodes dominated by emotion, and therefore in a range of contexts, which shape the way it is expressed. Context links to modality-different contexts favour different modalities. The strategy of using acted data is not suited to those aims, and has been supplemented by work on both fully natural emotion and emotion induced by various technique that allow more controlled records. Applications for that kind of work go far beyond the 'trouble shooting' that has been the focus for application: ‘really natural language processing’ is a key goal. The descriptions included in such a database ideally cover quality, emotional content, emotion-related signals and signs, and context. Several schemes are emerging as candidates for describing pervasive emotion. The major contemporary databases are listed, emphasising those which are naturalistic or induced, multimodal, and influential.

q 2005 Published by Elsevier Ltd.

Keywords: Database; Emotion; Multimodal; Elicitation; Naturalistic

Contents

1. Emotion and archetypes .................................................................000
2. The scope of databases ...............................................................000
  2.1. Types of emotionality ............................................................000
  2.2. Contexts and modalities of emotion ........................................000
  2.3. Obtaining samples of emotion ..............................................000
  2.4. Emotion and applications .....................................................000
3. Descriptive issues in databases ...................................................000
  3.1. Describing quality ...............................................................000
  3.2. Describing emotional content ..............................................000
  3.3. Analyses of emotion-related signals and signs .......................000
  3.4. Describing context ..............................................................000
4. Databases ..................................................................................000
5. Conclusion ...............................................................................000
6. Uncited references .....................................................................000
7. Acknowledgements ...................................................................000
8. References ..............................................................................000

It is a truism that a neural network can only achieve what its training data allows it to. In the area of research on emotion, the quality and quantity of data available is a major issue. This paper aims to inform the research community about the collections of data that are available, and to inform it about ongoing developments in the area.
One of the key developments is growing recognition that the task of assembling adequate databases is a major challenge, intellectual as well as practical, with links to fundamental issues in the theory of emotion. In effect, a database expresses a tacit theory of the domain it deals with, and learning extracts it (selectively). To choose a database, and still more to construct one, research teams need some sense of the theoretical issues in the area. That includes knowing when ideas that intuitively seem appealing—or even compelling—are old science or ‘folk psychology’ (Place, 1996) rather than credible contemporary science. The paper aims to draw out the scientific issues that are relevant to decisions about databases as well as describing the actual material available.

Reviewing the issues leads to a broad sense of the ground that database research might ideally aim to cover. The best convenient summary is ‘emotion in action and interaction’—indicating that the challenge is not to find the emotional equivalent of pure elements, but to represent emotion as it appears woven through everyday activities.

From that we move to list the significant data sources reported in the literature. For each source, we give the information most likely to be relevant to research teams aiming to identify useful sources or to evaluate work that has been done with them.

1. Emotion and archetypes

‘Emotion’ and related words are notoriously elusive (Russell & Barrett-Feldman, 1999). One of the key issues relates to a well known characteristic of words in general. When they are used in isolation, they tend to call up images that people regard as archetypal examples of the category in question (Mervis, Catlin, & Rosch, 1976). That is a useful device in its place, but it needs to be treated with proper caution. Certainly research should not slip into assuming that the empirical data needed to understand a topic equate to a collection of cases that reflect the archetypal images we associate with it.

In the domain of emotion, there are several kinds of pressure to make exactly that assumption. Understanding recent work on databases depends on recognising that it has been concerned to break their grip.

The first type of pressure to focus on archetypes comes from well-known theoretical concepts. The concept of primary emotions is particularly influential. The underlying idea, which goes back at least to Descartes (1911–1912), is that emotion is like colour. Certain emotions, like certain colours, are ‘pure’ examples of the phenomenon: others are ‘secondary’, made by combining them in different proportions. If that theory were correct, then the agenda for emotion research would be clear: investigate the primaries, and the secondary forms will fall into place.

The primary/secondary distinction has enduring appeal, but Scherer et al. (2004, p. 10) sum up the consensus within the emotion research community: ‘This distinction is especially problematic ... and should therefore be used only with caution.’ The same warning applies to the idea that the ideal basis for training is a database stocked with examples of ‘primary emotions’.

More recent is the concept of ‘basic emotions’. It expresses the idea that emotion consists of qualitatively distinct types of state, each related to a biologically important scenario; and (in the best known version of the idea) each with a direct biological link to particular expressive behaviours, vocal and more particularly facial. A famous list due to Ekman’s group (Ekman et al., 1987), is sometimes called the ‘big six’ (Cornelius, 1966) However, even Ekman no longer advocates that kind of attractively short list (Ekman, 1999). Longer alternatives are summarised by Cowie and Cornelius (2003). Correspondingly, databases composed of a few hypothetically basic states need to be treated with caution.

A second pressure to focus on archetypes is precisely that many databases do deal with a few archetypal states. The outstanding example is Ekman and Friesen’s (1975a,b) collection of still photographs, which is grounded in an early version of his ‘basic emotion’ theory. Various other collections, still or video, have been modelled more or less explicitly on it (e.g. http://www.media.mit.edu/).

These databases are intuitively appealing, but, as noted above, the theory behind them is dated. And even in its strongest form, Ekman’s account included qualifications that set limits to the likely uses of his database and others like it. It specified that in everyday life, the expressive gestures rooted in biology were overlayd, modified, and mimicked according to socially defined ‘display rules’ (Ekman & Friesen, 1969). Hence, the ‘pure’ examples of Ekman’s database were explicitly not the patterns that we should expect to see when humans actually interact, with each other or with machines; and training systems to deal with those interactions would need data that showed the effects of display rules.

A third type of pressure favours a different kind of archetype. There is a tradition of believing that there is a peculiar direct correspondence between emotion terms and physiological measurements. One version of the idea was introduced by William James (1884), who suggested that the basis of emotion was a set of bodily responses to a significant type of stimulus—change in heart rate, breathing, skin state, and so on. More recently there has been intense interest in relationships between emotion and brain states, such as activation of the amygdala (LeDoux, 1987) and change in the balance of electrical activity between the two cerebral hemispheres (Davidson, 1999). It is a natural step from those ideas to infer that some set of physical measurements represents the ‘ground truth’ behind the confusing surface of emotional behaviour, and it is the business of a database to ensure that those measures are available.

The Jamesian idea seems to contain some truth (Cornelius, 1966), but it is far from the whole truth. There is a large literature on the extent to which emotions can be
discriminated by bodily signs (e.g. Cacioppo, Klein, Berntson, & Hatfield, 1993). Some discriminations certainly can be made, but they are quite coarse, and they almost always depend on very carefully controlled experimental conditions. The reason is that most of the variables which are affected by emotion are also affected, as much or more, by commonplace activities such as moving, speaking, or even thinking. As a result, their value as discriminators diminishes rapidly as one moves towards free situations where people are likely to be moving, speaking, or thinking.

On one level, the jury is still out on the measurement of brain states. The reason is that many of the key measurement techniques require activity to be severely restricted, making it impossible to study whether normal activity would interfere with the detection of emotion or change the patterns associated with it. However, there are good reasons to doubt that reductionist approaches based on brain states will fare better than any others.

Saying that a person is in a particular emotional state carries implications about the way he or she attends to and perceives things, people, and events, which may be present, past, anticipated, or imagined; about his or her feelings; about his or her capacity to weigh options dispassionately; about the actions he or she is disposed or likely to take; about the moral status of his or her thoughts and actions; and at least in some cases about the reality of his or her situation (e.g. Sabini & Silver, 2005; Cowie, 2005).

Such a statement strongly suggests that ascriptions of emotionality implicate many systems, at least some of them rather a high order, and not all of them contained within the person to whom the emotion is attributed; and that they refer to subtle features of the ways in which systems operate rather than simply whether they are active or not. It is instructive that systems such as attention, which are part of the emotional pattern, are themselves based on highly complex and distributed neurophysiological substructures (Schneider, 1995). In such a situation, research cannot realistically expect brain imaging to identify tractable correlates of emotion terms as they are used in natural language. It may find correlates of something simpler that plays a role in emotionality, but that is another matter.

A common feature of the ideas covered in this section is that they suggest a two-stage approach to understanding emotion. In the first stage, a set of relatively simple core correspondences is established. It is assumed—perhaps implicitly—that analysing the core correspondences will pave the way to deal with the complexities of emotion as it appears in everyday action and interaction, and the attribution of everyday emotion words. It is a short step to the assumption that the business of databases is to supply data that allow these core correspondences to be recovered.

It would be misleading to say that two-stage approaches are discredited. Rather, it is no longer obvious that attempting to bypass the apparent complexity of the emotion domain is the best way to proceed. Emotion-related phenomena have refused to fall into place around a succession of plausible simplifying assumptions. It may be that current assumptions will pay off with additional research, or that assumptions yet to be articulated will pay off. But the alternative has to be taken seriously—that the phenomena are essentially complex, and will resist simplification indefinitely.

The idea that the phenomena are essentially complex suggests two conclusions. First, it is the responsibility of research on databases to supply data that fully reflect the complexity of the phenomena. Second, techniques like neural nets, which are inherently suited to addressing complex relationships, are likely to have a key place in describing the patterns that exist. The two define complementary parts of an emerging strategy.

2. The scope of databases

It is not self-evident what range of material emotion-related databases should contain, particularly if the decision is not pre-empted by archetypes. This section outlines major themes in recent thinking about the subject.

2.1. Types of emotionality

At least two broad types of phenomenon are described as emotional in common parlance. We have described them as ‘episodic’ and ‘pervasive’ emotion (Cowie & Schroeder, 2005).

The term ‘episodic emotions’ describes states where emotion dominates people’s awareness, usually for a relatively short time, by affecting their perceptions, feelings, and dispositions to act. The emotion may not determine the action that the person takes, but it requires effort to prevent it from doing so. Clear-cut episodic emotional states are generally brief, and relatively rare.

In contrast, ‘pervasive emotion’ refers to something that is an integral part of most mental states, including states where rationality seems subjectively to be in control. It is integral in the sense that emotional colouring is part and parcel of the way people experience situations that they are in, or remembering, or anticipating. It may be associated with the situation as a whole, or within individual elements of it, or with possible courses of action. However, the colouring is part of the background unless something triggers a kind of phase shift and propels people into a state of episodic emotion.

It is a major issue whether the business of databases is to assemble instances of episodic emotion, or to take a broader remit and represent pervasive emotion. Database users have the right to look for sources that consist of emotion in the strong sense formulated by Scherer. But if they do, they should be aware of the reasons why recent collections have moved away from that pattern. These are bound up with prevalence and application.

The key question about prevalence is whether episodic emotion is intermittent, but relatively common, or distinctly
that clear-cut bursts of episodic emotion will look and sound somewhat similar in most contexts—whatever the person was doing will be interrupted and replaced by emotion-specific behaviour. Thinking of pervasive emotion raises different issues. It will be integrated into the way ongoing actions and interactions are executed. The results will depend on the nature of those ongoing actions and interactions, and will be very different in different contexts—watching a football match, driving a car, buying groceries, standing waiting for a friend, chairing a meeting, writing a paper.

Database research has not generally paid systematic attention to the questions about context, but the issue is gaining recognition. In particular, the ongoing HUMAIN project has set out a programme involving two types of context-sensitivity—labelling for context in naturalistic data, and carrying out induction studies chosen to represent a range of strategically different contexts.

Context is inescapably linked to modality. Emotion is strongly multimodal in the sense that signs may appear in various different channels. However, not all types of sign tend to be available together, because context will affect the signs that are relevant or accessible. One of the most challenging tasks for database work is to capture records that clarify when and how which modalities come into play.

Facial signs are potentially available in a wide range of contexts, but they are also subject to a wide range of influences. These include culture-specific display rules and interactions with speech, involving both the lips and the eyebrows, They are reduced in the absence of social context, and can be feigned or dramatised, consciously or unconsciously. Databases provide systematic coverage of a small part of this range.

Gesture-related signs are not generally available when the hands are occupied by a task, and they are constrained when the arms are constricted (e.g. by a narrow space) or being used (e.g. to lean on or to lift something). They are also socially constrained (as every good child knows, it is rude to point), and the violation of social constraints may be a sign of emotion in itself. Databases showing emotion-related gestures reflect a very small part of that situation.

Similar comments apply to global posture (‘body language’). The signs that can occur in mobile, standing, seated, and reclining contexts are very different, as are postures shaped by engagement with an object and/or other people. The emotional significance of a posture may also depend on its orientation to another person or congruence with his or her posture.

The classical vocal signs of emotion involve prosody, voice quality, and timing. Their context is subtle. They are clearly linked to interactive contexts, but there seems to be antagonism between overt emotionality and well-formed speech communication, as reflected in phrases like ‘speech-speechless with emotion’ (Cowie & Cornelius, 2003). That suggests the natural setting for vocal signs of emotion is in a transition zone between controlled speech and inarticulate
vocalisation. Intrusion of non-speech sounds (Scherer, 2003) makes sense as part of a transitional state like that.

Capturing that kind of setting is an intriguing challenge in itself. Like the face, the voice can easily be dramatised. That presumably involves mimicking some of the signs that occur in the transition zone. In any case, dramatised speech is an integral part of the way people communicate their emotional states, and well worth representing in a database.

Traditional research on speech and emotion focused on vocal signs and deliberately excluded verbal variation by requiring subjects to express various emotions using the same word of phrase. In fact, though, verbal content naturally covaries with vocal signs, in several respects. Acted databases almost never capture that aspect of emotionality, and relatively little has been done to draw it out of the naturalistic databases that do reflect it.

Some attention has been paid to cues at the level of discourse (broadly speaking). Lee and Narayan (2003) classified utterances into rejections, repeats, rephrases, ask-start overs, and others. The discourse measures were less informative than acoustic measures, but nevertheless they did relate systematically to emotional state.

Beyond that, emotion affects the classical linguistic variables of syntax and vocabulary. A simple demonstration comes from the paper by Athanaselis et al. in this special issue, which shows that the verbal content of emotional speech can be recognised much better if the language model is based on a corpus biased towards utterances with some signs of emotional content. A large literature dating back several decades provides a rich source of ideas about the way emotion might influence choice of vocabulary—immediacy of expression, concreteness of terms, use of expletives, and so on (Berger & Bradac, 1982).

Intuitively, it seems obvious that emotion will usually be intimately linked to the semantic content of people’s speech, and that the most powerful cues to emotionality will usually come from content. Rich databases are the key to pursuing that idea systematically, and they have only begun to reach the point where that is possible.

Effects involving attention are similarly under-represented in mainstream databases, with the partial exception of extreme inattention in boredom (Cowie, McGuigan, McMahon, & Douglas-Cowie, 2003) and drowsiness (Hadfield & Marks, 2000). That is curious in the light of contemporary theory, which understands emotion in terms of appraisal processes concerned with forming selective, valenced representations of the environment (Scherer, 1999). Eye movement, head movement and posture all offer potential sources of information about the focus and manner of a person’s attention, and hence of the appraisals governing his or her emotional state. Again, rich databases are the key to pursuing the idea systematically.

Action ranks with attention as a variable that theory suggests should be pivotal. Aristotle understood emotion in terms of action tendencies (such as the impulse to revenge in anger), and the idea has remained an integral part of modern theory since Frijda (1986) re-established it. Very few databases are designed to record action patterns that might be symptomatic of emotion. An exception is work on driving (McMahon, Cowie, Kasperidis, Taylor, & Kollia, 2003), which offers a context for measuring behaviours associated with risk-taking and aggression.

Sensitivity to context is particularly crucial in the case of physiological measurements, peripheral and central. The optimal setting for discriminating emotional states on the basis of physiological measurements is a sedentary individual who is inducing pure and intense emotional states by acts of imagination (Picard, Vyzas, & Healer, 2001) or being exposed passively to powerful emotion-inducing images. Such experiments clarify points of theory by showing that connections exist between emotion and certain somatic and neural systems. It is a different issue to establish how they feature and may be used in the more complex contexts that are relevant to most applications.

There are contexts where it is reasonable to think that physiological measurement could be useful. Driving is a prime example, because simple physical effort is limited, and relevant environmental variables (temperature, humidity, etc.) can be monitored. There is some data on physiological correlates of emotion-related states in that context, and more work is under way (McMahon et al., 2003).

Forming an overview of contexts, and the ways in which emotion might be expressed within them, is a first step towards developing an analysis of pervasive emotion. The next step is to assemble records that establish what actually does happen in contexts that have been identified as significant. That is considered in the next section.

2.3. Obtaining samples of emotion

Most technological research on emotion continues to be based on recordings of actors, skilled or unskilled. The fundamental reason is linked to the previous sections. It is assumed that the proper target is episodic emotion; episodic emotion is much rarer than people tend to assume, and difficult to elicit; therefore, research teams turn to actors.

Blanket rejection of acted data is not a sensible option, for various reasons. However, there are good reasons to move away from uncritical reliance on acted data, and that has been a major theme in recent database work.

Several kinds of standard can be used to evaluate acted data. The one which is most often observed is that people can recognise the emotion being portrayed (Banse & Scherer, 1996; Ekman & Friesen, 1975a,b; Laukka, 2005). It is not a criterion that should be over-emphasised. On one hand, it would tend to rule out a good deal of genuine emotional material, because inter-observer agreement on it is often not particularly high. On the other, it does not rule
out material which is identifiable, but not at all like natural emotion—in short, ‘ham’ acting.

We have explored a technique which addresses the second problem. As a recording is played (audio, visual or audio–visual), observers use a sliding scale to indicate whether they think the material is spontaneously expressed emotion or acting, and how confident they are in their judgment. We compared ten extracts from the Belfast naturalistic database with renditions of the same material by actors in the Belfast structured database (Douglas-Cowie et al., 2003). On the basis of pilot studies, we chose the most convincing rendition of each extract, and one which was about average. Two measures were of interest. The first was the percentage of trials which eventually favoured the right choice (acted or not). 68% concluded that the less convincing renditions were acted, and 67% that more convincing renditions were. In contrast 79% concluded that the ones which were actually spontaneous were not acted.

The second measure was how quickly the difference between responses to the corresponding real and acted extracts reached statistical significance. For both convincing and unconvincing renditions, the difference reached significance within the first three seconds. Both measures indicate that even well acted material differs very noticeably from spontaneous emotionality.

To the best of our knowledge, no databases of acted material currently include measures of naturalness. However, it is easy to carry out the kind of study we have described above, and it would make sense to carry out checks of that sort before committing to using an acted database.

Even if acted material is far from natural, it could conceivably be well suited for training, if it presented the features that are relevant to everyday behaviour in sharper focus. There seem to be no direct tests of the idea, but the evidence that exists is not encouraging. Batliner, Fischer, Huber, Spiker, & Nöth (2003) have shown that the vocal signs of emotionality used by an actor simulating a particular human–machine interaction were different from, and simpler than, the signs given by people genuinely engaged in it. Kollias et al. (2005) have used the Ekman and Medialab databases of acted facial expressions as a basis for generating less extreme facial profiles, and are applying the approach to recognition in induced emotion data. Results to date suggest that the approach has some success with individuals who express themselves in a dramatic way, but not with individuals behaving more naturally.

Note that for some applications, acted material is actually the optimal source. For example, an artificial newsreader should probably simulate the conventionalised emotional signs offered by human newscasters rather than showing true grief at disasters or euphoria at triumphs.

Naturalistic data seems an ideal alternative to acted sources, but the reality is not straightforward. Problems of copyright and privacy are well recognised (Douglas-Cowie et al., 2003). More fundamental, but less often noted, is the fact that key tools needed to deal with naturalistic data are underdeveloped.

First, recordings of the kind seen in ‘reality’ TV could be obtained in reasonable quantity, but technology lacks the robust low-level analysis needed to handle them. Humans can recognise expressions of emotion from faces that are at an angle or at a distance, but the 3D modelling techniques needed to do that automatically have barely been explored (Kollias et al., 2004). Similarly, speech analysis lacks the tools to deal with varying distances from microphones, reverberation, and noise. The traditional strategy is to expect recordings to cater to the engineering limitations, but the usual methods of doing that conflict with free emotional expression, and the long term solution is clearly to address the engineering limitations.

Second, we lack the conceptual tools to deal with the diversity of natural emotional behaviours. The traditional strategy is to look for elements common to all or most instances of an emotion (Juslin & Laukka, 2002). From the perspective of the previous sections, that filters out a huge amount of evidence, because so much evidence lies in the way emotion modulates activities that carry their own constraints. If the expression of emotion has common elements in the contexts considered above, they are at a deeper level, involving attention, planning, risk-taking, and so on. We do not have the theory at that level that would allow us to draw out the common threads in a genuinely naturalistic collection. Sophisticated learning techniques may well be relevant to developing it.

Recordings from real human–machine interactions avoid both of these problems to some extent, because they involve built-in constraints and a kind of uniformity. If short-term applications are the target, then directly relevant databases are clearly the appropriate resource—not least because of the context-dependence that has been emphasised. However, focussing on short term applications may be counterproductive in the long term, because it postpones confronting the major issues, and finding relationships which are dependent on the structure of a particular task may not clarify the general picture. To illustrate, we studied bored subjects carrying out a naming task (Cowie et al., 2003), and found reduced pausing (because speakers ran descriptions of successive items together). However, we have since found that a small change in the structure of the task results in exactly the opposite trend, increased pausing (Cowie & Schroeder, 2005). It seems that the underlying effect is not on pausing, but on attention; and the overt signs of the attentional effect depends on task details. If the field were driven by data from many very specific tasks, it would risk being overwhelmed by that kind of highly context-specific contingency. Allied to that is the risk demonstrated by Yacoub, Simske, Lin, & Burns (2003). A recogniser trained on a data containing a narrow range of emotions is likely to respond quite inappropriately if it encounters an emotion outside the training range. That raises ethical as well as practical problems (Goldie et al., 2004).
In spite of the problems, major efforts have been invested in collecting naturalistic data (Batliner, Hacker, Steidl, Noth, D’Arcy, Russell, & Wong, 2004; Batliner, Hacker, Steidl, Noth, & Haas, 2004; Douglas-Cowie et al., 2003 De Villiers et al. this issue). Key sources are listed in the Appendix to this paper.

Between acting and pure naturalism lie various emotion induction techniques. It is not difficult to create situations which induce a specific emotion such as irritation (Bachorowski, 1999) or boredom (Cowie et al., 2003). It is more challenging to induce a range of emotions in comparable conditions, but there are various established methods—verbal methods such as the Velten technique (Velten, 1968); listening to emotive music (Clark, 1983; Kenealy, 1988); looking at emotive pictures (Center for the Study of Emotion and Attention, 1999; Lang, Bradley, & Cuthbert, 1999) or films (Gross & Levenson, 1995); and playing specially designed games (Scherer, Johnstone, & Bänziger, 1998). Cowie and Cornelius (2003) give a more detailed review of classical techniques and their status.

The traditional test of induction techniques is whether they evoke a state that genuinely qualifies as emotion. That is a reasonable criterion if one is looking for an inner core of phenomena that (hypothetically) accompany all emotions. However, the perspective developed here points to a different test, which is ability to induce emotion that carries through into action and interaction. Work on that problem is much more recent.

Scherer et al. (1998) carried out pioneering work with a computer game containing incidents designed to induce specific appraisals. Their effect was shown not only in transient facial responses, but also in a reading task which was part of the game. That taps strictly vocal signs of emotion in speech, though not effects at the level of phrasing, word selection, etc.

A technique developed by Auberge, Audibert, and Rilliard (2003) evokes freer expression. Students with an interest in language used a voice-driven language teaching program, and were given feedback indicating that their performance was outstandingly good or disturbingly bad. Their ongoing reactions were reflected in facial expression and vocal behaviour which included specific items to be pronounced and unscripted comments.

Our SAL technique (Cox 2004) is modelled on ‘chatbot’ systems. The user talks to characters whose conversation is defined by ‘scripts’ consisting of emotionally loaded stock phrases. The system contains different characters, each of which has a distinctive emotional outlook (in the current version, angry, happy, sad, and pragmatic), and each tries to draw the user towards its own outlook. Users do not automatically engage with SAL, but when they do, it allows them to express a range of emotions through a relatively wide range of channels, facial, vocal, discourse, and content.

One of the key difficulties for database research is that induction remains an uncertain art, seriously complicated by ethical issues. It is a telling sign that we are currently exploring ways to induce a range of genuinely emotional states that are persistent enough to affect driving behaviour (in a simulator). Transient emotion can certainly be induced by events of no enduring significance in subjects’ lives, but it seems to dissipate very quickly when they are presented with a task or a challenge. The same does not hold when the inducing events or situations are part of the fabric of people’s lives, but not many laboratories would manipulate those events to gather data.

2.4. Emotion and applications

Application is clearly an issue in the development of databases, but it is best addressed after the scope of the field has been indicated.

The obvious short-term application of neural nets in the domain of emotion is ‘trouble shooting’—detecting emotional states that may be troublesome in callers using automatic exchanges or websites, pilots, drivers, and so on (Cowie & Schroeder, 2005). Another possibility is ‘affective selection’—using emotion-related signs as evidence that a person might want to hear certain music, preserve certain experiences, or be shown certain types of option (Cowie & Schroeder, 2005). Neither the databases nor the computing techniques needed for these applications is particularly exciting, at least at first sight.

The longer term goal, though, is ‘really natural language processing’—communication between humans and machines that follows human norms in emotional terms as well as others. That means both registering and providing the kind of emotional colouring that is the norm in human interaction. In applications such as artificial companions and tutors, there is a clear incentive to do these things better than many humans do. It is to achieve targets like these that research needs databases of great scope and subtlety.

Correspondingly sophisticated computing is needed to exploit the data. The requirement in these contexts is not simply to label isolated patterns. It seems likely that sequence and timing, often crossing different modalities, are critical in the detection of emotion-related effects. Taylor’s ANNA (this issue) illustrates the reasonable supposition that mechanisms modelled on the human brain will prove the best way to handle that kind of problem. The same point applies to the inverse problem, which is to generate appropriately sequenced and timed arrays of signals.

3. Descriptive issues in databases

Emotion databases include not only recordings in various modalities, but also descriptions of the emotions involved and the signs relevant to them. This section summarises the main options in these areas.
3.1. Describing quality

Several preliminary issues should clearly be documented in a satisfying database, though they rarely are. The key examples are

- Is any emotion evident? If, for instance, one is exploring a large database that involves very little emotion, one of the most useful elements a database could have is a trace showing where there is emotionality of any sort is among.
- Is emotion being masked? This is linked to the Ekman concept of display rules (see Section 1).
- Is the emotionality that is displayed spontaneous or contrived? This is not a simple matter of ‘good’ and ‘bad’ samples. For instance, an element of dramatisation is part of everyday communication, and sometimes the target is to achieve acceptable formal representations of emotion (see Section 2.3).

3.2. Describing emotional content

Three main types of scheme could reasonably be used to describe the emotional content of a database: categorical; continuous; and appraisal-based.

Categorical schemes are rooted in everyday language. They involve assigning terms like angry, ashamed, jealous, and so on. There is no real doubt that terms like these have an enduring place in emotion-oriented technology, because they are the means of describing emotion that people find most natural. The difficulty is how to deal with the range of descriptors that everyday language offers and the kind of emotion that appears in relatively natural settings.

Numerous teams have tried to label relatively naturalistic material using schemes based on established psychological lists of emotions. Two examples illustrate the outcomes.

The Leeds-Reading database (Douglas-Cowie et al., 2003; Roach, 2000) produced fine labellings in which a single utterance might be described as moving from hate to vengeful anger. The result is a very low incidence rate of emotional states that appears in most naturalistic data. Two main types of scheme that involve assigning terms like angry, ashamed, jealous, etc. can be expressed in terms of everyday language.

A finer theoretical tool comes from appraisal theory, which is probably the most influential approach to emotion within psychology at present (Scherer, 1999). Appraisal theory offers a descriptive framework for emotion based on the way the person involved experiences the events, things, or people at the focus of the emotional state. This is described involving five broad headings:

- Novelty (involving suddenness, familiarity and predictability as distinct subcategories)
- Intrinsic pleasantness
- Goal significance (including the aspect of life to which the events are relevant, from the body to relationships to social order; the likelihood that there will be a significant outcome; how the events relate to expectation, and to the agent’s goals; and how urgent they are)
### Table 1

The emotion-related words rated most useful for describing three types of day (from most to least useful)

<table>
<thead>
<tr>
<th>Typical day</th>
<th>Good day</th>
<th>Bad day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amusement</td>
<td>Pleasure</td>
<td>Stress</td>
</tr>
<tr>
<td>Friendliness</td>
<td>Amusement</td>
<td>Sadness</td>
</tr>
<tr>
<td>Pleasure</td>
<td>Delight</td>
<td>Anxiety</td>
</tr>
<tr>
<td>Affection</td>
<td>Excitement</td>
<td>Frustration</td>
</tr>
<tr>
<td>Love</td>
<td>Satisfaction</td>
<td>Hurt</td>
</tr>
<tr>
<td>Politeness</td>
<td>Affection</td>
<td>Annoyance</td>
</tr>
<tr>
<td>Contentment</td>
<td>Content</td>
<td>Despair</td>
</tr>
<tr>
<td>Interest</td>
<td>Friendliness</td>
<td>Tension</td>
</tr>
<tr>
<td>Trust</td>
<td>Happiness</td>
<td>Worry</td>
</tr>
<tr>
<td>Excitement</td>
<td>Love</td>
<td>Disappointment</td>
</tr>
<tr>
<td>Calm</td>
<td>Interest</td>
<td>Powerlessness</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Hopefulness</td>
<td>Shock</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Modality</th>
<th>Elicitation method</th>
<th>Size</th>
<th>Language</th>
<th>Type of emotional description</th>
<th>Emotional content</th>
<th>Electronic contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belfast naturalistic database (Douglas--Cowie et al., 2003)</td>
<td>AV</td>
<td>Natural: 10–60 s long ‘clips’ taken from television chat shows, current affairs programmes and interviews conducted by research team</td>
<td>239 clips; 209 from TV recordings, 30 interview recordings. 125 subjects; 31 M, 94 F</td>
<td>English</td>
<td>Dimensional labelling/Categorical labelling</td>
<td>Wide range</td>
<td><a href="http://emotion-research.net/ws/wp5/databasesLQ.ppt">http://emotion-research.net/ws/wp5/databasesLQ.ppt</a></td>
</tr>
<tr>
<td>EnoTV corpus</td>
<td>AV</td>
<td>Natural: 4–43 s long ‘clips’ from French television channels</td>
<td>51 clips from French TV channels; 48 different individuals</td>
<td>French</td>
<td>Categorical labelling</td>
<td>Wide range</td>
<td><a href="http://emotion-research.net/ws/wp5/SALAS.ppt">http://emotion-research.net/ws/wp5/SALAS.ppt</a></td>
</tr>
<tr>
<td>SAL (sensitive artificial listener) database</td>
<td>AV</td>
<td>Induced: subjects talk to artificial listener and emotional states are changed by interaction with different personalities of the listener</td>
<td>Approx. 10 h; study 1, 20 subjects; study 2, 4 subjects × 20 min sessions each</td>
<td>English</td>
<td>Dimensional labelling</td>
<td>Wide range of emotions/emotion related states but not very intense</td>
<td><a href="http://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom">http://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom</a>;</td>
</tr>
<tr>
<td>SMARTKOM</td>
<td>AVG</td>
<td>Human machine in WOZ scenario: solving tasks with system</td>
<td>224 speakers; 4/5 min sessions</td>
<td>English, German</td>
<td>Categorical labelling holistic and for each channel in turn Intensity also labelled</td>
<td>Joy, anger, gratification, irritation, helplessness, pondering, reflecting, surprise, neutral</td>
<td><a href="http://www.phonetik.uni-muenchen.de/BasPublications/Steininger-02-LREC.pdf">http://www.phonetik.uni-muenchen.de/BasPublications/Steininger-02-LREC.pdf</a></td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Name</th>
<th>Modality</th>
<th>Elicitation method</th>
<th>Size</th>
<th>Language</th>
<th>Type of emotional description</th>
<th>Emotional content</th>
<th>Electronic contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>DaFEx Database of Kinetic Facial Expressions (Batocchi &amp; Pianesi, 2004)</td>
<td>AV</td>
<td>Acted: instructions given to actors. Use of ‘scenarios’, one for each emotion</td>
<td>1008 clips (4–27 s each) coming from recordings with 8 professional actors (4 male, 4 female) total recording time: 2 h 50 min 13 s</td>
<td>Italian</td>
<td>Categorical emotional labelling</td>
<td>Ekman 6 emotions plus neutral expression, 3 intensity levels, 2 acting conditions: ‘Utterance’ and ‘Non-Utterance’.</td>
<td><a href="mailto:batocchi.pianesi@itc.it">batocchi.pianesi@itc.it</a></td>
</tr>
<tr>
<td>eWiz database (Auberge et al., 2003)</td>
<td>AVP</td>
<td>Induced using a WoZ scenario simulating a language teaching exercise</td>
<td>17 subjects (10 general, 7 actors) 3400 single words plus comments 5 h speech, 15 h of non-verbal</td>
<td>French</td>
<td>Labels chosen by subjects (category and dimensional)</td>
<td>Concentration, positive, negative, stressed, satisfaction, boredom, surprise, disappointment, anxious, confused</td>
<td><a href="http://www.icp.inpg.fr/EMOTION">http://www.icp.inpg.fr/EMOTION</a></td>
</tr>
<tr>
<td>Speech motion face database</td>
<td>AV</td>
<td>Acted</td>
<td>Three different expression variations for each emotion and 1326 utterances: frustrated 59, happy 85, neutral 86, sad 100, angry 112</td>
<td>English</td>
<td>Pre-defined</td>
<td>Frustrated, happy, neutral, sad, angry</td>
<td><a href="http://www.cs.ucla.edu/~abingcao/Documents/Cao_TOG_small.pdf">http://www.cs.ucla.edu/~abingcao/Documents/Cao_TOG_small.pdf</a></td>
</tr>
<tr>
<td>Polzin and Waibel (Polzin &amp; Waibel, 2000)</td>
<td>AV</td>
<td>Acted: sentence length segments taken from acted movies</td>
<td>Unspecified no of speakers: Segment numbers 1586 angry, 1076 sad, 2991 neutral</td>
<td>English</td>
<td>Choice from a given set of emotion tags</td>
<td>Anger, sadness, neutrality (other emotions, but in insufficient numbers to be used)</td>
<td><a href="http://www.cs.ucla.edu/~abingcao/Documents/Cao_TOG_small.pdf">http://www.cs.ucla.edu/~abingcao/Documents/Cao_TOG_small.pdf</a></td>
</tr>
<tr>
<td>Banse and Scherer (Banse &amp; Scherer, 1996)</td>
<td>AV</td>
<td>Acted: actors were given scripted eliciting scenarios for each emotion, then asked to act out the scenario. (visual info used to verify listener judgements of emotion)</td>
<td>12 subjects (6 M, 6 F)</td>
<td>German</td>
<td>Expert (peer) rating of authenticity; rating from 14 emotional categories</td>
<td>Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame</td>
<td></td>
</tr>
<tr>
<td>AIBO ROBOT DATABASE (Erlangen database) (Batliner, Hacker, Steidl, Noth, D’Acre, Russell, &amp; Wong, 2004)</td>
<td>AV/A</td>
<td>Task directions to robot</td>
<td>51 German children, 51,393 words</td>
<td>German</td>
<td>Categorical labelling of emotion by several labellers</td>
<td>Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral</td>
<td><a href="http://pfstar.itc.it/public/publications/lrec04.pdf">http://pfstar.itc.it/public/publications/lrec04.pdf</a></td>
</tr>
<tr>
<td>Amir, Ron, and Laor (2000)</td>
<td>AP</td>
<td>Induced: subjects asked to recall personal experiences involving each of the emotional states</td>
<td>140 subjects; 60 Hebrew speakers, 1 Russian speaker</td>
<td>Hebrew, Russian</td>
<td>Categorical labelling</td>
<td>Anger, disgust, fear, joy, neutrality, sadness</td>
<td><a href="http://www.qub.ac.uk/en/isca/proceedings/pdfs/amir.pdf">http://www.qub.ac.uk/en/isca/proceedings/pdfs/amir.pdf</a></td>
</tr>
<tr>
<td>Study</td>
<td>Data Source</td>
<td>Number of Subjects</td>
<td>Type of Interaction</td>
<td>Duration</td>
<td>Language(s)</td>
<td>Emotions Studied</td>
<td>Methodology</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>--------------------</td>
<td>---------------------</td>
<td>----------</td>
<td>-------------</td>
<td>------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ORESTEIA database</td>
<td>Balomenos, Cowie, Malinescos, McMahon, and Kasderidis (2003)</td>
<td>77 subjects; 61 Korean speakers, 6 Americans</td>
<td>range of topics including sad and joyful moments in their lives</td>
<td>29 subjects, 90 min sessions per subject</td>
<td>English</td>
<td>Dimensional labeling/Categorical labeling</td>
<td>English Performance in tasks used as index of subject state</td>
</tr>
<tr>
<td>Kim, Bang and Kim</td>
<td>P Ledbetter and Kim</td>
<td>29 subjects, 90 min sessions per subject</td>
<td>English</td>
<td>n/a</td>
<td>Stress, irritation, shock</td>
<td>Pre-defined emotional statuses</td>
<td></td>
</tr>
<tr>
<td>The HUT facial expression database</td>
<td>AV</td>
<td>2 actors</td>
<td>n/a</td>
<td>n/a</td>
<td>Pre-defined</td>
<td>Anger, disgust, fear, surprise happiness, sadness</td>
<td></td>
</tr>
<tr>
<td>Reading-Leeds database</td>
<td>A Ledbetter and Kim</td>
<td>5 h</td>
<td>English</td>
<td>n/a</td>
<td>Emotion labels</td>
<td>Range of full blown emotions</td>
<td></td>
</tr>
<tr>
<td>Campbell CREST database (ongoing)</td>
<td>A Ledbetter and Kim</td>
<td>1000 h over 5 yrs at 2004, 500 h</td>
<td>English, Japanese, Chinese</td>
<td>Three levels of label: speaker state, speaking style, voice quality</td>
<td>Stock exchange: 100 dialogues, 5000 speaker turns; Capital Bank: 250 dialogues, 5000 turns of word extracts</td>
<td>Two annotators using categorical labels</td>
<td></td>
</tr>
<tr>
<td>Capital Bank Service and Stock Exchange Customer Service (used by Devillers &amp; Vasilescu, 2004)</td>
<td>A Ledbetter and Kim</td>
<td>115 subjects 48 females, 67 males, controls (therapists), 38, patients 77</td>
<td>English</td>
<td>DSM-IV diagnosis of depression</td>
<td>Natural: therapy sessions and phone conversations. Post therapy evaluation sessions were also used to elicit speech for control subjects.</td>
<td>Dysthemic disorder; major depression (single episode and recurrent, no psychiatric features)</td>
<td></td>
</tr>
<tr>
<td>France et al. (France, Shiava, Silverman, Silverman, &amp; Wilkes, 2000)</td>
<td>A Ledbetter and Kim</td>
<td>110 dialogues, 29,200 words (i.e. tokens, not vocabulary)</td>
<td>German</td>
<td>Word-based emotional user states</td>
<td>Human machine dialogue system</td>
<td>Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic</td>
<td></td>
</tr>
<tr>
<td>SYMPAFLY (as used by Batliner, Hacker, Steidl, Noth, D’Arcy, Russell, &amp; Wong)</td>
<td>A Ledbetter and Kim</td>
<td>110 dialogues, 29,200 words (i.e. tokens, not vocabulary)</td>
<td>German</td>
<td>Word-based emotional user states</td>
<td>Human machine dialogue system</td>
<td>Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic</td>
<td></td>
</tr>
</tbody>
</table>

For more information, please visit: [http://www.iece.org/Publish/Journals/Prof-Journ/MBEC/20043886.pdf](http://www.iece.org/Publish/Journals/Prof-Journ/MBEC/20043886.pdf)
<table>
<thead>
<tr>
<th>Name</th>
<th>Modality</th>
<th>Elicitation method</th>
<th>Size</th>
<th>Language</th>
<th>Type of emotional description</th>
<th>Emotional content</th>
<th>Electronic contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belfast Boredom database (Cowie et al., 2003)</td>
<td>A</td>
<td>Human machine dialogue system (Extracts from recordings of simulated interactions with a call centre)</td>
<td>13,187 utterances, average length about 2.75 words, 1750 considered emotional</td>
<td>English</td>
<td>Categorical emotional labeling by 5 students from 1 of 7 possible labels</td>
<td>Frustration, annoyance, neutral, tired, amused, other, not applicable</td>
<td><a href="http://www.speech.cs.cmu.edu/Communicator/">http://www.speech.cs.cmu.edu/Communicator/</a></td>
</tr>
<tr>
<td>DARPA Communicator corpus (as used by Ang et al., 2002)</td>
<td>A</td>
<td>Induced: subjects give verbal responses to maths problems in simulated driving context</td>
<td>Data reported from four subjects</td>
<td>English</td>
<td>Performance in tasks used as indicator of subject state</td>
<td>Stress</td>
<td></td>
</tr>
<tr>
<td>Fernandez &amp; Picard (2003)</td>
<td>A</td>
<td>Dialogue with automated system</td>
<td>1187 calls, total 7200 utterances</td>
<td>English</td>
<td>Tagged: negative and non-negative</td>
<td>Negative, non-negative</td>
<td></td>
</tr>
<tr>
<td>Yacoub et al.</td>
<td>A</td>
<td>Induced: cognitive stress induced through slides containing logical problems; emotional stress induced through slides of human bodies showing skin disease/accident injuries</td>
<td>60 subjects (33 male, 27 female)</td>
<td>German</td>
<td>Low versus high stress</td>
<td>Stress (both cognitive and emotional)</td>
<td></td>
</tr>
<tr>
<td>Tolkmitt and Scherer (1986)</td>
<td>A</td>
<td>Contextualised acting: subjects asked to read passages written with appropriate emotional content</td>
<td>Eight subjects reading paragraph length passages (20–40 mm s each)</td>
<td>Spanish</td>
<td>Emotional categorical labelling</td>
<td>Desire, disgust, fury, fear, joy, surprise, sadness</td>
<td></td>
</tr>
<tr>
<td>Iriondo et al. (2000)</td>
<td>A</td>
<td>Contextualised acting: subjects asked to read passages written with appropriate emotional content</td>
<td>Three subjects reading eight semantically neutral sentences</td>
<td>Dutch</td>
<td>Emotional categorical labelling</td>
<td>Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage,</td>
<td></td>
</tr>
<tr>
<td>Database/Source</td>
<td>Methodology</td>
<td>Participants</td>
<td>Language</td>
<td>Emotional Labelling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------</td>
<td>----------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McGilloway Database (McGilloway, 1997)</td>
<td>Contextualised acting: subjects asked to read passages written in appropriate emotional tone and content for each emotional state (each repeated three times)</td>
<td>40 subjects reading 5 passages each</td>
<td>English</td>
<td>Emotional categorical labelling (Anger, fear, happiness, sadness, neutrality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish Emotional Speech Database</td>
<td>Acted (material not emotionally coloured)</td>
<td>4 subjects read 2 words, 9 sentences and 2 passages in range of emotions</td>
<td>Danish</td>
<td>Emotional categorical labelling (Anger, happiness sadness, surprise, neutrality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groningen ELRA corpus number S0020</td>
<td>Acted: Scripted (material selected to be semantically neutral)</td>
<td>238 subjects reading 2 short texts</td>
<td>Dutch</td>
<td>Unclear (Database only partially oriented to emotion)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin database (Kienast &amp; Sendlmeier, 2000)</td>
<td>Acted 10 subjects (5 M, 5 F) reading 10 sentences each</td>
<td>10 subjects</td>
<td>German</td>
<td>Emotional categorical labelling (Anger, hot, boredom, disgust, fear-panic, happiness, sadness-sorrow, neutrality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leinonen and Hiltunen (Leinonen &amp; Hiltunen, 1997)</td>
<td>Acted 483 samples selected down to 120 (one word utterances)</td>
<td>483 samples</td>
<td>Finnish</td>
<td>Emotional categorical labelling (Angry, frightened, commanding, pleading content (naming), astonished, scornful, admiring, sad)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nakatsu et al. (Nakatsu, Tosa, &amp; Nicholson 1999)</td>
<td>Utterances acted in various emotions then imitated by speakers</td>
<td>100 words, uttered (by 100 speakers) once for each emotion, totaling 800 words</td>
<td>Japanese</td>
<td>Emotional categorical labelling (Neutrality, anger, sadness, fear, disgust, playfulness, surprise, happiness)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nogueiras et al. (2001)</td>
<td>Acted 6 sentences, paragraphs, words, numbers by each of 2 Spanish, French Slovenian, English speakers</td>
<td>14 actors (8m, 6f)</td>
<td>Spanish, French, Slovenian, English</td>
<td>Emotional categorical labelling (Anger, disgust, fear, joy, sadness, surprise, neutrality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Media Team corpus in Finnish</td>
<td>Acted Finnish passage of approx 100 words read</td>
<td>14 actors (8m, 6f)</td>
<td>Finnish</td>
<td>Emotional categorical labelling (Happiness/joy, sadness, fear, anger, boredom, disgust, neutral)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Modality</td>
<td>Type of emotional description</td>
<td>Size</td>
<td>Language</td>
<td>Electronic contact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>----------</td>
<td>--------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUFACS</td>
<td>Naturalistic/Induced</td>
<td>Aiming for 400–800 min each. Aiming for 100 subjects, 2.5 min each. Aiming for 100 subjects, 2.5 min each.</td>
<td>100 subjects</td>
<td>German</td>
<td><a href="http://mplab.ucsd.edu/databases/databases.html">http://mplab.ucsd.edu/databases/databases.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The AR face database</td>
<td>Posed</td>
<td>154 subjects (82 male, 74 female) 26 pictures per person</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html">http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Japanese Female Facial Expression Database (JAFFE) Database</td>
<td>Posed</td>
<td>10 subjects 60 raters using 6 emotion adjectives (sadness, happiness, surprise, anger, disgust, fear)</td>
<td>n/a</td>
<td>English</td>
<td><a href="http://www.mis.atr.co.jp/~mlyons/jaffe.html">http://www.mis.atr.co.jp/~mlyons/jaffe.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CMU PIE Database</td>
<td>Posed</td>
<td>68 subjects 60 subjects using 6 emotion adjectives (joy, surprise, anger, fear, disgust, and sadness)</td>
<td>n/a</td>
<td>English</td>
<td><a href="http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Psychological Image Collection at Stirling</td>
<td>Posed</td>
<td>Aberdeen: 116 subjects, Nottingham scans: 100, Nott-faces-original: 100, Stirling faces: 36</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://pics.psych.stir.ac.uk/">http://pics.psych.stir.ac.uk/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORATOR</td>
<td>Acted</td>
<td>117 samples produced by professional actors, 14 non-actors produced 28 samples</td>
<td>n/a</td>
<td>German</td>
<td><a href="http://mplab.ucsd.edu/databases/databases.html">http://mplab.ucsd.edu/databases/databases.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CMU Facial Expression Database (Cohn-Kanade)</td>
<td>Posed</td>
<td>210 subjects (158 male, 52 female) 30 pictures per person</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Yale Face Database A</td>
<td>Posed</td>
<td>15 subjects 60 raters using 6 emotion adjectives (sad, sleepy, surprised)</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://cvc.yale.edu/projects/yalefaces/yalefaces.html">http://cvc.yale.edu/projects/yalefaces/yalefaces.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CMU PIE Database</td>
<td>Posed</td>
<td>68 subjects 60 subjects using 6 emotion adjectives (joy, surprise, anger, fear, disgust, and sadness)</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Psychological Image Collection at Stirling</td>
<td>Posed</td>
<td>Aberdeen: 116 subjects, Nottingham scans: 100, Nott-faces-original: 100, Stirling faces: 36</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://pics.psych.stir.ac.uk/">http://pics.psych.stir.ac.uk/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORATOR</td>
<td>Acted</td>
<td>117 samples produced by professional actors, 14 non-actors produced 28 samples</td>
<td>n/a</td>
<td>German</td>
<td><a href="http://mplab.ucsd.edu/databases/databases.html">http://mplab.ucsd.edu/databases/databases.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CMU PIE Database</td>
<td>Posed</td>
<td>210 subjects (158 male, 52 female) 30 pictures per person</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Yale Face Database A</td>
<td>Posed</td>
<td>15 subjects 60 raters using 6 emotion adjectives (sad, sleepy, surprised)</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://cvc.yale.edu/projects/yalefaces/yalefaces.html">http://cvc.yale.edu/projects/yalefaces/yalefaces.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CMU PIE Database</td>
<td>Posed</td>
<td>68 subjects 60 subjects using 6 emotion adjectives (joy, surprise, anger, fear, disgust, and sadness)</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Psychological Image Collection at Stirling</td>
<td>Posed</td>
<td>Aberdeen: 116 subjects, Nottingham scans: 100, Nott-faces-original: 100, Stirling faces: 36</td>
<td>n/a</td>
<td>n/a</td>
<td><a href="http://pics.psych.stir.ac.uk/">http://pics.psych.stir.ac.uk/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Coping potential (covering how much control anyone could exercise over the event, how much power the agent in particular has to control them, and how easily the agent can adapt by adjusting his or her own goals)

• Compatibility with standards (covering how far the events conform to or violate moral or ethical standards, held either by the agent personally or by society as a whole)

Work is under way on translating this scheme into a practical tool. One of the main motivations is that the individual items may be linked to component gestures in the expression of emotion (such as eyebrow raising, mouth movement, etc.). That opens the way to deal both with the sequencing of gestures, and with situations where some, but not all of the appraisal elements associated with full-blown emotion are present.

These descriptive schemes are far removed from the pattern that someone entering the field might expect to find in a database. Understanding how the descriptions involved relate to signs of emotion is also a much more interesting task.

3.3. Analyses of emotion-related signals and signs

Not many teams working on neural nets will be able to use data in the form of direct records in the audio, visual, or physiological domain. Most will need material to be analysed on at least one of two other levels, feature extraction and sign description. Feature extraction refers to the output of signal processing operations (perhaps corrected by hand). For example, existing routines can extract positions of key joints from video records. Sign description refers to the ways in which humans spontaneously interpret incoming evidence, usually involving intentionality—for instance, pointing at a car. Automatic transition from extracted features to signs is very difficult.

There are various levels at which neural network research could engage with sequences of signals, features, and signs, given appropriate data. There is still a need for general, robust solutions to many of the basic problems in feature extraction. The mapping of features to signs is an extremely challenging problem, as is the mapping of signs to descriptions of emotion. Neural nets may well also be well suited to some aspects of the inverse problem, synthesis of sign sequences that are appropriate to express a sequence of emotional states.

There is currently rapid evolution in the provision of data relevant to these tasks. The Erlangen group is in the process of making available the AIBO database, with hand corrected transcriptions of verbal content and pitch. SAL data collected under the ERMIS project includes FAPS (Balomenos et al., 2004), basic speech features (pitch and intensity contours and pause boundaries) and higher order features calculated by the ASSESS system. ASSESS features are also available for the Belfast naturalistic database. The EmoTV Corpus contains rich descriptions of signs. Programs under way use motion capture to provide machine-readable descriptions of body movements relevant to emotion (see e.g. Pollick, Hill, Calder, & Paterson 2003).

Records are often not clear about the level of analysis available, and the situation is often evolving. Hence the Appendix does not try to detail the level of analysis associated with data sources. Some additional information is available via the listing of databases on the HUMAINE website (http://emotion-research.net).

3.4. Describing context

Despite the importance of context, there are no generally used schemes for describing it. A recent HUMAINE report included a proposal that at least the following issues should be specified

• Agent characteristics (age, gender, race)
• Recording context (intrusiveness, formality, etc.)
• Intended audience (kin, colleagues, public)
• Overall communicative goal (to claim, to sway, to share a feeling, etc.)
• Social setting (none, passive other, interactant, group)
• Spatial focus (physical focus, imagined focus, none)
• Physical constraint (unrestricted, posture constrained, hands constrained)
• Social constraint (pressure to expressiveness, neutral, pressure to formality)

It is proposed to refine this scheme through work with the HUMAINE databases as they develop. Parts are already implemented in the EmoTV Corpus described in the Appendix.

4. Databases

The Appendix lists the databases that seem most likely to be of interest to the neural net community (and to research on machine learning in general). Items are included for three main reasons.

First and foremost is the use of data that is naturalistic or induced in credible ways. These are covered as fully as possible.

The second is multimodality. Multimodal databases are still rare, but the picture has changed considerably since a relatively recent review (Douglas-Cowie et al., 2003).

The third is influence. Databases which are neither multimodal nor naturalistic have nevertheless been influential for a variety of reasons, one of which may be that they are convenient to access and use. It would be remiss not to include influential sources in a systematic review.
5. Conclusion

Ekman and Friesen’s database of emotional faces remains for many people the example of what they expect an emotion database to be. The databases listed in the first part of the Appendix are very different. It seems likely that the future trend will be towards databases more like the latter than the former. One of the main aims of this paper has been to convey the background against which that move has occurred, so that people who are inclined to expect something like the Ekman and Friesen resource understand why they are being pointed to something so different.

It is also the case that assembling databases has not traditionally been considered a high-profile or intellectually challenging area. Good quality recording and large balanced samples tend to be thought of as the basic requirements, with the human side assumed to be relatively straightforward.

A little thought shows that in the domain of emotion, that cannot be the case. The human race expends a huge proportion of its resources trying (with mixed success) to direct people out of some emotional states and into others. If it were easy to achieve the shifts, there would be no need for whole industries and cities to exist.

As a result, capturing a faithful, detailed record of human emotion as it appears in real action and interaction is a massively challenging task. Nevertheless, the payoff is also large. At root, it is enlisting computers to co-operate in the old task of directing people away from some emotional states and into others. The lure of technologies capable of doing that is enough to keep the enterprise going in spite of the difficulties.

6. Uncited references

Douglas-Cowie et al. (2004); Lee and Narayanan (2005); Tsapatsoulis et al. (2002); Ward (2003).

Acknowledgements

Preparation of this paper was supported by EC projects ERMIS (IST-2000-29319) and HUMAINE (IST-2003-507422).

Appendix

List of key databases for computational research on emotion. The abbreviations in the ‘modality’ column are as follows: A = audio; V = video; G = gesture; P = physiology. Column 1 cites published sources where they are available: the corresponding references are in the general list at the end of the paper.

References

Evaluation and comparison of the Velten and musical procedures. 
Advances in Behavioral Research and Therapy, 5, 27–49.

Cornelius, R. (1966). The science of emotion: Research and tradition in the 

Cowie, R. (2005). What are people doing when they assign everyday 

Cowie, R., & Cornelius, R. (2003). Describing the Emotional States that are 

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., & 
words. In N. Mastorakis (Ed.), Computational intelligence and 
applications (pp. 109–114). World Scientific Engineering Society, 
109–114.

Speech in the process of becoming bored Proceedings of 15th 
international congress of phonetic sciences, Barcelona.

In S. Bengio, & H. Bourlard, Machine learning for multimodal 
interaction: Lecture notes in computer science (Vol. 3361) (pp. 305– 

Cox, C. (2004). SALAS (Sensitive Artificial Listener induction tech-
niques). Paper presented to HUMAINE Network of Excellence Summer 
net/ws/summerschool1/SALAS.ppt#259.

emotion in dialogue AAI spring symposium on exploring emotion and 
affect in text: Theories and applications. Stanford University: AAI 
Press pp. 44–49.

Davidson, R. J. (1999). Neuropsychological perspectives on affective styles 
and their cognitive consequences. In T. Dalgleish, & M. Power (Eds.), 
Handbook of cognition and emotion (pp. 103–125). New York: Wiley, 
103–125.

Descartes, R. (1911–1912). The passions of the soul (1649). In E. L. 
Haldane, & G. R. Ross (Eds.), The philosophical works of descartes. 
Cambridge: Cambridge University Press.


speech: Towards a new generation of databases. Speech Communication, 
40(1–2), 33–60.

Douglas-Cowie et al. (2004). Multimodal data in action and interaction: A 
library of recordings and labelling schemes. Report for EC funded IST 
Network of Excellence 507422 (HUMAINE—Human–Machine Inter-
emotion-research.net/deliverables/D5d%20potential%20examples 
%20databases.pdf.

Handbook of cognition and emotion (pp. 301–320). New York: Wiley, 
301–320.

Categories, origins, usage and coding. Semiotica, 1, 49–98.

Consulting Psychologists’ Press.

recognizing emotions from facial clues. Englewood Cliffs, NJ: Prentice-
Hall, Inc.

Ekman, P., Friesen, W. V., O’Sullivan, M., Chan, A., Diacoynni-Tarlatzis, 
I., Heider, K., et al. (1987). Universals and cultural differences in the 
judgments of facial expressions of emotion. Journal of Personality and 

Speech Communication, 40, 145–159.

Acoustical properties of speech as indicators of depression and suicidal 
risk. IEEE Transactions on Biomedical Engineering, 47(7).

Press.

Goldie, P. et al. (2004). Report on ethical prospects and pitfalls for emotion-
oriented systems. Report for EC funded IST Network of Excellence 
507422 (HUMAINE). Retrieved 22nd March 2005 from http://emotion-
research.net/deliverables/D10a.pdf.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., 
et al. (1995). Representation of prosodic and emotional features in a 
spokean language database Proceedings XIIIth international congress of 

and speech: Methodological issues in naturalistic approaches. 
Language and Speech, 43, 355–375.

Cognition and Emotion, 9, 87–108.

Hadfield, P., & Marks, P. (2000). This is your captain doing. New Scientist, 
1682267, 21.

Iriondo, I., et al. (2000). Validation of an acoustical modelling of emotional 
expression in spanish using speech synthesis techniques Proceedings of 
the ISCA workshop on speech and emotion, Belfast 2000.


expression and music performance. Psychological Bulletin, 129(5), 
770–814.


and temporal changes in emotional speech. In R. Cowie, E. Douglas-
Cowie, & M. Schroeder (Eds.), Speech and emotion: Proceedings of the 

Kollias, S. et al. (2004). The facial feature extraction subsystem. Report 
(Deliverable D10) for EC funded fifth framework IST project IST-
2000–29319 ERMIS (Emotionally Rich Man-machine Intelligent 

Kollias, S. et al., 2005. Presentation to the final review of the ERMIS 
project (EC funded Fifth Framework IST project IST-2000-29319), 
Athens, May 2005.

affective picture system IAPS: Technical manual and affective ratings 
The center for research in psychophysiology. Gainesville, FL: 
University of Florida.

Laukka, P. (2005). Verbal expressions of emotion: Discrete emotions and 
dimensional accounts. Acta Universitatis Upsaliensis, Comprehensive 
Summaries of Upsala Dissertations from the Faculty of Social 
Sciences, 141.

neurology. Section I: The nervous system. Higher functions of the 


dialogs. IEEE transactions on speech and audio processing, 13(2), 
293–303.

connotations with a one-word utterance. Journal of the Acoustical 
Society of America, 102(3), 1853–1863.


McGilloway, S. (1997). Negative symptoms and speech parameters in 
schizophrenia. Unpublished doctoral thesis, Queen’s University, 
Belfast.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielens, S., Westerdijk, M., 
(Eds.), Speech and emotion: Proceedings of the ISCA workshop. 


Sweeney, F. (2005). The words people use to describe their emotional lives. BSc thesis: School of Psychology, Queen’s University Belfast.


