Issues in data labelling

Roddy Cowie, Cate Cox, Jean-Claude Martin, Anton Batliner, Dirk Heylen, Kostas Karpouzis

Angaben zur Veröffentlichung / Publication details:

Cowie, Roddy, Cate Cox, Jean-Claude Martin, Anton Batliner, Dirk Heylen, and Kostas Karpouzis. 2011. "Issues in data labelling." In *Emotion oriented systems: the humaine handbook*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, 213–41. Heidelberg [u.a.]: Springer. https://doi.org/10.1007/978-3-642-15184-2_13.

Issues in Data Labelling

Roddy Cowie, Cate Cox, Jean-Claude Martin, Anton Batliner, Dirk Heylen, and Kostas Karpouzis

Abstract Labelling emotion databases is not a purely technical matter. It is bound up with theoretical issues. Different issues affect labelling of emotional content, labelling of the signs that convey emotion, and labelling of the relevant context. Linked to these are representational issues, involving time course, consensus and divergence, and connections between states and events. From that background comes a wealth of resources for labelling emotion, involving not only everyday emotion words but also affect dimensions, and labels for combination types, appraisal categories, and authenticity. Resources for labelling signs of emotion cover linguistic, vocal, face descriptors, plus descriptors for gesture, and relevant physiological variables. Resources for labelling context are developing.

1 Introduction

The aim of this is to give readers a sense of the challenge and of the kinds of partial solution that are now available.

Labelling is a challenge because emotion-related phenomena are massively complex and surrounded by uncertainty. The first chapter of this handbook conveys some of the complexity and that is taken for granted as background in this chapter.

Pulling against the challenge of complexity is set the need to formulate tasks that are simple enough for a machine to carry out. Given a brief video clip of a person in an emotionally coloured episode, a human being can develop a commentary of enormous richness, which is unique to that particular clip. In contrast, groups concerned with practical machine learning are likely to be unhappy if more than one label is used for each clip or more than a dozen for the whole database (two being much the preferred number).

R. Cowie (✉)
Department of Psychology, Queen’s University Belfast, Belfast, Northern Ireland, UK
e-mail: roddy.cowie@qub.ac.uk

Working labelling schemes are always the outcome of tension between the two sets of factors, one pulling towards complexity and the other towards simplicity. As a result, they are likely to frustrate both engineers, who want something simpler, and philosophers and social scientists, who deplore their lack of richness.

This chapter rests on the assumption that there is space for a sub-discipline between the extremes. Constructing a credible emotion database is a major undertaking. The undertaking is hardly worthwhile unless the result can be used by many teams. Hence, it is not optimal to invest effort in a minimal labelling, which is totally driven by the needs of a particular synthesis or analysis project. It makes more sense to construct a labelling that is rich enough to support any foreseeable use of the basic material. Using a rich coding scheme means that researchers can pick up and adapt what they are interested in according to their application requirements. Simplifying a rich initial labelling is not trivial, but it is possible. Enriching a poor initial labelling is likely in practice to mean restarting from scratch. That is why this chapter does not restrict itself to labelling schemes that are likely to be used directly in learning or synthesis applications in the immediate future.

The chapter is divided into two main sections. The first considers general concepts, theoretical and technical, that create the framework for work on databases. The second deals with specific resources that exist for labelling and studies that illustrate them. The ideas that influenced the HUMAINE database are naturally prominent, but not to the exclusion of others.

It is worth stressing that the chapter is very far from comprehensive. Labelling a rich emotional database potentially involves an enormous number of skills, and it seems fair to say that they have rarely if ever all been drawn together. The labelling of the HUMAINE database (described in a later chapter) is probably the most ambitious exercise of its kind, and the chapter reflects the expertise that undertaking has been drawn together. But there are still many areas where expertise exists that is not reflected here or where the expertise to handle what are clearly significant problems simply does not exist.

2 General Concepts

Some terms are used here in ways that reflect practice within HUMAINE and that may not be immediately obvious.

The term ‘records’ is used to describe the video, audio, physiological, or other signals that form the raw material of the database. The term ‘clip’ is used to describe an individual record which has been edited to capture a natural unit.

Terms related to emotion and emotion-related states are used in line with the opening chapter of this handbook. When ‘emotion’ is used without a specific modifier, it is used in its broad sense to cover the whole range of phenomena that distinguish life as it normally is from the life of an agent who is emotionless. These phenomena include moods, stances, and bonds, as well as the particular states that are called emotion in a narrow sense – both brief episodes of emergent emotion and established emotions, which shape the way a person feels and acts over months or years.

Emotion labellings are symbols that express the emotion or emotional colouring in a clip (whether it is truly present, perceived to be). Sign labellings identify evidence in the records that is relevant to attributing emotion labels – facial expressions, attributes of speech, gestures, etc. They exclude context labels. These identify features of the situation that are relevant to understanding why a particular emotion might be felt, or why it might be expressed in a particular way.

2.1 Theoretical Foundations

There is a well-known dictum that all data are theory laden (Hanson, 1958). It is most certainly true of data created by labelling. Experience suggests that people coming into the field often do not register that apparently simple choices are bound up with questions on which there are large literatures. The aim of this section is to alert people to the theoretical dimension of practical decisions about labelling. It deals first with emotion and emotion-related states, then looks at sign labellings, then at descriptions of context.

2.1.1 Theoretical Foundations of Emotion Labelling

The first chapter of this handbook reviews the main resources that theories of emotion offer. This section does not repeat the ground. Its aim is just to make the connection between particular types of theory and emotion descriptors that a database may use.

Most people are at home with atheoretical descriptions of emotion, based on everyday emotion words. There is a school of thought that assumes those descriptions are the best we have and the natural basis for labelling. It is a defensible position. The important thing is to register that it is by no means obviously right.

A theoretical position with very broad appeal proposes that a few everyday categories (six in the best known version) correspond to the elements of emotional life. From the point of view of labelling, it would be very convenient if emotion came in half a dozen 'basic' flavours, each corresponding to an everyday category. Unfortunately, even the main advocate of that position no longer defends it (Ekman, 1999).

An alternative framework was formulated long before basic emotion theories. It describes emotion in terms of dimensions. Russell and Barrett-Feldman (1999) have recently updated the idea, suggesting that the classical dimensions capture an element of emotion which he calls core affect. On that view, the natural format for labelling is to attach parameters describing core affect to each significant unit.

Appraisal theories give rise to approaches that are more sharply distinct from common sense. They suggest that emotion is rooted in a distinctive way of weighing up situations that impinge on an individual and provide lists of issues that are involved in the weighing up. That suggests that labelling could describe the relevant issues rather than using everyday emotion categories.

A very durable version of appraisal theory was proposed by Ortony et al. (1988) and is commonly known as OCC. It considers emotions as valenced reactions. The theory divides possible types of reaction first in terms of whether they focus on aspects of objects, actions of agents, or consequences of events. The next level of division hinges on whether issues at stake relate to the subject himself or herself, self or to others. Below that, states are classified in terms of whether they are concerned with fortunes, prospects, well-being, attribution, or attraction. The point here is not to set out the theory but to convey the kinds of concept that it suggests might be used rather than commonplace emotion words.

OCC is by no means the only appraisal-based framework. Probably the most fully developed alternative is Scherer's (Sander et al., 2005). He regards emotions as appraisals concerned with the novelty, intrinsic pleasantness, goal significance, and compatibility with standards of focal events, and also the potential for the subject to cope with them. Each of these categories is subdivided again.

There are several ways in which these theories (and others) might be useful to people concerned with labelling databases.

Economy. The sheer number of everyday words that describe emotion-related states makes them an unwieldy descriptive system. Several theories hold out the prospect of doing essentially the same job with a smaller number of well-chosen primitives.

Consistency. Investigators sometimes try to achieve economy by using 'cover classes' based on the way everyday labels appear to cluster in their data set. The difficulty is that different data sets suggest different cover classes, and that limits the prospect of transferring of insight across studies. Basing cover classes on theory reduces the number of options that are likely to be used.

Intermediates. Despite the number of everyday emotion terms, naturalistic databases often show states for which there is no exactly appropriate everyday term. Several theory-based schemes offer ways of describing these elusive states (in terms of co-ordinates or lists of appraisal-related features).

Semantics. Appraisal-based theories in particular offer ways of representing the meaning associated with a label, and that is important for the prospects of using the label in a functional system (recognising not only the state but what it may lead to, allowing analysis of the circumstances to influence the states that are synthesised, and so on). Hence there are advantages to using labels for which that kind of semantics is available.

Natural classes. One of the functions of a theory is to divide emotion-related phenomena into classes that make logical sense rather than ad hoc groupings. Presumably using well-chosen classes will in the long run allow systems to work better. That said, it is not obvious that the same classes are suitable for all functions. It may be important to distinguish envy and anger morally but inappropriate to separate them for the purpose of a database concerned with training systems to recognise emotions. Choosing the right classification for the right purpose is a subtle problem.

There is a large body of literature about the issues that have been sketched above. However, there are many questions where there clearly is a theoretical dimension, but much less material is available. The most important are sketched briefly below.

Part of the debate over basic emotions is whether emotion-related phenomena are continuous or divide into discrete types (Ortony and Turner, 1990; Ekman, 1992). That is relevant to a substantial issue in labelling, which is whether to use continuous or categorical descriptions. The issue is complicated because practical schemes can combine both. For instance, what has been called 'soft labelling' uses categorical labels, but each category is associated with a number that conveys how much of that emotion type is present (Steidl et al., 2005). Conversely, accounts that are inherently dimensional often describe the range of a variable in terms of a few categories (Craggs and Woods, 2004). Evaluating schemes like these depends both on the way the underlying systems actually work and on the practicalities of obtaining labellings.

The theory that has been mentioned so far is predominantly concerned with emergent emotion. That may well not be the most important kind of emotion-related phenomenon for computing to deal with. HUMAINE developed an approach to describing other emotion-related states, which is summarised in the first chapter of this handbook. Around the same time, a taxonomy developed by Baron-Cohen (2007) attracted increasing interest in the computational community.

Cutting across these is the distinction that has been drawn between 'cause'- and 'effect'-type descriptions (Cowie et al., 2001). Cause-type labelling sets out to associate records of a person generating certain signs with descriptions of the states that actually led them to produce the signs – whether he/she was actually angry, despite a calm surface appearance, or calm, despite giving signs of anger. Effect-type labelling sets out to associate the records with descriptions of the impressions that the signs would be expected to produce in an observer. Within that, a distinction has been drawn between the result of considered judgment and 'perceived flow of emotion' – the kind of impression that somebody faced with that person would be expected to form in real time.

There is a tendency to look automatically for cause-type labelling. That poses difficult problems, because there is no easy way to establish what a person is truly feeling. Practically, though, effect-type labelling will often be the more appropriate target conceptually as well as the easier one. The aim of an expressive agent will generally be to create a specified effect on people who encounter it, and at least in ordinary interaction, the natural target for recognition is to interpret signs as a person would.

Individual differences are a key issue for effect-type labelling. Differences between cultures are increasingly well documented (Matsumoto, 2001). What strikes one culture as overt anger will strike another as mild perturbation. Hence it is important to consider and report the culture of the people who generated an effect-type labelling. Even within a culture, it is clear that different people 'read' signs of emotion differently. Differences in simple sensitivity are known from research on emotional intelligence (Mayer et al., 2000). Standard variables like gender and introversion–extraversion probably have quite considerable effects (Hall and Matsumoto, 2004). More complex differences hinge on different judgments about what is genuine and what is being concealed or simulated.

These points impact on well-known technical issues surrounding validity and reliability. Technologists often equate validity with the provision of a 'ground truth'. In the context of emotion, demanding ground truth is tantamount to focusing on a very special subset of the domain. Ground truth is neither known nor needed in most everyday emotionally coloured interactions, and it is difficult to imagine how it could be provided, hence insisting on ground truth rules those interaction 'out of bounds' for research. Paradoxically, acted emotion is sometimes seen as more acceptable, because the actor's intention is known – despite the fact that there is no truth at all. If the field is to engage with subtle phenomena, it needs subtler notions of validity.

Similarly, it is standard to take statistical measures of reliability (specific techniques are discussed below) and to reject material where coefficients are low. Again, blind application of that strategy has the effect of excluding important kinds of material, particularly material whose character means that observers do and should either differ systematically or agree that there is uncertainty. There are some numerical strategies that can be used for that kind of material; again, they are discussed below.

These issues relate to the selection of labellers. A strong commitment to effect-type description would suggest that labellers should ideally be a representative sample of the population, preferably not contaminated by training that would make them less representative. In practice, it is commoner to use a strategy that fits logically with a cause-type approach, which is to use a few 'experts' whose judgments are supposed to be more reliable than average. That is partly because of concern that genuinely naïve raters may simply not understand the task, and therefore scatter in the data will reflect plain confusion as well as genuine variation. Consensus will probably develop over what can and cannot be done with truly naïve raters.

2.1.2 Theoretical Foundations of Sign Labelling

General theoretical issues are less complex than in the area of sign labelling, but some points should be made.

The area spans very different levels of interpretation. At one extreme are patterns that can be derived automatically from the record, such an F0 contour or a particular degree of eyebrow raising. At the other extreme are descriptions that identify the intention behind patterns in the record, for instance, describing a hand movement as raising a fist or voice quality as strained. These may not be much easier to derive from the raw signals than judgements about the person's emotional state.

When humans explain how they attribute emotion, the descriptors that they use are often quite deeply interpretive. It is debatable how relevant labels of that kind are in a database. On one side, they may not identify features that can be recognised as a preliminary step towards identifying a global emotional state. On the other, they may identify action patterns that are relevant to synthesising a particular state. They may also contribute to less sequential recognition strategies, where the system is making multiple kinds of high-level attribution simultaneously and favours interpretations that ‘hang together’ in a coherent constellation.

Descriptors based on shallow interpretation carry their own problems. Most obviously, they are often tied to the specific situation in which they were developed and take on a different meaning when a person is carrying out a different activity, or even when camera angle is changed.

There are also long-standing difficulties over the status of schemes that have been developed to describe significant types of event for different purposes. In the speech domain, there are established systems for describing prosody (the so-called British system and ToBI). In the domain of action, there are schemes that were developed to describe dance and sign language. These have the attraction of being standard. The difficulty is that they may abstract away details that are irrelevant to the original application, but highly relevant to emotion. For instance, ToBI discards harmonic information (such as information about major and minor intervals), but there is a growing interest in the idea that they may be relevant to the expression of valence (Schreuder et al., 2006).

2.1.3 Theoretical Foundations of Context Labelling

People often doubt the need for context labelling. However, there are good reasons why some form of it is needed.

Emotion is very often a reaction to eliciting events – in the famous example from James (1884), a bear emerging from the woods. There are at least many cases where it would be very difficult to read the meaning of emotional signs without knowing about the eliciting circumstances (as anyone will know who has watched tears streaming down the face of a winning sports team). Conversely, it is clearly important to understand human ability to infer the eliciting event from emotion-related signs, without independent information. That is, for instance, what allows people to use another person’s face to orient to a danger (which is an ability that could be very important to emulate in dangerous situations). On a fine level, there is good reason to think that the timing of signs relative to eliciting events is integral to natural expression of emotion (Sander et al., 2005). Databases cannot be used to study any of these issues unless they contain the relevant contextual information.

There is also a body of work on the way social context affects signs of emotion. The best known approach uses the concept of display rules (Ekman and Friesen, 1975). For example, the presence of in-laws is likely to have radical effects on the way couples express feelings towards each other. The display rule concept may oversimplify the relationship between social context and emotion, but it is not in doubt that the relationship is a powerful one. Again, unless databases contain appropriate contextual information, techniques based on them cannot be sensitive to these issues.

Physical context also affects signs. Hand movements are an archetypal example. They depend heavily on what the hand is holding: nothing, a pencil, a hammer, a shopping bag, another person's hand. Trying to interpret their emotional significance without that context would be very difficult. Background noise and distance from an interlocutor have similarly large effects on speech.

For some of these effects, it is natural to assume that information will be attached to a clip rather than labelled in the usual sense (weeping sportspeople are a good example). But others can change very quickly, and timing information may be critical. Hence it is at least sometimes natural to integrate information about context fully within the labelling framework.

2.2 Representational Issues

This section covers issues that are relevant to all the types of labelling that are being considered, and that involve a mix of practical and theoretical issues.

2.2.1 Representing Time Courses

Emotional life involves events that unfold over time. There are various ways in which labelling can reflect the temporal aspect.

The simplest option is to attach a label to a clip as a whole. In the chapter "The HUMAINE Database", we have called that kind of label global. In and of itself, a global label says nothing about any variation in time within the clip. That kind of labelling is very useful for indexing or reviewing the content of a database. For attributes that change slowly, it may be all that is needed. It is, for instance, the natural way to handle information about elicitation techniques, technicalities of recording, and some kinds of context. It would be very useful to move towards a standard approach to global labelling, but it is not a problem that has received much explicit attention.

The use of global labelling is related to the issue of choosing clip boundaries, which is discussed elsewhere in this handbook. For some purposes, it makes sense to divide basic records into small clips, such as individual words. Most labelling will then be global in the sense used here – attached to the clip as a whole. The cost is that more slowly changing variables have to be dealt in terms of relationships between clips. It remains to be established where the overall advantage lies.

The archetypal labelling strategy is to associate labels with discrete time periods – for instance, to attach the label 'happy' to a particular time period and identify the beginning and end of the period. That can be called quantised labelling.

The natural alternative to quantised labelling is what we have called trace labelling. A trace specifies how a measure associated with a particular label varies moment by moment. That general description covers a great variety of possibilities. One trace might show the ratio of maximum horizontal to maximum vertical body extent at it varied from moment to moment; another might show rise and fall in the overall intensity of emotion that a person was experiencing; another might show how a particular emotion (such as happiness) fluctuated over time.

These different representations are often (though not necessarily) associated with different kinds of labelling process. Quantised labelling tends to be done by playing and replaying recordings to find the best point to draw a boundary. Trace labelling tends to be done in real time by moving a cursor (or some other device) while a clip is playing.

Linked to these are semantic differences. The different labelling processes tap into different kinds of percept: in the terms introduced earlier, considered judgment and perceived flow of emotion, respectively. They also suit different kinds of entity. If something has a definite onset and offset, then quantised labelling allows them to be located. If it actually shifts gradually from a negligible role to a central one, then it is artificial to specify a sharp cut-off.

It seems likely that in the long run, it will be normal to use a mixed diet of representations. Establishing which to use for which purpose depends on accumulating experience.

2.2.2 Representing Consensus and Divergence

It has been argued that divergence between perceivers is an intrinsic feature of emotionality, not noise to be eliminated. It is not immediately obvious how divergence should be dealt with. Averaging is an obvious option. It can be done even with categorical data by constructing 'soft vectors', which associate each of the relevant labels with a number indicating how often it was used.

There are two problems with averaging, though. Firstly, data are quite likely to be bimodal some raters think the person rated is in state A, most others that he/she is in state X, but nobody thinks he/she is midway between the two – and yet the midpoint is what the average describes. Secondly, averaging masks sequential effects. These would occur if raters who thought the person began in state A thought he/she moved to state B, raters who thought the person began in state X thought he/she moved to state Y, but nobody thought that the person began in state A and moved to state Y. Thirdly, averaging masks relationships between different types of labels – raters who thought the person was looking at M thought his/her emotion was A, raters who thought the person was looking at N thought his/her emotion was B.

At present, there is no obvious way to deal with these issues except to include each individual rater's data. In the long run it should be possible to pick out overlapping trends and associations statistically, but it is not straightforward.

2.2.3 Representing Connections

As the opening chapter of this handbook pointed out, emotion is richly connected to people, things, and events, present, past, and foreseen. One of the most difficult problems in the area is how to integrate those connections into a labelling.

It is quite typical that a person will feel positive towards collaborators in a situation, negative towards rivals, doubtful about him/herself, daunted by the task ahead, and glad to have overcome a previous obstacle; and that the balance among these feelings will shift as the task proceeds, and new elements become salient.

Not all ways of representing emotional state demand that connections like that be addressed, but some are very unsatisfying unless they are. Schemes involving appraisal are an example. Various appraisal concepts apply to various elements of the kind of scenario that has been described (intrinsically positive, goal, goal obstructive, etc.). Declaring relevant objects in a global labelling is a partial solution, but it is far from ideal.

One kind of solution that one might imagine is to embed descriptions of relevant objects and events with emotion labels (Devillers and Martin, 2008). A more radical approach is an annotation where timelines describe emotionally significant niches (for instance, a strongly valenced object, an emotionally significant goal, a goal obstructive object, etc.); and entries specify what if anything occupies a given niche at any given time. The result would be a description of the person's affective world, so to speak, rather than his or her emotion as such. The main point here is that dealing with these connections satisfyingly is the kind of problem that may still precipitate quite basic shifts in the way labelling schemes are structured.

2.3 Coda: How Difficult Does Labelling Need to Be?

The range of issues covered in the theoretical part of this chapter may seem quite daunting. They are not exercises in pure intellectual gymnastics, though. Almost all of them originate in a confusion or a disagreement within HUMAINE, which eventually turned out to rest on a theoretical issue that had not been articulated.

The point of articulating the theoretical issues is not to make people address them all. It is to let them make informed decisions about practical decisions, including decisions about the kinds of simplification that make sense and to understand that people with different goals may rationally make different decisions.

3 Specific Resources

This part of the chapter concentrates on specific resources, giving basic technical information, identifying sources that give more information, and commenting where the information is available on functional characteristics (such as procedural issues and reliability).

3.1 Resources for Labelling Emotion

This section first outlines different types of descriptive resource separately and then considers issues that cut across them, such as methods of validation and comparisons between the techniques. For reasons of presentation, it is convenient to begin with description based on dimensional concepts.

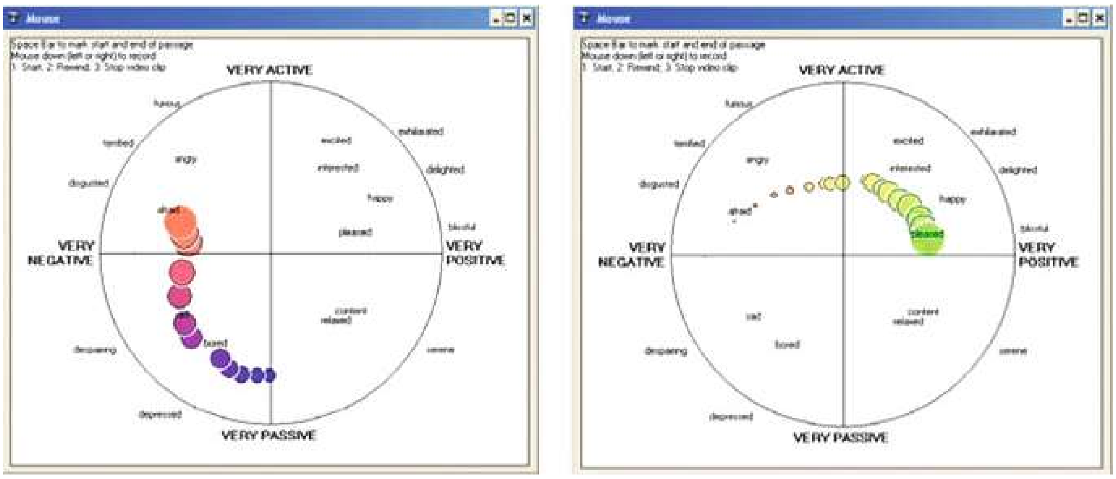
3.1.1 Affect Dimensions

The ideas behind using affect dimensions were introduced above. There are many ways of translating the concepts into labelling schemes.

The FEELtrace technique (Cowie et al., 2000) reflects dimensional concepts very directly. A rater using FEELtrace watches a computer screen with the clip playing on one side and a circle on the other. The axes of the circle correspond to the standard ‘core affect’ dimensions, valence (horizontal) and activation (vertical). Ratings are made by moving a cursor within the circle, and its colour changes with position in a way that people find easy to relate to emotion (red for pure negative, green for pure positive, yellow for maximum activation, dark blue for minimum activation). Selected words are shown within the circle to indicate the kinds of emotion that occupy particular positions in the space. Figure 1 shows snapshots of screens, in one of which the cursor has been moved from deep passivity to fear, in the other from fear to pleased. The output of a tracing session is a file containing three columns. In each row, the first figure specifies time, and the second and third specify the valence and the activation level, respectively, that the rater attributed to the person P being rated at that time. These correspond to the *x* and *y* co-ordinates, respectively, of the cursor at that time. A sequence of valence or activation measures traces the perceived flow of P’s affect over time.

The technique depends on making the demands of the task as low as possible so that users can concentrate on watching the clip. That is achieved by making the display as intuitive as possible and by giving raters a thorough training session. When those precautions are taken, ratings with the trace system show higher reliability than ratings using words (Savvidou, submitted). Reliability can be quite low if those precautions are not taken.

FEELtrace is attractive because it produces in real time what some theories suggest is a full description of affective content. It may provide all the information that is needed for some applications, particularly applications involving voice, where


Space Bar to mark start and end of passage
Mouse down (left or right) to record
1: Start, 2: Rewind, 3: Stop video clip

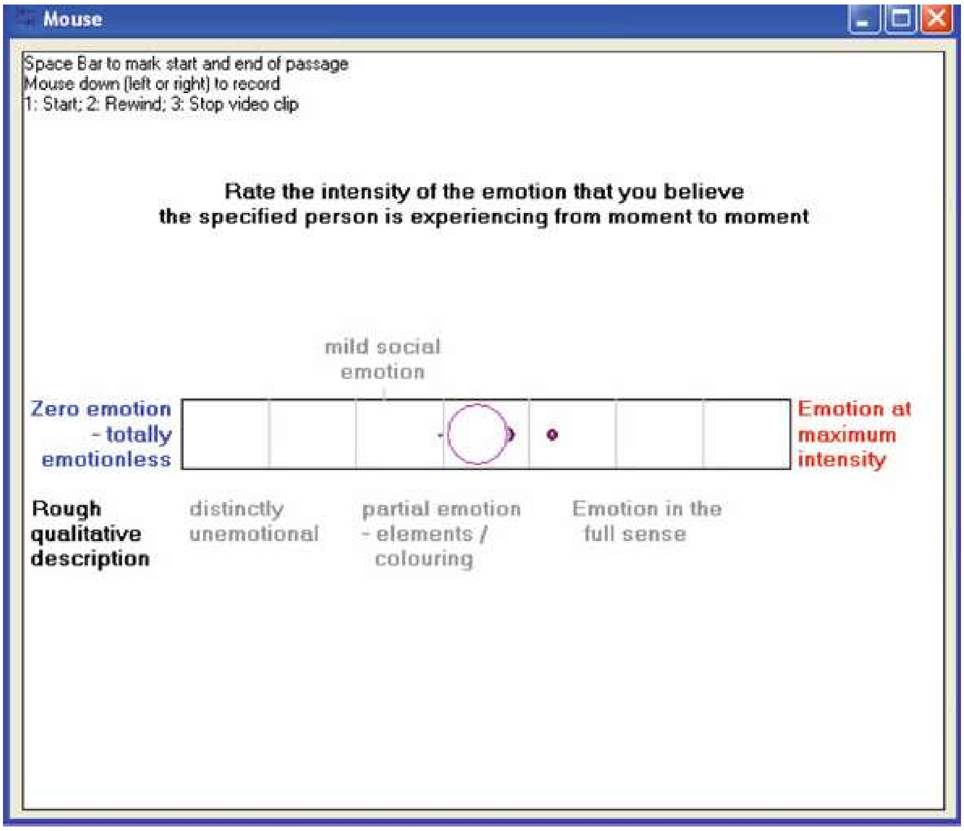
Space Bar to mark start and end of passage
Mouse down (left or right) to record
1: Start, 2: Rewind, 3: Stop video clip

Fig. 1 Examples of FEELtrace screens during tracing. The largest coloured circle shows the current position of the cursor: *smaller circles* mark previous positions

activation is the main kind of information available (Bachorowski, 1999). However, richer information is clearly needed for some purposes. The trace approach has been adapted to provide that by creating a family of one-dimensional traces, each dealing with a particular dimension of emotionality. As with FEELtrace, their reliability depends on training and displays that are carefully constructed to convey what it means to have the cursor in a particular position. Figure 2 illustrates the kind of display that has been used; the program in question is for rating the intensity of the emotion apparently being experienced by the person being rated.

The HUMAINE database has used a family of one-dimensional trace programs to map perceived affect, reflecting the dimensions reported in the 'grid study' (see the first chapter of this handbook). In addition to intensity, illustrated above, they cover valence and activation (separately, not in the single display used by FEELtrace), power/powerlessness, and predictability/unpredictability.

There is no necessary connection between trace techniques and labellings concerned with affect. The 'Self-assessment Mannikin' (SAM) is a standard paper tool used to give discrete ratings (on three seven-point scales) for the affect dimensions of valence, activation, and power (Bradley and Lang, 1994). It could be used to rate a single frame or a relatively homogeneous episode, though in practice labellers appear not to have used it. A related tool, which simplifies the classical dimensions to a few scale points, was developed specifically for labelling by Craggs and Wood (2004).



Space Bar to mark start and end of passage
Mouse down (left or right) to record
1: Start; 2: Rewind; 3: Stop video clip

Rate the intensity of the emotion that you believe the specified person is experiencing from moment to moment

mild social emotion
Zero emotion - totally emotionless
Emotion at maximum intensity
Rough qualitative description
distinctly unemotional
partial emotion - elements / colouring
Emotion in the full sense

Fig. 2 Screen from one-dimensional trace program to measure intensity of emotion (INTENStrace)

3.1.2 Specific Emotion Words

The most obvious approach to labelling is to use words like ‘anger’, ‘happiness’, and so on. They are called ‘specific’ here because it is natural to think of the states they describe as analogous to species in biology (‘lions’, ‘horses’, etc.).

The list of specific emotion-related words is vast. The web is a very useful resource in that context, since it gives lists that reflect contemporary usage rather than academic theory. Sites that give reasonably well-considered lists of words or stock phrases used to describe feelings or emotions include the following:

http://www.angelfire.com/in/awareness/feelinglist.html
http://www.searchingwithin.com/journal/abptb/feel.html
http://lightisreal.com/positiveemotionlist.html
http://en2.wikipedia.org/wiki/List_of_emotions
http://www.umpi.maine.edu/~petress/feelinga.pdf
http://www.psychpage.com/learning/library/assess/feelings.html
http://eqi.org/fw.htm
http://www.preciousheart.net/empathy/Feeling-Words.htm
http://marriage.about.com/library/blfeelingwords.htm

Between them, these sources include nearly 3,000 words or standard phrases that the authors regard as describing feelings or emotions. Of those, 280 occur in four sources or more. Table 1 lists them.

That kind of exercise highlights one of the main problems with labelling based on unconstrained use of terms from everyday language, that is, the sheer number of categories that are relatively common. If raters are left to choose labels freely, and the material is at all complex, they are likely to use a large number of labels, each applying to a small amount of material, and that kind of result is difficult to use directly.

A response that is often used is to allow an unrestricted set of labels in the first instance and then to group them into ‘cover classes’. Examples of that approach are in Douglas-Cowie et al. (2005), Batliner et al. (2006), and Auberge et al. (2006). It might seem that reducing a free labelling to a few cover classes was not so much part of the database as a device used later to reduce the information in a database to a form that suited (for instance) a recognition algorithm. However, grouping may also be necessary to answer questions about the database itself, in particular, questions about the reliability of the labelling.

Cover classes are usually formed ad hoc in the context of a particular database. There are various reasons to be concerned about that. For example, it makes comparability across databases difficult, and it will inflate measures of reliability if terms are assigned to the same class because different raters apply them to the same material. Scherer’s group has developed a tool, based on Scherer’s theoretical framework, which reduces a wide range of terms to 36 broader classes (http://www.unige.ch/fapse/emotion/resmaterial/GALC.xls). Even there, though, it remains a concern how naturally everyday terms do divide into cover classes.

Table 1 Everyday words that Web sites regularly list as descriptors of emotion

Abandoned	Blissful	Cowardly	Energised	**Happy**	**Joyous**	Positive	Sorrowful
Accepted	Bold	Cross	Engaged	Hateful	Jubilant	Powerful	Sorry
Accused	**Bored**	Crossed	Engrossed	Haunted	Jumpy	**Powerless**	Spirited
Adequate	Brash	Cruel	Enraged	Heartbroken	Keen	Pressured	Spiteful
Admired	Brave	Crushed	Enthusiastic	Helpful	Kind	Productive	Splendid
Adored	Breathless	Curious	**Envious**	**Helpless**	Liberated	**Proud**	Squashed
Affectionate	Bright	Daring	Evasive	Hesitant	Lifeless	Provoked	Strong
Afflicted	Bruised	Deceived	Exasperated	Honoured	Light	Puzzled	Stubborn
Afraid	Burdened	Defeated	**Excited**	**Hopeful**	Lonely	Quiet	Stunned
Aggressive	**Calm**	Dejected	Excluded	Hopeless	Lost	Reassured	Suffering
Agitated	Capable	**Delighted**	Exhausted	Horrified	Loved	Rebellious	Sulky
Agonised	Captivated	Depressed	Exhilarated	Hostile	**Loving**	Receptive	Sullen
Agreeable	Carefree	Deserted	Exploited	Humble	Low	Refreshed	Supported
Alarmed	Caring	**Despair**	Exuberant	Humiliated	Lucky	Regretful	Sure
Alert	Carried away	Desperate	Fantastic	**Hurt**	Mean	Rejected	**Surprised**
Alienated	Cautious	Determined	Fascinated	Hysterical	Melancholy	**Relaxed**	Suspicious
Alive	Certain	Devoted	Fearful	Ignored	Merry	Released	Sympathetic
Alone	Challenged	**Disappointed**	Festive	Important	Mischievous	**Relieved**	Tenacious
Aloof	Charmed	Discouraged	Fidgety	Impulsive	Miserable	Remorse	Tender
Amazed	Cheated	Disgraced	Firm	Inadequate	Nervous	Renewed	**Tense**
Amused	Cheerful	**Disgusted**	Flustered	Incapable	Odd	Resentful	Terrible
Angry	Cherished	Dismayed	Foolish	Indecisive	Offended	Reserved	Terrified
Annoyed	Clean	Distant	Fortunate	Independent	Open	**Sad**	Thankful
Anxious	Clever	Distracted	Free	Indifferent	Optimistic	Safe	Threatened
Apathetic	Close	Distraught	**Friendly**	Indignant	Outraged	**Satisfied**	Thrilled
Apologetic	Cold	Distressed	Frightened	Inept	Overjoyed	Scared	Timid
Appalled	Comfortable	Disturbed	Frisky	Infuriated	Overwhelmed	Secure	Tormented
Apprehensive	Comforted	Dominated	**Frustrated**	Innocent	Pained	Sensitive	Torn
Ashamed	Compassionate	**Doubtful**	Fulfilled	Inquisitive	Pampered	**Serene**	Tortured

Table 1 (continued)

Assertive	Competitive	Dull	Furious	Insecure	Panicky	Settled	Trapped
Attractive	Complacent	Dynamic	Generous	Inspired	Paralyzed	Sexy	Uneasy
Aware	Complete	Eager	Gentle	Insulted	Paranoid	Shaky	Upset
Awed	Concerned	Ecstatic	Glad	Intense	Passionate	**Shamed**	Useless
Awkward	Confident	Edgy	Gloomy	**Interested**	Peaceful	**Shocked**	Warm
Bad	Confused	**Elated**	Good	Intimidated	Peeved	Shy	Weak
Beautiful	Considerate	**Embarrassed**	Gratified	Intrigued	Perplexed	Silly	Weary
Belittled	Conspicuous	Empty	Great	**Irritated**	Pessimistic	Sceptical	Wonderful
Betrayed	**Contempt**	Enchanted	Grieving	Isolated	Petrified	Smothered	**Worried**
Bewildered	**Content**	Encouraged	Guarded	Jealous	Playful	Soothed	Zealous
Bitter	**Courageous**	Energetic	**Guilty**	Jolly	**Pleased**	Sore	

Term A may belong with term B in one respect, and term C in another; and terms D and E may be very close in meaning and yet fall on opposite sides of the most natural dividing line. Forming cover classes may still be the most practical procedure in particular cases, but the difficulties should not be underestimated.

The alternative is to develop lists which are designed to offer raters a range of alternatives that are wide enough to be useful, but not totally unmanageable. Motivation for selection varies. A list chosen on theoretical grounds is given as an Appendix in Scherer et al. (1988a) and can be retrieved from the web. It is unique in that (approximately) equivalent words are given in English, German, French, Italian, and Spanish. Teams in HUMAINE concerned with labelling developed a list designed mainly to cover states observed in naturalistic data (Cowie et al., 1999; Banziger et al., 2005; Devillers et al., 2006). It includes the terms shown in bold in Table 1, plus four others: stress, politeness, empathy, and trust.

At the other extreme is the ‘big six’ – fear, anger, happiness, sadness, disgust, and contempt. Comparison with Table 1 suggests why it is not ideal for all purposes. Some of the terms can be regarded as approximations to one of the ‘big six’, but many cannot.

It is natural to use specific emotion terms in a quantised labelling, by associating a given label with a period when it applies. However, that is not the only option. Trace techniques can be applied as easily to words as to dimensional concepts. What they set out to show is how strongly a particular emotion is being felt at any given time. Figure 3 shows how traces (based on the kind of 1D scheme shown in Fig. 2) can reflect the way different named emotions come to the fore and recede in a complex episode, involving anticipation, a surprise, a reunion, and a shared reminiscence.

3.1.3 Generic Emotion Labels

Emotion-related states can also be divided at a different level, as explained in the opening chapter of this handbook. This involves distinguishing, for instance, brief episodes of emergent emotion from moods or emotions that are part of a person’s long-term make-up (such as enduring shame or anger over something that happened long ago). The practical point of labelling at that level is that an automatic system may need to deal differently with different types. Anger at an immediate event (say an unhelpful helpline) calls for a different response from an ongoing bad mood or anger rooted in a long-standing grievance. In at least some cases the signs are clearly different – for example, a system trained on the signs of full-blown anger is very likely to misinterpret the signs of suppressed anger.

Studies in HUMAINE have produced a list of generic types which appear to cover most of emotional life. It is explained in the first chapter of this handbook. A simplified summary is given below (Table 2).

A version of that system has been applied to the clips in the HUMAINE database.

A related development based on a different approach has been exploited by El Kaliouby and Robinson (2004). They used a taxonomy developed by Baron-Cohen (2007). Its distinctive feature is the prominence that it gives to states which are not emotion in a strong sense. Baron-Cohen describes them as epistemic mental

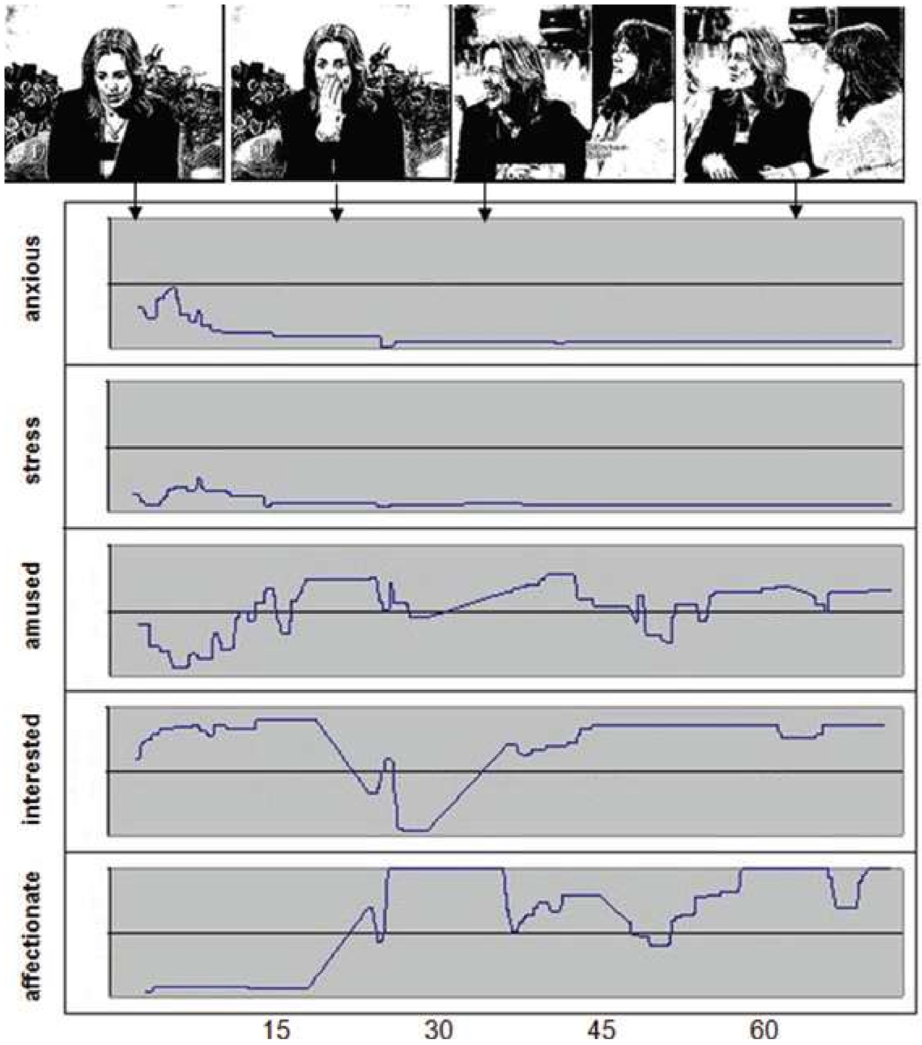


Fig. 3 Traces showing how named emotions are judged to fluctuate through a surprise meeting. Pictures mark key moments: they are stylised to protect identity

Table 2 Generic types of emotion

Emotion-like		Established emotion
		Emergent emotion (suppressed)
		Emergent emotion (full-blown)
	Mood-like	Transitional emotion (shifting between mood and emergent)
		Mood
		Altered state of arousal
		Altered state of control
Stance-like		Altered state of seriousness
		Stance towards object/situation
		Interpersonal stance
		Interpersonal bond
		Emotionless

Table 3 Baron-Cohen (2007) taxonomy of epistemic mental states with an emotional dimension

Group	Class	Concepts included
Sure	Agreeing	Assertive, committed, convinced, knowing, persuaded, sure
Unsure	Unsure	Baffled, confused, puzzled, undecided, unsure
Interested	Concentrating	Absorbed, concentrating, vigilant
	Interested	Asking, curious, fascinated, impressed, interested
Unfriendly	Disagreement	Contradictory, disapproving, discouraging, disinclined
Thinking	Thinking	Brooding, choosing, fantasising, judging, thinking, thoughtful

states with an emotional dimension. El Kaliouby and Robinson pointed out that these states are highly relevant to potential applications such as teaching, and they developed techniques for recognising them. Table 3 shows the categories that they considered. The class level terms in particular appear to be a useful addition to the range of labels worth considering.

Using everyday categories effectively, at various levels, is one of the major challenges for labelling. Good strategies need to be worked out empirically. That depends on setting out the options and testing them in the context of a suitable body of data. That is, the process that the last two sections have tried to facilitate.

3.1.4 Combination Types

It is a feature of naturalistic data that emotion is likely to occur in various kinds of combination rather than as a 'pure' single emotion (Douglas-Cowie et al., 2003; 2005; Devillers et al., 2006). That has led to two types of development, both described by Devillers et al. (2006).

To describe the kinds of combination that occur in a clip, the following set of labels has been developed:

- Unmixed
- Simultaneous combination (distinct emotions present at the same time)
- Sequential combination (single episode which moves through a sequence of related emotions)

To describe the components of a complex, raters should enter more than one emotion term at a time (sad and angry, for instance). 'Soft vectors' (Batliner et al., 2006) can then be used to reflect the strengths of the various components. That kind of information falls out automatically if trace-type descriptions are used to describe the time course of each relevant type.

3.1.5 Appraisal Categories

Appraisal categories aim to reflect the way the person being considered weighs up emotionally critical events or people around. The labelling in one of the earliest naturalistic databases, the Leeds-Reading database, used the appraisal theory due to

Ortony Clore and Collins (1988); it described emotions in terms of an OCC class and the associated object. Research within HUMAINE began by trying to apply the version due to Scherer (see. e.g. Sander et al., 2005), but pilot work showed that relatively few of the descriptors could be assigned with any degree of reliability (Devillers et al., 2006). The labels that could be assigned reliably were as follows:

- Goal conduciveness (the situation offers the person an opportunity to achieve a significant goal)
- Goal obstructiveness (the situation presents an obstacle to the person achieving a significant goal)
- Power/powerlessness (the extent to which the person feels he/she has the power to affect or control emotionally significant events)
- Expectedness (the extent to which the person anticipated emotionally significant events or was taken unawares by them).

Disagreement on the other appraisal categories was partly due to the presence of multiple emotion-related events. That highlights the need to specify events, things, people, and so on that are material to the emotion. Thinking about archetypal cases like James' bear in the woods conceals the fact that events (etc.) may have several different roles in relation to a single emotional episode. The system used by Devillers et al. distinguished four. An example helps to distinguish them. I hear in the morning that a major grant application has been rejected. Walking to work I trip on a loose paving stone and become very angry with the council. The categories involved are as follows:

Cause. The term is not ideal, but it is hard to do better. It refers to an event that may precede the particular emotion by some time, without which it would not have happened. In the example, it is the rejection.

Aspiration. This is relatively self-explanatory. In the example, the whole scenario depends on my aspiration to get the grant.

Trigger. This is an event which immediately precipitates an emotion – in the example, tripping.

Focus. This is what an emotion is about – in the example, the council. It is also called the object of the emotion.

The scheme is not ideal, but it provides some advance over struggling to say what an emotion is about when several factors of different kinds are obviously relevant.

3.1.6 Authenticity

It is natural to assume that authenticity is a single issue. However, inspection makes it clear that there are at least two issues to consider. The first is whether or not the subject appears to be simulating an emotion that he or she does not feel. The second is whether the subject in question is masking his/her emotion. Pilot work on television data (particularly on ‘reality shows’ where subjects are facing major challenges in competitive situations) suggests that masking of emotion is common.

These are addressed using two trace techniques in the Belfast database. In one, coders are asked to rate the data on a scale from ‘no acting of emotion’ to ‘extreme acting of emotion’. In the other, they mark the level of masking on a scale from ‘no concealment of emotion’ to ‘total concealment of emotion’.

3.1.7 Reliability and Variation

It is impossible to give a deep treatment of these issues in a short section, so the aim here is simply to point out that there is a variety of issues to be faced and to note some known solutions.

The most straightforward question is how consistently a particular piece of material has been labelled. Where the material is divided into segments that are meant to be relatively homogeneous, and each labeller assigns each segment one label chosen from a relatively small set, the natural measure is the proportion of raters who agree on the commonest label. For pure dimensional labellings, a comparable kind of information is provided by the standard deviations of ratings (Cowie et al., 2000). The two can be compared by finding dimensional co-ordinates for each label (Whissell, 1989 provides an extensive list) and using those to derive standard deviations (Cowie and Cornelius, 2003). The same kind of replacement can be used to give a comparable measure of agreement in cases where there are many categories to choose from, some of which are closer together than others.

At more global level, there are standard techniques for measuring how reliably different raters assign a particular set of labels. With a set of qualitative labels, the standard statistic is kappa (Cohen, 1960). Where ratings are continuous (as with trace measures), the standard statistic is Cronbach’s alpha. To illustrate expected values, Devillers et al. (2006) report average kappa of just above 0.6 when raters are assigning a set of 15 everyday labels, and Savvidou (submitted) reports alpha values above 0.9 for both dimensions (valence and activation) in a series of studies with FEELtrace.

Note, though, that simple application of these measures is not always appropriate, and techniques have been devised to deal with the issue. The usual form of kappa is not appropriate when it is possible to assign more than one label (for instance, when there is a blend of sadness and anger). Rosenberg and Binkowski (2005) have proposed an extension that handled multi-element responses. Their approach still gives low values when raters use multiple labels even if they agree perfectly, though. Devillers et al. (2006) further adapted Rosenberg and Binkowski’s approach to give a kappa measure whose value is (as expected) 1 when raters choose the same multiple labels.

Alpha is an adjusted average of the correlations between pairs of outputs. Like other averages, it can be misleading if the distribution is not straightforward. An immediate response to that kind of issue is to plot correlations between all pairs of

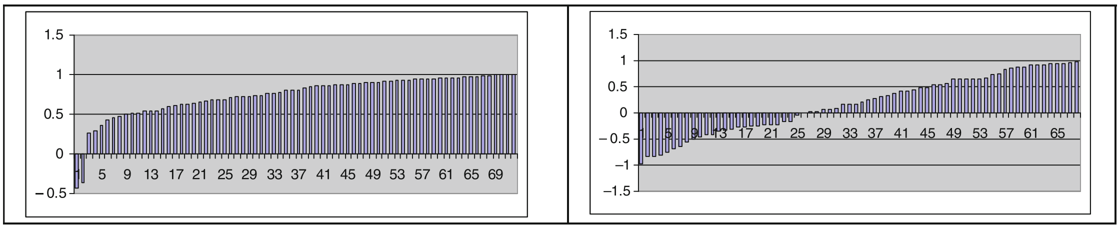


Fig. 4 Plots of the correlations between traces made by multiple raters (from Devillers et al., 2006). Each bar corresponds to the correlation coefficient obtained by comparing two traces of the same passage. The *left-hand panel* summarises ratings of intensity and the *right panel* summarise ratings of power

raters. Figure 4 shows the kinds of picture that emerge. They illustrate an intriguing and quite common pattern, which is that in some cases, raters are not simply uncorrelated; they take positively opposing views of what is happening.

That kind of problem is classically recognised, and the classical prescription is to carry out factor analysis as a way of separating the different patterns that are overlaid in the data set as a whole (Tabachnik and Fidell, 2001). Figure 5 shows patterns inferred from trace data in that way (Savvidou, in preparation). The procedure is not wholly satisfying, but it provides at least some protection against concluding that a measure is unreliable because it reveals that people read the material in different ways.

The ability to represent variation is particularly important in evaluating an artificial recogniser. One of the key tests of a good recogniser is how the distribution of responses that it gives compares to the range given by human raters. The standard measure, percentage correct, is not a satisfying way to do that; it may not be clear what ‘correct’ is, and there may be several responses that are quite commonly given even if they are not correct, or even the commonest choice for humans. Steidl et al. (2005) have developed an entropy measure that addresses these issues.

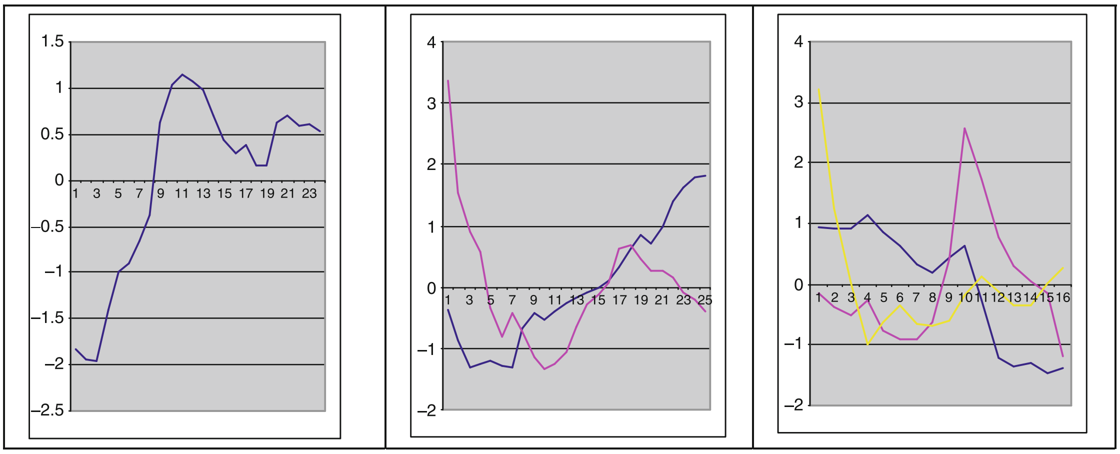


Fig. 5 Factors in ratings of valence for clips showing one, two, and three readings of the material, respectively

3.2 Resources for Labelling Signs of Emotion

Many of the issues related to sign labelling are dealt with in depth elsewhere in the handbook. Broadly speaking, two main areas are involved. On the one hand, there are the signal processing techniques covered in Part II ‘Signals to Signs’; on the other, there are constructs that are used to synthesise signs, which are covered in Part IV ‘Emotion in Interaction’. Hence, the aim in this section is generally to give a broad overview rather than technical detail.

It would be quite wrong if the labelling of signs in databases was divorced from these other areas, because databases are a resource for teams concerned with recognition and synthesis. For teams concerned with recognition, databases provide the material that they need to evaluate recognition algorithms. To be practically useful, the database has to provide the algorithms with input variables that are machine recoverable (or could be recovered given foreseeable improvements in signal processing). For teams concerned with synthesis, the database has to indicate the kinds of gesture (in the broadest sense) that are relevant to conveying a particular kind of emotion in a particular context, and that only useful, if the description of the gesture corresponds to something that can be generated.

On the other hand, there are types of labelling that may be significant, but that cannot at present be either recovered or generated by machine. Those receive particular attention here, because they are not covered elsewhere.

Descriptors are divided into five main areas: linguistic; vocal (including paralinguistic); facial; gesture; and physiological.

3.2.1 Linguistic

Orthographic transcription is a key resource, most obviously because the content of what people say usually conveys a great deal about their emotions. Even crude words spotting adds materially to emotion recognition (Fragopanagos and Taylor, 2005). Time alignment gives added value. It is not obvious how much phonetic transcription adds. Some of the information that it provides can be derived in other ways (for instance, duration of vowels can be derived using detection of voicing). However, indicators like reduction in the vowel space are not easy to access automatically.

Linguistic descriptions of intonation (such as ToBI) have a similar status. Some prosodic signs of emotion can depend on the way observed patterns of pitch and stress relate to underlying linguistic patterns (Ladd et al., 1985; Mozziconacci, 1998), but it is not clear how often the effect matters.

Discourse-related units may also be important. For instance, Craggs and Wood (2003) argue that grounding behaviour is more protracted in discussions about a subject about which they feel emotional. They consider possible schemes for emotion-relevant annotation of dialogue features. Campbell (2004) describes a well-developed coding scheme for dialogue acts that has been used in the ESP corpus.

Several sources suggest that linguistic features are not overwhelmingly important for coarse judgments of emotion. There is a shortage of good data, but it seems clear that people have considerable ability to gauge the emotions of people speaking unfamiliar languages. There is also relatively little shift in judgments when speech is filtered so that prosody remains, but individual words are unintelligible (Douglas-Cowie et al., 2005).

On the other hand, there are emotionally charged types of communication that it is difficult to imagine grasping without language, such as irony. The function of the database, and the nature of the material, probably dictates the value of including sophisticated linguistic descriptions.

3.2.2 Vocal

No database can reasonably aspire to include all the acoustic measures that have a claim to be relevant to detecting emotion. As the chapter by Batliner et al. on recognition from audio signals (in Part II 'Signals to Signs') stands to indicate, there are simply too many. The best that can be hoped for is to include raw descriptions from which the most important measures can be derived.

There is a strong case for databases to include F0, particularly corrected F0. Uncorrected F0 contours are notoriously prone to octave jumps, particularly when speech is emotional (hence the HUMAINE Web site contains a tool for correcting F0). Corrected pause boundaries are useful for a similar reason. Intensity may well be as important, but if distance to the microphone is fixed, it is trivially easy to recover and if it is not, it is almost impossible to achieve much precision.

Recognition algorithms also tend to use coefficients associated with the presence or the absence of voicing, the spectrum, formants, and MFCCs. It is not obvious how much would be gained by including these in a database.

At a higher level, Douglas-Cowie derived a set of descriptors by listening to the Belfast naturalistic database (Douglas-Cowie et al., 2003). The original system contained many labels, but for use in the HUMAINE database, these have been reduced to a core set of items which the tests indicate are strongly characteristic of emotion and can be applied reliably.

The labels address four descriptive categories and raters can assign a number of labels within these levels:

- Paralanguage
  - Laughter, sobbing, break in voice, tremulous voice, gasp, sigh, exhalation, scream
- Voice quality
  - Creak, whisper, breathy, tension, laxness
- Timing
  - Disruptive pausing, too long pauses, too frequent pauses, short pause + juncture, slow rate
- Volume
  - Raised volume, too soft, excessive stressing

These have been used in the HUMAINE database (described later in this handbook).

3.2.3 Face Descriptors

Face is discussed in several other chapters of this handbook. It is an area where there is strong convergence. The facial action coding system (FACS) is generally accepted as a standard (Ekman et al., 2002). It underpins the system of coding in terms of facial action packages (FAPs) adopted by MPEG, which is used extensively in recognition, generation, and databases. Fuller's descriptions are given in Ioannou et al. (2005) and the chapters on image and video processing and multimodal emotion recognition are given in Part II 'Signals to Signs'.

3.2.4 Gesture Descriptors

Gesture is also discussed in several other chapters of this handbook. In contrast to the work on faces, several very different schemes lie in the background, which have been developed for different purposes – distinguishing types of communicative device, annotating dance, transcribing sign language. There is momentum to develop a satisfying composite scheme.

An example of recent developments is a scheme developed at LIMSI-CNRS. It is a manual annotation scheme and has been used for annotating multimodal behaviours in two corpora: EmoTV (Martin et al., 2005) and EmoTABOO (Martin et al., 2006). The scheme is currently being used for studying the relations between multimodal behaviours and blends of emotions (Devillers et al., 2005).

The scheme uses the following dimensions to annotate gesture:

- Classical dimensions of gesture annotation (McNeill, 1992; Kita et al., 1998; Kipp, 2004; McNeill, 2005), to allow exploratory study of the impact of emotion on these dimensions:
- Gesture units (e.g. in order to study how much gesture there is in an emotional corpus)
- Phases (e.g. to study if there are long duration of holds due to the interactive game and thinking behaviours of the subject)
- Phrases
- Categories (e.g. to study the frequency of adaptors and compare it to other corpora)
- Lemmas adapted from a gesture lexicon (Kipp, 2004) (e.g. Doubt=Shrug)
- Expressivity: annotated at the level of the phrases and at the level of the gesture units. The scheme considers six expressivity parameters used in research on movement quality (Hartmann et al., 2005; Wallbott, 1998): activation, repetition, spatial extent, speed, strength, fluidity.

This scheme is implemented in a module in the Anvil tool (Kipp, 2001). The chapter on the HUMAINE database (in Part III) shows it applied to selected clips. Improvements are under active discussion.

3.2.5 Physiological Descriptors

It makes sense for databases to include the relatively small core of physiological descriptors that are known to vary with emotion. They are as follows:

- Heart rate
- Blood pressure
- Skin conductance
- Respiratory effort
- Skin temperature
- EMG at key sites

There are many standard ways of deriving measures from these basic signals, involving differences, standard deviations, filtering, and various other operations. Code that can be used to carry out standard transformations can be accessed via the portal from the Augsburg Biosignal Toolbox (AuBT), which was developed for HUMAINE (see chapter 'Multimodal Emotion Recognition from Low-Level Cues' in Part II 'Signals to Signs')

The major problem with these measures is that they are sensitive to many types of variables which are not directly related to emotion (including physical and mental effort, speech, and health-related factors). To control for these statistically, they need to be recorded.

3.3 Resources for Labelling Context

An enormous range of issues could be covered under the heading 'context'. The description here is mainly drawn from the way the HUMAINE database has attempted to systematise the issues.

At a basic level, there is factual data on the subject's personal characteristics (age, gender, race), on technical aspects of recording (recording style, acoustic quality, video quality), and on physical setting (degree of physical restriction, posture constriction, hand constriction, and position of audience).

A second set of labels deals with communicative context. The elements addressed in the HUMAINE database fall into two categories:

(i) Coders are asked to rate the purpose or the goal of the communication – to persuade, to create rapport, to destroy rapport, or just a pure expression of emotion, for example, somebody laughing.

(ii) Coders record their perceptions of the social setting of the clip, whether, for example, there is a clear interaction happening between two or more people or whether one person is communicating to another who has a very passive role in the interaction. Social pressure is considered under this heading. Coders rate whether they think that the situation puts the person they are rating under pressure to be formal (e.g. in a court), or to be freely expressive (e.g. at a party), or that there is little pressure either way.

Table 4 Classification of antecedents to happiness from Scherer (1988)

Good news (immediate social context). Example: an unexpected job offer

Good news (mass media). Example: cheering news in newspapers or on TV

Continuing relationships with friends and permanent partners. Example: pleasure from contact with friends

Continuing relationships with blood relatives and in-laws

Identification with groups (actual and reference). Examples: pleasure in belonging to a club; returning to your own country after a holiday

Meeting friends, animals, plants. Examples: seeing one's dog again; meeting one's friend for dinner

Meeting blood relatives or in-laws

Acquiring new friends

Acquiring new family members. Examples: birth of a baby; marriage of one's brother

Pleasure in meeting strangers (short-term chance encounters). Example: talking to a stranger on a train

Pleasure in solitude. Example: being left alone with one's own thoughts

New experiences. Examples: adventures; planning a holiday

Success experiences in achievement situations. Example: passing an examination

Acquiring some material for self or other (buying or receiving). Examples: presents from others; buying something nice for oneself or others

Ritual. Examples: religious, academic ceremonies, festivals, birthdays

Natural, also refined, noncultural pleasures. Examples: sex, food, nature, landscape

Cultural pleasures. Examples: art, music, ballet, etc.

Acquiring nonmaterial benefits (emotional support, altruism). Example: helping an old lady cross the road

Happiness without reason

Schadenfreude. Example: malicious pleasure in another person's misfortune

Description of appraisal could be counted here. The HUMAINE database includes a very compressed description, which has been covered already in this chapter in Sect. 3.1.3. Much fuller descriptions have been offered in the literature. Table 4 forms part of the coding system reported in Scherer (1988b), classifying the types of antecedent to happiness. Similar schemes are offered for sadness, anger, and fear.

4 Proposals to Advance the State of the Art

The research priority at this stage is reasonably clear. A rich body of ideas about labelling emotion databases has accumulated. The ideas need to be applied systematically to a database that is sufficiently large, and challenging to test what they can contribute, and evaluated. In the first instance, evaluation involves several levels. Reliability (with qualifications noted above) is the first. The second is redundancy: to what extent do different descriptors duplicate the same information? That kind of question has been asked in the context of isolated word (and the results form the basis of dimensional accounts). However, it has not been asked in the context of real instances of emotion. The third is adequacy for synthesis: how well do various sets of parameters allow an ECA to reproduce the emotional content of an original clip? The fourth is predictive power: how well do various combinations of sign and context labels allow emotion labels to be inferred, and (with some differences) vice versa? The task goes beyond reconstructing labels in one stream from labels in another at an adjacent time. It critically involves identifying possibilities to be considered and factored into future choices – which kinds of intervention are advisable given these signs in this context, and which should be avoided at any price?

These questions add up to a research agenda for a number of decades. It depends on the availability of suitable sets of records and cumulative work on applying possible labels to them. The HUMAINE database, described in the next chapter, is a substantial step in that direction.

On the other hand, applied research projects will test how useful subsets of the repertoire are in specific applications. It is an interesting challenge to ensure that applied and academic evaluations connect.

Some people will be frustrated by the implication that genuinely satisfying prescriptions for labelling lie at the end of a long-lasting research program. It is an understandable reaction. However, understanding it does not change the reality that that is the timescale.

References

Auberge V, Audibert N, Rillard A (2006) Auto-annotation: an alternative method to label expressive corpora. In: Proceedings LREC 2006, Genoa, pp 45–46

Bachorowski JA (1999) Vocal expression and perception of emotion. Curr Dir Psychol Sci 8(2): 53–57

Bänziger T, Tran V, Scherer KR (2005) The Geneva Emotion Wheel: a tool for the verbal report of emotional reactions. In: Poster presented at ISRE 2005, Bari

Baron-Cohen S (2007) Mind reading: the interactive guide to emotions – version 1.3. Jessica Kingsley, London

Batliner A, Steidl S, Schuller B, Seppi D, Laskowski K, Vogt T, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2006) Combining efforts for improving automatic classification of emotional user states. In: Erjavec T, Gros J (eds) Language technologies, IS-LTC 2006. Infornacijska Druzba (Information Society), Ljubljana, Slovenia, pp 240–245

Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59

Campbell N (2004) Extra-semantic protocols: input requirements for the synthesis of dialogue speech. In: Andre E, Dybkjaer L, Minker W, Heisterkamp P (eds) Affective dialogue systems. Lecture notes in artificial intelligence. Springer, Berlin, pp 221–228

Cohen JA (1960). A coefficient of agreement for nominal scales. Educ Psychol Meas, 20(1):37–46

Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Commun 40:5–32

Cowie R, Douglas-Cowie E, Apolloni B, Taylor J, Romano A, Fellenz W (1999) What a neural net needs to know about emotion words. In: Mastorakis N (ed) Computational intelligence and applications. World Scientific Engineering Society, Dallas, TX, pp 109–114

Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schroeder M (2000) 'FEELTRACE': an instrument for recording perceived emotion in real time. In: Proceedings of the ISCA ITRW on speech and emotion: developing a conceptual framework, Newcastle, 5–7 Sept 2000, Textflow, Belfast, pp 19–24.

Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor J (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1):32–80

Craggs R, Wood M (2003) Annotating emotion in dialogue. In: Proceedings of the 4th SIGdial workshop on discourse and dialogue, Sapporo

Craggs R, Wood M (2004) A 2 dimensional annotation scheme for emotion in dialogue. In: AAAI spring symposium on exploring attitude and affect in text: theories and applications. Stanford University, AAAI Press, pp 44–49

Devillers L, Abrilian S, Martin J-C (2005) Representing real life emotions in audiovisual data with non basic emotional patterns and context features. In: Proceedings of the 1st international conference on affective computing and intelligent interaction (ACII'2005), Beijing, 22–24 Oct. Spinger, Berlin, pp 519–526

Devillers L, Cowie R, Martin J-C, Douglas-Cowie E, Abrilian S, McRorie M (2006) Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. In: Proceedings LREC 2006, Genoa

Devillers L, Martin J-C (2008) Coding emotional events in audiovisual corpora. In 6th international conference on language resources and evaluation (LREC 2008). Marrakech (Morocco)

Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40(1–2):33–60

Douglas-Cowie E, Devillers L, Martin J-C, Cowie R, Savvidou R, Abrilian S, Cox C (2005) Multimodal databases of everyday emotion: facing up to complexity. In: Proceedings of the Interspeech 2005, Lisbon, pp 813–816

Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200

Ekman P (1999) Basic emotions. In: Dalgleish T, Power MJ (eds) Handbook of cognition and emotion. Wiley, New York, pp 301–320

Ekman P, Friesen W (1975) Unmasking the face: a guide to recognizing emotions from facial clues. Prentice-Hall, Englewood Cliffs, NJ

Ekman P, Friesen WC, Hager JC (2002). Facial action coding system. The manual on CD ROM. Research Nexus Division of Network Information Research Corporation, Salt Lake City, UT

El Kaliouby R, Robinson P (2004) Mind reading machines: automated inference of cognitive mental states from video. In: IEEE international conference on systems, man and cybernetics, vol 1, The Hague, pp 682–688

Fragopanagos N, Taylor J (2005) Emotion recognition in human–computer interaction. Neural Netw 18:389–405

Hall JA, Matsumoto D (2004) Gender differences in judgments of multiple emotions from facial expressions. Emotion 4(2):201–206

Hanson NR (1958) Patterns of discovery: an inquiry into the conceptual foundations of science. Cambridge University Press, Cambridge, MA

Hartmann B, Mancini M, Pelachaud C (2005) Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture Workshop (GW'2005), Vannes, France, pp 1095–1096

Ioannou S, Raouzaiou A, Tzouvaras V, Mailis T, Karpouzis K, Kollias S (2005) Emotion recognition through facial expression analysis based on a neurofuzzy network Neural Netw, 18, 423–435

James W (1884) What is emotion? Mind 9:188–205

Kipp M (2001) Anvil – a generic annotation tool for multimodal dialogue. In: 7th European conference on speech communication and technology (Eurospeech'2001), Aalborg, Denmark, 3–7 Sept, pp 1367–1370

Kipp M (2004). Gesture generation by imitation. From human behavior to computer character animation, Florida, Boca Raton, Dissertation.com. 1581122551

Kita S, van Gijn I, van der Hulst H (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. Gesture and sign language in human computer interaction: Proceedings/international gesture workshop, Bielefeld, 17–19 Sept. Springer, Berlin, Heidelberg

Ladd S, Silverman K, Bergmann G, Scherer K (1985) Evidence for independent function of intonation contour type, voice quality, and F0 in signalling speaker affect. J Acoust Soc Am 78(2):435–444

Martin J-C, Abrilian S, Devillers L (2005). Annotating multimodal behaviors occurring during non basic emotions. In: 1st international conference on affective computing and intelligent interaction (ACII'2005), Beijing, 22–24 Oct. Spinger, Berlin, pp 550–557

Martin JC, Devillers L, Zara A, Maffiolo V, LeChenadec G (2006) The EmoTABOU corpus. Humaine Summer School, Genova, pp 22–28

Matsumoto D (2001) Culture and emotion. In: Matsumoto D (ed) Handbook of culture and psychology. Oxford University Press, New York, NY, pp 171–194

Mayer JD, Salovey P, Caruso D (2000) Models of emotional intelligence. In: Sternberg R (ed) Handbook of intelligence. Cambridge University Press, Cambridge, MA, pp 396–421

McNeill D (1992) Hand and mind – what gestures reveal about thoughts. University of Chicago Press, Chicago, IL

McNeill D (2005) Gesture and thought. University of Chicago Press, Chicago, IL

Mozziconacci SJL (1998). Speech variability and emotion: production and perception. Ph.D. thesis, Eindhoven

Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge, MA

Ortony A, Turner TJ (1990) What's basic about basic emotions? Psychol Rev 97:315–331

Rosenberg A, Binkowski E (2005). Augmenting the kappa statistic to determine interannotator reliability for multiple labelled data points. In: Proceeding of the HLT-NAACL, Boston

Russell J, Barrett-Feldman L (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 5:37–63

Sander D, Grandjean D, Scherer KR (2005) A systems approach to appraisal mechanisms in emotion. Neural Netw 18:317–352

Savvidou S (submitted) Validation of the Feeltrace tool for recording impressions of expressed emotion. Thesis submitted for Ph.D., Queen's University, Belfast.

Scherer KR (ed) (1988a) Facets of emotion: recent research. Erlbaum, Hillsdale, NJ. Appendix F. Labels describing affective states in five major languages, revised by the members of the Geneva Emotion Research Group retrieved Dec 2007 from http://www.unige.ch/cisa/gerg/research.html

Scherer KR (ed) (1988b) Appendix B. Antecedent and reaction codes used in the "Emotion in Social Interaction." In: Scherer KR (ed) Facets of emotion: recent research. Erlbaum, Hillsdale, NJ, pp 241–243. http://www.unige.ch/cisa/gerg/research.html

Schreuder M, van Eerten L, Gilbers D (2006) Music as a method of identifying emotional speech. In: LREC Research Workshop on Corpora on Emotion and Affect, Genoa

Steidl S, Levit M, Batliner A, Nöth E, Niemann H (2005) "Of All Things the Measure is Man" – automatic classification of emotions and inter-labeller consistency. In: Proceedings of the ICASSP 2005, Philadelphia, pp 317–320

Tabachnik BG, Fidell LS (2001) Using multivariate statistics, 4th edn. Allyn & Bacon, Boston, MA

Wallbott HG (1998) Bodily expression of emotion. Eur J Soc Psychol 28:879–896

Whissell C (1989) The dictionary of affect in language. In: Plutchik R, Kellerman H (eds) Emotion: theory, research and experience: vol 4, The measurement of emotions. Academic, New York, NY, pp 113–131