

# HUMAINE

**D5c**

**Preliminary plans for exemplars:  
Databases**

**Ellen Douglas-Cowie and WP5 members**



**Version 1.0**

**Date: 28 May 2004**

<b>IST project contract no.</b>	507422
<b>Project title</b>	<b>HUMAINE Human-Machine Interaction Network on Emotions</b>
<b>Contractual date of delivery</b>	<i>month 5</i>
<b>Actual date of delivery</b>	<i>28 May 2004</i>
<b>Deliverable number</b>	<i>D5c</i>
<b>Deliverable title</b>	Preliminary plans for exemplars: Databases
<b>Type</b>	Public Report
<b>Number of pages</b>	68
<b>WP contributing to the deliverable</b>	WP5
<b>Task responsible</b>	QUB
<b>Author(s)</b>	Ellen Douglas-Cowie and WP5 members
<b>EC Project Officer</b>	Philippe Gelin

Address of author:

School of English, Queen's University Belfast, BT7 1NN, Northern Ireland, UK.

Institutions contributing:

QUB, DFKI, ICCS, Paris8, KTH, GERG, TAU, CNRS, FAU, BARI, TCD

## Table of Contents

<b>1. THE STATUS OF THIS REPORT .....</b>	<b>5</b>
<b>2. THEMATIC DEFINITION OF THE WORKPACKAGE AND THIS REPORT .....</b>	<b>6</b>
<b>A. Databases in the Humaine Context .....</b>	<b>6</b>
<b>B. The core topics.....</b>	<b>7</b>
<b>C. Conclusion .....</b>	<b>9</b>
<b>3. REVIEW OF KEY CONCEPTS IN THE THEMATIC AREA .....</b>	<b>10</b>
<b>A. Emotion.....</b>	<b>10</b>
<b>B. Modalities.....</b>	<b>11</b>
1) Speech .....	11
2) Face.....	11
3) Gesture.....	11
4) Physiological .....	13
5) Brain imaging .....	13
6) Action .....	13
<b>C. MPEG.....</b>	<b>14</b>
<b>D. Emotion elicitation.....</b>	<b>15</b>
1) Acted /posed data .....	15
2) Reading.....	15
3) Naturalistic data.....	15
4) Call center data.....	15
5) Induced data.....	16
<b>E. Episodes in databases .....</b>	<b>16</b>
<b>F. Labelling and format .....</b>	<b>16</b>
1) Encoding of emotion-related content .....	16
2) Encoding of the signs of emotion.....	18
<b>4. REVIEW OF KEY ACHIEVEMENTS IN THE THEMATIC AREA.....</b>	<b>20</b>
<b>A. Data: the state of the art.....</b>	<b>20</b>

---

<b>B. Labelling and Format .....</b>	<b>39</b>
<b>5. REVIEW OF KEY PROBLEMS IN THE THEMATIC AREA.....</b>	<b>41</b>
<b>6. ASSESSMENT OF THE KEY DEVELOPMENT GOALS IN THE THEMATIC AREA .....</b>	<b>43</b>
<b>7. RELATION TO OTHER WORKPACKAGES .....</b>	<b>46</b>
<b>A. WP3 Theories of emotion .....</b>	<b>46</b>
<b>B. WP4: Signals to signs and vice versa.....</b>	<b>48</b>
<b>C. WP6: Interaction.....</b>	<b>49</b>
<b>D. WP10: Ethics .....</b>	<b>49</b>
<b>E. WP 3, 4, 6 Shared task to develop appropriate labellings.....</b>	<b>49</b>
<b>8. PRELIMINARY IDEAS ABOUT AN EXEMPLAR.....</b>	<b>51</b>
<b>9. CONCLUSIONS AND WAY FORWARD.....</b>	<b>57</b>
<b>10. REFERENCES .....</b>	<b>57</b>
<b>A. References cited in the text.....</b>	<b>57</b>
<b>B. General bibliography.....</b>	<b>64</b>

## 1. The status of this report

Joint research in HUMAINE aims to produce 'exemplars'. We chose the term exemplars to convey that above all, the systems that we develop should embody sound principles. The systems may be working models or 'in principle' specifications. Embodying sound principles means not only that they should exemplify good ways of addressing individual problems, but also that the set of exemplars taken as a whole defines a rational ways of partitioning the overall problem of developing emotion-sensitive systems. Arriving at a satisfying partition is a major part of the challenge that HUMAINE faces, requiring iteration and consultation between groups dealing with different thematic areas.

The Technical Annex sets process that is designed to meet that challenge. It will begin with production by each thematic group of a review of key concepts, achievements and problems in its thematic area; and drawn from the review, an assessment of the key development goals in the area. This review and assessment will be circulated to the whole network for discussion and comment, aimed both at building understanding of basic issues across areas, and at identifying the choices of goal that would be most likely to let the different groups achieve complementary developments. That consultation phase will provide the basis for deliverables in month 11, which will specify a range of exemplars that deserve serious consideration. A further round of consultation will follow before concrete plans for each workpackage are drawn up and shared at the 18 month plenary meeting.

This report is the review that defines the starting point of the process for work package WP5, whose brief title is 'Data and databases'.

## 2. Thematic definition of the workpackage and this report

### A. Databases in the Humaine Context

HUMAINE is about the development of multimodal emotion-sensitive interfaces. The aim of WP5 is to define the kind of database that is needed to underpin that development, and to produce exemplars that will serve as concrete models for the collection and labelling of such databases. This first report is focused on assessing what has been achieved in the field, identifying the challenges, and proposing a number of candidate exemplars.

In summary, there has been substantial progress in the field, particularly in the last decade or so, and there are many databases that are relevant in one way or another. Nevertheless, very few genuinely meet the needs of a program like Humaine.

In particular, two fundamental requirements are not adequately addressed in existing databases. These are:

#### (i) Multimodal representation of emotion

Humaine is concerned with both emotion and multimodality, but it is rare to find databases that include both to any degree. Although there is a growing interest in multimodal databases and an increasing number of them, emotion rarely features in them, and although a growing number of databases feature some level of emotion, they are rarely multimodal. Section 4 of the report provides an extended review of the relevant material.

As always there are exceptions. For example the Belfast Naturalistic Database is an audiovisual database of 140 subjects expressing a wide range of emotions (Douglas-Cowie et al. 2000, 2003). The Geneva lost luggage database (based on customer reactions in Geneva airport when their luggage goes missing) is also audiovisual (Scherer & Ceschi 1997, 2000). Other examples include some that record speech and the associated physiology (Amir et al. 2000), and the recent SMARTKOM database ([www.smartkom.org](http://www.smartkom.org), Schiel et al. 2002; Steininger, Schiel & Glesner 2002; Steininger et al. 2002 ). But even these databases are by no means sufficient for Humaine. It is rare to find combinations of more than two modalities at any one time; and some modalities, such as gesture and body language are largely ignored.

#### (ii) Naturalism

Humaine is also committed to dealing with emotions and emotion-related states as they occur in everyday life (see 'Project Objectives' in the Technical Annex).

That entails that the data should be 'naturalistic'. However, many databases that are described as having emotional content are in fact acted or posed (such as the databases of static faces from Yale <http://cvc.yale.edu/projects/yalefaces/yalefaces.html> or of Ekman & Friesen 1978, and the Berlin database of emotional speech Kienast & Sendlmeier, 2000). These have played a key role in developments to date, but there is evidence to show that data of this type does not bear a straightforward relationship to emotion in everyday life. A study by Batliner et al (2003) shows that techniques which work well for recognising acted emotion do not transfer to spontaneous speech. That is not surprising given their auditory distinctness. A recent experiment at Queen's, Belfast studied listeners' ability to discriminate between samples of

spontaneous emotional speech and acted renditions of them. Listeners knew after a few seconds which category a clip belonged to.

The data should also reflect the type of emotion that occurs in everyday life. Existing databases are often oriented to extreme representations of a few emotions (often based on traditional lists of ‘primaries’), and although these do indeed occur in everyday life, much of our life is taken up with less intense emotion (irritation rather than anger, pleasure rather than elation) or emotion-related states (such as friendliness, interest, satisfaction, stress, anxiety). Humaine needs databases that are representative of the range and degree of emotion that takes place around us.

As before, there are exceptions. Recent work on speech and emotion has been particularly alert to problems with acted data, and has explored various methods of obtaining more naturalistic data. Campbell et al. (in Japan), for example, make recordings of people as they go about their everyday lives, and use the term ‘expressive’ speech rather than emotional speech to cover the type and range of emotion related behaviour that occurs in everyday life (Campbell 2002; Douglas-Cowie et al 2003). Call center speech databases are also currently a popular way of trying to obtain more naturalistic data (Ang et al 2002; Devillers et al 2002, 2003, 2004). These and other studies in a similar vein acknowledge that there is a core issue about using acted data, and provide a useful basis for moving forward. But again, they tend to be limited in ways that are problematic for Humaine. They are not multimodal, and extending them to include multimodality is not straightforward. And the call centre data is often very context specific and linguistically constrained; these factors make any kind of generalisation difficult.

## **B. The core topics**

Five topics are fundamental to the report. This section provides a brief introductory overview for each of the topics.

(i) Emotion and emotion-related states in everyday life: The word ‘emotion’ has a range of meanings, and it is essential for database work to be clear where it stands in relation to the various states and qualities to which the term is applied. The most basic point has already been made – HUMAINE is not restricted to the archetypal ‘primary’ emotions. That leaves the task of forming a conception of the broad domain that databases should cover, and finding adequate ways to describe the phenomena that fall within the domain.

(ii) Multimodality: The literature indicates that we can expect emotion to be registered in speech (both in what is said and how it is said), face, gesture, body language, visceral signals (blood pressure, skin response etc), brain activity, and action (rushing about, behaving aggressively, etc). The key question is which modalities should we consider including in databases? In sheer practical terms, if we have to wire people up to obtain a wide range of modalities, we may not get any data that is emotional, let alone naturalistic. So we need to consider what kind of balance can be struck between multimodality and naturalistic emotional data. All-inclusive multimodal databases may not be realistic options.

(iii) Elicitation methods: As indicated above, we have good reason to be suspicious of emotional data generated by actors. That raises the question of how are we to elicit emotion of the type that occurs in everyday interactions. The literature provides various some ideas and possibilities including the use of data from chat shows, facilitating recall of emotional events, wiring people up to a radio microphone over long periods, recording conversations between friends, inducing emotional states in semi laboratory situations (e.g. showing them unpleasant

picture to induce stress, recording them while playing computer games), making use of call center data, recording play situations with robots. These provide a useful basis for moving forward, but they also raise problems that have to be solved. Some are not particularly suited to multimodality; some offer potential but need development to be of wider use (e.g. there many laboratory techniques for inducing stress, but not as many techniques for inducing a wider range of emotional behaviour).

Two other factors bear on how we elicit emotion. One is to do with the quality of recording that we need. Some methods are attractive in that they produce naturalistic data, but they may be too noisy to be of any practical use. Some methods may produce an adequate quality for one modality but not another: for example naturalistic emotional behaviour may produce speech which can be analysed but may lead to a lot of head movement which is a problem for reliable face recovery.

There are also ethical and legal considerations which bear on how we elicit emotion. At the end of the day we must have elicited emotion according to ethical and legal codes that allow us to make it fully accessible to members of Humaine and a wider research community.

(iv) Scope, scale and design of databases: The focus on naturalistic data and on everyday emotional expression raises issues of scope, scale and design.

The traditional focus on 'primary' emotions suggested roughly what the emotional scope of a database should be. But everyday emotional behaviour involves a much wider range of states - Cowie and Cornelius (2003) indicate that lists of emotion terms in English tend to contain 100-200 words. Hence, we need to determine what kind of range of emotional behaviour it is desirable and/or practical to consider.

Traditionally the total length (in time) of databases was also quite low, often because acted representations of emotion could be captured in a short time frame. If we are going to capture everyday emotional behaviour in more naturalistic settings, we are probably going to need to collect much larger amounts of data. A significant factor is highlighted by work on emotion in large call center databases: with that kind of naturalistic resource, only a very small proportion of the data collected can be considered emotional in any strong sense (Ang et al. 2002).

There is also a question about the numbers of subjects to be used. Two strategies have been used in the past. One is to collect large quantities of data from a few subjects, or even one. That has enormous advantages from the point of view of analysis, but raises questions about generalisation. The other is to collect small amounts of data from each of a large number of subjects. That runs the risk of obscuring anything but the 'lowest common denominator'. If the work based on the data is to be of wider value, it needs to be based on statistically sound samples, and that may well mean ensuring that both within- and between- subject variability can be studied.

There are other issues of design which are not currently well addressed in existing databases and which become more salient when faced with naturalistic data and everyday emotion. In particular we need to address issues of context and time. Emotion in naturalistic data involves a wide range of contexts; it occurs as part of a surrounding context rather than in isolation; and it fluctuates and shades into other states over time. Existing data is often highly context dependent: the material is sometimes linguistically constrained or highly selective (for example at word level in call center data). It is often also decontextualised: people read emotional passages in isolation, or emotional episodes are culled from their surrounding context. The result is databases that present emotion as if it consisted of discrete episodes of

relatively constant, high intensity, instead of a phenomenon that modulates over time in rather complex ways.

Humaine databases also need to include data that allow comparisons of the expression of emotion across genders and cultures.

A final issue is whether we want to extend databases to include animal emotion or to create databases of emotional music. There are particular expertises in these areas in the Humaine network, and they are of interest both in terms of theory (for instance data on animal emotions are highly relevant to standard claims about the evolutionary basis of emotion) and in applied terms (music is a highly practical way to influence people's emotional state).

(v) Labelling and format of databases: There are two levels of labelling to be addressed. The first is the encoding of emotion-related content. The second is the encoding of the signs of emotion. Although existing databases go some way towards addressing these, there is much work still to be done, particularly in the context of the focus of Humaine on everyday emotion and multimodality. We cannot assume that the types of labels we have used until now (mainly in the context of acted data and 'primary' emotions) will transfer, and so development on this level is needed.

In terms of the encoding of emotion-related content, categorical labelling has been the dominant method used in existing databases. That is appealing when dealing with a restricted set of emotions. But if we are dealing with everyday emotional expression, we know from the literature that hundreds of categorical terms are potentially relevant (Cowie & Cornelius 2003). That raises problems for labelling.

In terms of the encoding of the signs of emotion, there is a standard system for the labelling of facial signs, and there is quite a wide range of literature covering the relevant signs of emotion in speech (Roach 2000, Cowie et al. 2001). But there is no reason to believe that these systems designed around posed/acted data will transfer well to everyday emotion. And even within areas such as speech where there is a considerable body of data on the signs of emotion, there is disagreement on the natural unit for labelling and inconsistency of findings. There are also some modalities where the signs have been less well explored, though that is developing.

Finally, we need to think through the electronic format of databases, the platform to be used, the annotation within a database and across databases. Multimodality arises some interesting issues here, as the time unit relevant to one modality may not transfer across modalities.

## C. Conclusion

The core message of this section has been that Humaine requires new databases. A lot of experience has accumulated in the field, but it needs to be transferred into a context that prioritises multimodality and everyday emotional behaviour. This report looks at what has been achieved and what we can use and learn from our experiences. It also sets out the problems and challenges, and identifies the goals.

### 3. Review of key concepts in the thematic area

As Section 2 has indicated, Humaine involves working across a wide range of sub areas – psychology, experimental techniques, speech, face, gesture, physiology etc. Members of Humaine and other readers of this report may well be familiar with one sub area but not necessarily the others. This section is intended to facilitate understanding and interaction by providing a brief guide to key terms and concepts in all the relevant areas. The guide below is structured according to the topics set out in section 2.

#### A. Emotion

The literature around the term ‘emotion’ is complex. But some core terms and concepts are a useful guide when trying to understand it, and especially when trying to think in practical terms of the scope of emotional behaviour that databases might cover.

When people hear the term ‘emotion’, they tend to think first of brief episodes when rationality is at least partially suspended and patterns of action and cognition rooted in evolution take control. There is no universally agreed term for these episodes. Two useful candidates are ‘*fullblown emotions*’ (Scherer 1999) and *modal emotions*, i.e. distinctive modes of operation in which emotion is the dominant factor (Scherer, 1994). Terms such as ‘*primary emotions*’ (Plutchik, 1984), ‘*basic emotions*’ (Stein and Oatley, 1992), and ‘*acute emotions*’ (Lazarus, 1994) are related, but they carry specific theoretical implications.

It is obvious from the discussion under section 2 above that Humaine is concerned with more than ‘fullblown emotions’, and it is useful to be aware of the terms that people have used to describe and categorise the emotional behaviour that lies outside this domain.

Cowie and Cornelius (2003) provide an overview and a working vocabulary. They argue that fullblown emotions are extreme cases of a much more general phenomenon – states where the balance of a person’s priorities and dispositions and perceptions are shifted in a characteristic way from a notional ‘norm’. They call these ‘*emotional states*’. Scherer (2000) offers a taxonomy for some of the main examples - moods, interpersonal stances with an affective component (e.g. warm), attitudes (e.g. considering something valuable or repugnant), and personality traits (e.g. being morose or benign). Others could reasonably be added - complex states with an emotional element (e.g. remorse), emotional undertones that colour a person’s state of mind rather than dominating it, and so on.

Emotional states as such shade into states which tend to co-occur with emotional states, and which have some properties in common with them, though they contrast in others. Cowie and Cornelius use the term ‘*emotion-related*’ to describe these states on the periphery of the domain of emotion. Examples include alertness, stress, boredom, agitation and so on.

The general position of HUMAINE is that it should not exclude states with a claim to belong in the broad domain of emotional and emotion-related states unless there is good reason to do so. If nothing else, it is hard to evaluate an interface that is supposed to detect anger without checking whether it discriminates anger from stress. That evaluation cannot be achieved without databases that include comparable samples of stress and anger.

## B. Modalities

There are six modalities that might be included in the list relevant to emotional expression – speech, face, gesture/body language, physiology, brain images, and actions.

### 1) Speech

This is often divided into two levels – *linguistic* and *non linguistic*, linguistic being the actual words used, the way they are put together and the meaning they express, and *non linguistic* being the manner in which they are produced via the pitch, intensity, quality of the voice, timing, and so on.

At the linguistic level, the term *propositional content* is used to refer to the meaning intended or expressed. *Emotion lexicons* specify the emotional content of words considered in isolation.

At the non linguistic level, the term *paralinguistic* is often loosely used to refer to a variety of features that combine to constitute the manner in which the words are produced. These include reflex noises such as gulps, sobs etc, pitch, intensity, duration, timing, voice quality. The term *prosodic* is often used to refer specifically to variations in intensity and pitch and other features that have a linguistic function: for example particular patterns of pitch (*intonation*) may have a grammatical or semantic function. However an exact distinction between paralinguistic and prosodic is not always drawn.

### 2) Face

Technical work on facial expression has a long history. The *Duchenne smile* was identified in the 19<sup>th</sup> century as a distinctive pattern associated with genuine happiness. Ekman and Friesen (1978) developed that approach and produced a system for describing all visually distinguishable facial movements called the *Facial Action Coding System (FACS)*. FACS is an anatomically oriented coding system, based on the definition of ‘action units’ (*AUs*) of a face that cause facial movements. Each *AU* may correspond to several muscles that together generate a certain facial action. Ekman and co-workers have generated a dictionary called *EMFACS* which lists certain key *AUs* and the actions that can co-occur with them to signify one of seven archetypal emotions. They also provide a database called *FACSAID* which can serve as a platform for translating *FACS* scores into emotion measurements.

Ekman’s work strongly influenced the MPEG standard (see below).

### 3) Gesture

There are extensive discussions of gesture in two other deliverables which accompany this one (from WP4 and WP6. Here we present a minimal outline of points which relate directly to database issues.

Methods for the encoding of gestures can be classified on a continuum ranging from purely semantic representations (related only to the meaning of the gesture) to purely physical formats (related only to the form of the gesture). Most existing representation languages for

computational systems are based on semantically oriented systems developed in psychology and research on sign language.

In semantic terms, hand movements are broadly classified with respect to their function:

- *Semiotic*: these gestures are used to communicate meaningful information or indications
- *Ergotic*: manipulative gestures that are usually associated with a particular instrument or job and
- *Epistemic*: these are concerned with providing tactile feedback that helps to confirm what or where an object is.

Semiotic hand gestures are considered to be connected, or even complementary, to speech. They may be used to convey both linguistic content and emotions (often simultaneously). Two major subcategories are particularly associated with linguistic functions, namely *deictic gestures* (pointing at something that is being talked about) and *beats*, i.e. gestures that mark or emphasise phrasal structure. Cutting across that classification, certain gestures are considered spontaneous, free form movements of the hands during speech (gesticulation), while others, termed *emblems*, are conventionalized signs of emotion or action, such as an insult or a nod of agreement.

Physical capture and description raises a completely different set of issues. Physically, gestures may be represented by both temporal hand movements and static hand postures. Capturing these movements and postures is a difficult task. A good deal of work has been based on mechanical capture, notably with *glove-based devices*. *Optical tracking* is useful for large-amplitude gestures and whole-body movement. Vision-based techniques are less intrusive, but demand sophisticated processing. At a basic level, *localisation* of eg a hand is a substantial problem. Two types of solution are often used in the localization process, based on skin color and motion cues. Capturing gesture as such involves another layer of analysis, which depends on fitting *dynamic 3-D models* to image features. Because many temporal gestures involve motion trajectories and hand postures, they are in some respects more complex than speech signals.

There is a good deal of common ground in the key research issues that emerge from both perspectives. Semantically oriented research is interested in developing a *gesticon*, that is, the gestural equivalent of a lexicon. It would comprise definitions of gesture types rather than concrete body-specific instances. That has close links to the physical problem of finding a suitable methods of modeling the hand and the movement patterns that it may display, which is still an open research problem.

Similarly, both sides have strong interests in defining possible and meaningful sequences of gestures: the aim is sometimes described as a *gesture grammar*.

This outline identifies some of the key issues for database collection. It needs to be considered whether WP5 will include a systematic effort to document a wide range of gestures, and if so how they will be recorded (eg using gloves, tracking, or video; and if with video, what the issues of lighting and camera position are). It also needs to be considered whether there will be a systematic effort to annotate databases with gestural coding.

Those issues relate to the objectives of other workpackages. Developing a gesticon or a gesture grammar would be interesting objectives, but adopting them would have major implications for the database effort.

#### 4) Physiological

One of the most famous theorists of emotion, William James, proposed in the 19<sup>th</sup> century (1884) that emotion was simply awareness of bodily changes precipitated by highly significant events (often called *visceral changes* to distinguish them from physiological changes in the brain). As a result, people who have never heard of James continue to believe that physiology is the true measure of emotion. In fact, the correlations between emotion and patterns of physiological activity are fairly gross. Physiological measurements that are known to be relevant include *heart rate, blood pressure, temperature, respiration, galvanic skin response* (a measure of skin conductivity linked to secretion of sweat), and electromyographic activity in certain areas, notably the muscles associated with frowning (Levenson 1992; Picard et al. 2001).

#### 5) Brain imaging

*Brain imaging techniques* potentially add another dimension to physiological observation, and form another modality. They fall into two main groups.

*EEG (Electroencephalography)* A long-established group of techniques measure electrical activity in various frequency bands. The best known correlates of emotionality found with these techniques involve prefrontal asymmetry – that is, positive affect is associated with greater activity in the left prefrontal region than the right, negative affect with the reverse.

*fMRI (functional Magnetic Resonance Imaging)* provides much higher resolution images of regions where activity is occurring. One of its main outcomes is to highlight activity in the *amygdala* as a correlate of emotional involvement. The amygdala are part of the limbic system, a relatively old brain structure in evolutionary terms. They have rich inputs from sensory systems, and appear to be involved in learning the reward values of stimuli – so it is natural to interpret them as a key site in evaluating situations as positive or negative. Other areas that imaging implicates in emotionality are orbitofrontal cortex (which is involved in preparing behavioural responses and autonomic responses, so it is natural to link its function to action tendencies); and the basal forebrain, which has widespread effects on cortical activation, and direct links to autonomic nuclei: that suggests a role in arousal.

An important practical difference between these techniques is that EEG can be measured with electrodes mounted in a cap or band; whereas fMRI requires the subject to be enclosed in a metal tube which generates a powerful magnetic field. This has implications for elicitation of emotion.

#### 6) Action

The term ‘action’ covers a range of measures concerned with deliberate movements that are or may be influenced by emotionality. These include error rates in speeded response tasks, ‘exceedances’ in driving (ie excursions beyond the normal safe limits in terms of position on the road, rate of acceleration, etc), excited waving of the arms, and many others.

## C. MPEG

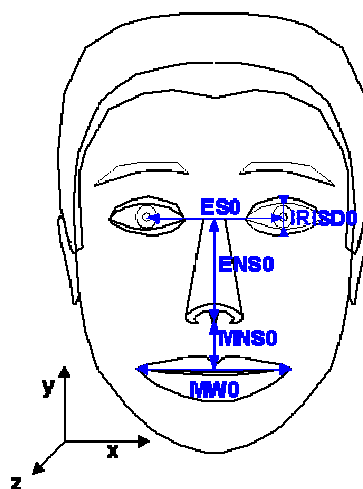
*MPEG* is an acronym for Moving Picture Experts Group – a group which has a major effect on standards in IT in general, and specifically on contemporary coding schemes. Particularly relevant is MPEG-4.

*MPEG-4* provides an “object profile” to describe the syntax and the “coding/decoding tools” for a given “media object”. Profiles exist for various type of media content (audio, visual and graphics) and for scene descriptions. The “Face and Body Animation” tools allows one to transmit parameters that can define, calibrate or animate models. The models themselves are not standardized in MPEG-4. Only the parameters and their transmission are part of the standard.

In particular, MPEG-4 defines a coding scheme to drive the animation the animation of synthetic facial and body models by defining several sets of parameters:

- Facial Animation Parameters (FAPs): describe the facial movements at a low level (for e.g. move horizontally the outer right lip corner) and at a high level (visemes and expressions of emotions)
- Facial Definition Parameters (FDPs): defines the structure of the face. These parameters may be used to modify the geometry of the facial model or to encode the necessary information to transmit a new model.
- Body Animation Parameters (BAPs): define the joints value of the body. The body joints may follow the H-Anim specification.

Two sets of parameters describe and animate the 3D facial model: facial animation parameter set (FAPs) and facial definition parameter (FDP). The FDPs define the shape of the model while FAPs define the facial actions. When the model has been characterized with FDP, the animation is obtained by specifying for each frame the values of FAPs. The first FAP defines visemes while FAP 2 corresponds to expressions of emotion. All other FAPs (the remaining 66) correspond to single facial parameters.



**Figure 1: A face model in its neutral state and the feature points used to define FAP units (FAPU).**

The shape, texture and expressions of the face are generally controlled by the bit stream containing instances of Facial Definition Parameter (FDP) sets and Facial Animation Parameter (FAP) sets. Upon initial or baseline construction, the Face Object contains a generic face with a neutral expression: the Neutral Face. The face is capable of receiving the FAPs from the bit stream, which will produce the animation of the face. If FDPs are received, they are used to transform the generic face into a particular face determined by its shape and (optionally) texture.

In order to define face animation parameters for arbitrary face models, MPEG-4 specifies 84 'feature points' associated with relevant parts of the human face. They are subdivided in groups and labeled with a number, depending on the particular region of the face to which they belong.

## **D. Emotion elicitation**

### **1) Acted /posed data**

These are terms that are used to describe data collected by asking actors to simulate particular emotions. Acted tends to be used of speech data; posed of facial data. In terms of speech, data collected in this manner is sometimes in the form of isolated words, of *semantically neutral content*. Where there is more speech it is usually in *monologue* form and is often *scripted*. Facial data is often *static* rather than moving or *kinetic*.

### **2) Reading**

Reading is often used to obtain speech data. Questions arise about the value of that approach because it is well known that the speech produced by reading differs at multiple levels from spontaneous speech (Johns-Lewis 1986).

### **3) Naturalistic data**

The term *naturalistic* is used refer to a range of data that is collected in non laboratory conditions. It is unscripted and often involves *dialogue / interactive discourse*. *Sociolinguistic fieldwork techniques* are sometimes employed to elicit it, for example minimising the role of the observer by letting participants record themselves (Campbell 2002) or increasing the informality of the situation by bringing friends together to talk or using interviewers to steer the conversation towards emotionally heated topics.

### **4) Call center data**

Call center data is a particular form of naturalistic data that has a high profile at the minute in the area of speech and emotion. This data is naturalistic in the sense that it is not collected in laboratory conditions, but it is far from the spirit of sociolinguistic fieldwork. For a start it

involves *human-machine* interactions, and although it is unscripted (for the human participant), it is scripted for the machine participant, which forces the human participant into a very linguistically limited discourse.

## 5) Induced data

*Induced* data is obtained in laboratory conditions, but is not acted. Emotions are induced by stimulating subjects to produce an emotional response. *Stimuli* may take the form of pieces of music or pictures or subjects may be helped through various techniques to recall a time when they were in an emotional state.

## E. Episodes in databases

We use the term *emotional episode* to describe the basic units of emotional life that make up a database. The literature shows that very varied types of unit have been used, and the rationale is not often articulated. In acted data an emotional episode is often a discrete, constant and quite short episode. But in naturalistic data, emotional behaviour may not be discrete, constant or short and so decisions may have to be made about how to demarcate the relevant emotional portion from a bigger context. Call center data tends to be analysed in *turns*, ie sequences of words spoken by one party and bounded by speech from the other party: that is almost inevitable, since user utterances tend to last a few words and to be separated by system utterances. Dealing with human-human conversations, which are much more fluid, Douglas-Cowie et al (2003) took the decision to select what they called *emotional clips* from the surrounding relatively neutral context. Clips were selected to reflect the build up from a non emotional state to a peak and back to a relatively neutral state. We will call these *beb clips* (short for baseline-emotional-baseline) when it is necessary to indicate that the term clip is being used in this specific sense. Not all studies have used beb clips as units – for instance, the Reading-Leeds study (Roach 2000) pulled out the emotional peak from the surrounding context.

## F. Labelling and format

### 1) Encoding of emotion-related content

A number of ways of encoding emotion-related content are available.

The one that has dominated work on speech and emotion and work on face and emotion is *categorical labelling*. That is well suited to the focus on full-blown emotions that are produced in acted data; emotions of this type may have an inherently categorical structure. Labellers often have to make a *forced choice* decision– that is they are given a limited number of categories to choose from (the same as the target categories that the actors were asked to produce) and they have to label the emotion they hear with the most appropriate category. Sometimes judges are allowed *open-ended* choice.

If categorical descriptions are used, it is not advisable to revert to the ‘folk science’ that there are five ‘primary’ emotions from which all others are derived. Table 1 below gives a well known summary of labels that have been considered in some sense fundamental by significant research groups.

Table 1  
A Selection of Lists of “Basic” Emotions

Reference	Fundamental emotion	Basis for inclusion
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Relation to action tendencies
Ekman, Friesen, & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda (personal communication, September 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
McDougall (1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer (1960)	Pain, pleasure	Unlearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

*Note.* Not all the theorists represented in this table are equally strong advocates of the idea of basic emotions. For some it is a crucial notion (e.g., Izard, 1977; Panksepp, 1982; Plutchik, 1980; Tomkins, 1984), whereas for others it is of peripheral interest only, and their discussions of basic emotions are hedged (e.g., Mowrer, 1960; Weiner & Graham, 1984).

More recent work on naturalistic data has moved from categorical labelling to broad *dimensional* labelling – a concept introduced by Wundt (1903), refined by Schlossberg (1954), and widely used since. Statistical analyses indicate that emotion concepts reflect two main dimensions usually called *evaluation* (how positively or negatively a person regards his/her situation) and *activation* (how strongly he/she is disposed to take action), and a number of others which are less important such as power or approach/avoidance

A key attraction of dimensional labelling is that it is more statistically tractable than categorical labelling, particularly when, as tends to happen in naturalistic data, a wide range of states is present (categorical labelling can result in a plethora of labels which are statistically intractable). In a similar way, it also deals with fluctuating and non discrete emotion better. In practical terms, dimensional descriptions can be assigned using the *FEELTRACE* program (Cowie et al 2000, Cowie & Cornelius 2003). *FEELTRACE* allows a rater to listen to (and/or watch) a recording, and simultaneously to move a cursor to a point on the screen whose co-ordinates reflects the speaker’s current state in terms of activation and evaluation.

A third approach is *logical labelling*. In effect it locates the emotional state being considered at the end of a branching tree. That conceptual approach was pioneered by Ortony et al. (1988), Roseman (1991) and others.

A particularly interesting form of logical labelling is based on the *appraisal* model. The approach is associated with theorists who argue that distinct types of emotion correspond to distinct ways of appraising the situation that evokes the emotion. Work in that tradition attempts to identify logical primitives from which descriptions of appraisals can be constructed, and combinations of elements that capture the distinctive appraisals underlying particular emotional states. Scherer (1999) provides a wide-ranging review of that approach and a model that could in principle be translated into a labelling system.

## 2) Encoding of the signs of emotion

We use the term *signal* to refer to a fluctuating activity in a physical channel. In contrast, *signs* are high-level patterns, corresponding roughly to words, which have the same significance even if irrelevant details are varied. Databases in other areas are often annotated to display the signs that they carry. We introduce the term *Interface Language for Signs of Emotion (ILSE)* to describe an annotation which would correctly define the signs in a database of emotional behaviour.

Considering potential models of an ILSE highlights a curious divergence. One style of encoding predominates when humans are describing emotion, and it is also associated with synthesis procedures (signs to signals); but it is very unlike representations involved in proceeding from signals to signs (analysis).

In the context of human encoding and synthesis, it is standard to use *Markup Languages*. Markup language is a general term for a formalised system used to annotate descriptions of events. They may serve several purposes. For instance, they make information searching more effective (in the Semantic Web), enable defining media and device-independent structures of documents (in natural language and multimedia generation) and annotate information content and structure of natural corpora.

Emotion-related markup languages may be defined at several 'abstraction levels'. At the highest level, they may denote the mental state of the message producers, in terms of their beliefs, desires, intentions and affective state. Going down in this hierarchy, they may denote message content: its communicative goal, focus, relation with other parts of the message, and so on. At the lowest level, tokens provide a symbolic representation of the surface structure of the message: linguistic features, facial expressions, gestures, speech characteristics and other forms that people perceive as the means by which other people express the mental state of the sender and the message content.

In contrast, analysis often naturally ends with patterns that have direct correlates in the signal, but not in human perceptual experience – such as (in the speech domain) measures of F0 (which is almost pitch, but not quite), formant tracks, cepstral parameters, etc.. Even events that people think of as wholly physical (such as pauses) are not easily mapped onto actual physical measurements. Specifying the features that seem to be relevant to the emotional effect of a single phrase may involve tens, or even hundreds of numbers. That raises the question of whether the information contained in standard markup languages is of the right order of magnitude to specify emotion.

The question has been posed in terms of speech, but it applies to any modality. It is not in dispute that standard coding schemes can support simulations of fullblown emotions that

people can categorise with high levels of success, but that is not the ultimate measure of a coding scheme.

The view taken in this report is that these divergences reflect deep unknowns. It is not clear what kind of ILSE should stand between the level of emotional states and the level of measurable physical signals. Hence, databases should not have written into them intermediate symbols whose meaning is at present quite unclear. That theme is taken up again in section 7H.

## 4. Review of key achievements in the thematic area

### A. Data: the state of the art

The three tables that follow summarise the state of the art with regard to databases. Table 1 lists multimodal databases; Table 2 lists speech databases; and Table 3 lists face databases. Data on gestures and physiological signals is contained within Table 1, as this type of data tends to occur within a multimodal context.

The tables are not intended to be a definitive statement of all the relevant databases that exist. But they do include the key databases and are indicative of the type of data that is available.

In most cases the term database is used to refer to a body of data that has been collected (or in some cases ongoing) rather than to a fully marked-up and annotated corpus.

Each table is arranged from left to right according to the same format – identifier for the database, modalities recorded (where more than one), description of how the data was elicited, indicator of size, further information regarding the type of data and finally, any information on cultural/linguistic range.

#### (i) Multimodality and emotion

The dates of databases in this domain indicate that work on multimodality and emotion is very recent and also that the area is gaining ground.

However the core point to be made is that databases of multimodality and emotion are still unusual. Most common within the multimodal domain is audiovisual data which focuses on face and speech. The biggest database of this type is the Belfast Naturalistic Database (Douglas-Cowie et al. 2003) which consists of 125 speakers each recorded speaking in a neutral state and also in at least one emotional state. Data on gestures and on physiological signals is rare as are databases that combine more than 2 modalities. The exception is the ORESTEIA database (McMahon et al. 2003, see also <http://manolito.image.ece.ntua.gr/oresteia/>) which records speech, face and physiological measurements from subjects on a driving simulator. The subjects encounter various problems while driving (deliberately positioned obstructions, dangers, annoyances ‘on the road’). These are intended to induce emotional responses. The SMARTKOM (Schiel et al. 2002; Steininger, Schiel & Glesner 2002; Steininger et al. 2002) database also addresses more than two modalities (speech, facial expression, gestures).

The range of emotions covered in recent multimodal databases tends in the direction of everyday emotional behaviour rather than full blown emotions. Recent work on the SALAS database (Sensitive Artificial Listeners Association, work in progress at Queen’s Belfast as part of ERMIS project IST-2000-29319) and on the SMARTKOM database is very much in this spirit. The SALAS database contains a range of emotional states and emotional-related states that are induced from subjects using an ‘artificial listener’. The system is not yet automated, and at the minute a live person stands in for the artificial listener. The ‘artificial listener’ has 4 different personalities – Pippa, Spike, Obadiah and Prudence. Each personality has set phrases reflecting his/her particular personality (commonsense, angry, sad, happy) which are used to conduct a ‘conversation’ with the subject and to pull the subject towards that personality. The system does not have natural language capabilities. The emotional states

produced have been coded on a dimensional emotional scale and are not full-blown emotions. The SMARTKOM database is also built from listeners' responses to a 'machine'. In fact the machine is actually two humans in another room operating in a Wizard of Oz type situation. Users are asked to solve a range of tasks. The emotional states and related states recorded range from much wider than the traditional 'primary' emotion group. They are joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral. Other multimodal databases reflect the same trend with the Geneva lost luggage database containing examples of good humour and indifference among the emotion-related states listed and the Belfast naturalistic database covering a wide range of emotional states with a wide spread on a two-dimensional representation of emotion (based on the dimensions of evaluation and activation).

Some multimodal databases do not include any emotional content. Two large projects XM2VTSDDB ([www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/](http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/)) and ISLE (<http://isle.nis.sdu.dk/>, IST project IST-1999-10647) do not appear to have emotional data, though ISLE in particular has developed labelling systems especially for multimodal work. Some databases appear to be multimodal (face and speech), but focus has been on the speech (Polzin & Waibel 2000; Banse & Scherer 1996).

In terms of the naturalness of the data, there has clearly been a considerable effort put into moving away from acted data to more naturalistic data. A variety of methods have been used. Some use real life situations such as lost luggage offices (Scherer & Ceschi 2000) or television chat shows (Douglas-Cowie et al. 2003; Chung 2000); some use induction techniques to induce a range of emotions (SALAS database, Belfast Boredom database –see Cowie et al. 2003, SMARTKOM). These techniques have increasingly moved away from an earlier tendency to focus on stress-related states and are instead being developed to produce a much wider range of emotions. SMARTKOM and SALAS have already been described. The boredom database is produced by getting subjects to carry out a boring task on a computer. It consists of asking them to name a series of shapes on a computer screen over and over again.

Despite the obvious progress that has been made in recent years towards multimodal databases of emotion, there are clear problems and gaps. First most of the work is not really truly multimodal – it is at best usually focused on two modalities, with gesture and physiology mostly ignored. Second, the data is flawed or underdeveloped in terms of naturalness. The 'truly' natural data tends to be noisy and a problem for some types of analysis. In addition there are real problems of copyright leading to the material not being freely available for researchers. The induced data has many limitations at the moment. SMARTKOM clearly produces an interesting range of emotional-related states, but it is not clear how context dependent they are. SALAS is in the early stages of development: it has yet to be seen how convincing the emotional behaviour is that is produced. Work is in progress on developing a Greek language version of SALAS.

Current developments not listed in the table indicate that some of the problems and gaps are beginning to be addressed. Hook et al in KTH, Sweden are experimenting with a new naturalistic technique (which they have called the Shit Study – see [www.dsv.su.se/~petra/Shitstudy/](http://www.dsv.su.se/~petra/Shitstudy/)) designed to elicit gestures (one of the missing modalities) as well as speech. They have piloted it in Sweden with Swedish speakers. Young females are brought together to 'bitch' about pop stars and how they dress etc. There are plans currently under way to replicate the study in France and Ireland. Vered Aharonson (personal communication) in the Tel Aviv group in Humaine has conducted an experiment which brings together speech and physiological measures and aims to tap into a range of 'everyday' emotions. Subjects play a computer game, in which they gamble with various sums of money.

The game is voice operated - basically the subjects repeat the same 2 phrases ("open this door", "close this door") over and over. Three emotions are expected: anticipation, disappointment and content. 8 recordings have been made so far. The SALAS scenario is being developed to be run in a WOZ situation so that more natural recordings may be collected. Suitable voice-overs for the different personalities have also been recorded, and these will be used in the WOZ situation so that users enter into the spirit of the personalities and 'play' with them or relate to them.

Table 1: Multimodal databases (Note: in Column 2 'audiovisual' refers to speech and face unless otherwise indicated)

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Nature of material	Language
Belfast Naturalistic Database (Douglas-Cowie et al 2000, 2003)	Audio- visual	Wide range (see Fig. 1 in this report )	Natural: 10-60 sec long 'clips' taken from television chat shows, current affairs programmes and interviews conducted by research team	125 subjects; 31 male, 94 female	Interactive unscripted discourse	English
Geneva Airport Lost Luggage Study (Scherer & Ceschi 1997; 2000)	Audio-visual	Anger, good humour, indifference, stress sadness	Natural: unobtrusive videotaping of passengers at Geneva airport lost luggage counter followed up by interviews with passengers	109 subjects	Interactive unscripted discourse	
Chung (Chung 2000)	Audio-visual	Joy, neutrality, sadness (distress)	Natural: television interviews in which speakers talk on a range of topics including sad and joyful moments in their lives	77 subjects; 61 Korean speakers, 6 Americans	Interactive unscripted discourse	English and Korean
SMARTKOM <a href="http://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom">www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom</a>	Audio-visual, (+gestures)	Joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, neutral	Human machine in WOZ scenario: solving tasks with system	224 speakers; 4/5 minute sessions	Interactive discourse	German
Amir et al. (Amir et al, 2000)	Audio + physiological	Anger, disgust, fear, joy, neutrality, sadness	Induced: subjects asked to recall personal experiences involving each of the emotional states	61 subjects 60 Hebrew speakers 1 Russian speakers	Non interactive, unscripted discourse	Hebrew Russian
SALAS database ( <a href="http://www.image">http://www.image</a> .	Audio-visual	Wide range of emotions/emotion	Induced: subjects talk to artificial listener & emotional	Pilot study of 20 subjects	Interactive discourse	English (Greek

<a href="http://ntua.gr/ermis/">ntua.gr/ermis/</a> IST-2000-29319, D09)		related states but not very intense	states are changed by interaction with different personalities of the listener		Subjects unscripted Machine scripted	version being developed)
ORESTEIA database (McMahon et al. 2003)	Audio + physiological  (some visual data too)	Stress, irritation, shock	Induced: subjects encounter various problems while driving (deliberately positioned obstructions, dangers, annoyances 'on the road')	29 subjects, 90min sessions per subject	Non interactive speech: giving directions, giving answers to mental arithmetic etc	English
Belfast Boredom database (Cowie et al. 2003)	Audio-visual	Boredom	Induced	12 subjects: 30 minutes each	Non interactive speech: naming objects on computer screen	English
XM2VTSDB multi-modal face database <a href="http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/">http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/</a>	Audio-visual	None	n/a	295 subjects ; Video	High quality colour images, 32 KHz 16-bit sound files, video sequences and a 3d Model + profiles ( left-profile and one right profile image per person, per session, a total of 2,360 images), scripted 4 sentences	English
ISLE project corpora ( <a href="http://isle.nis.sdu.dk/">http://isle.nis.sdu.dk/</a> , IST project IST-1999-10647)	Audio-visual  + gesture	None	n/a	unclear		
Polzin (Polzin, 2000)	Audio- visual (though only audio channel used)	Anger, sadness, neutrality (other emotions as well, but in insufficient numbers to be used)	Acted: sentence length segments taken from acted movies	Unspecified no of speakers. Segment numbers 1586 angry, 1076 sad, 2991 neutral	Scripted	English

Banse and Scherer (Banse and Scherer 1996)	Audio- visual (visual info used to verify listener judgements of emotion)	Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame	Acted: actors were given scripted eliciting scenarios for each emotion , then asked to act out the scenario.	12 (6 male, 6 female)	Scripted: 2 semantically neutral sentences (nonsense sentences composed of phonemes from Indo-European languages)	German
--	--	---	---	-----------------------	--	--------

### (ii) Speech and emotion

There has been a considerable body of audio data collected for speech and emotion studies, as reflected in Table 2 which shows more corpora than either of the other tables.

Much of the data has three characteristics: the emotion in it is simulated by an actor (not necessarily trained); the actor is reading preset material; and he or she is aiming to simulate full-blown emotion (Yacoub et al 2003, Kienast and Sendlmeier 2000, Paeschke and Sendlmeier 2000). Other examples in the general vein include Leinonen et al 1997, Nakatsu et al 1999, Juslin and Laukka 2001, Nogueiras et al 2001, Murphy 2003, Oudeyer 2003. There are sometimes attempts to make the data more natural by contextualising the emotion, for example using material to read that is inherently emotional in content (Examples in Table 2 are Mozziconacci 1998, McGilloway 1997).

At the other extreme, there are a few speech databases which are focused on naturalistic data. The Reading-Leeds database, collected by Roach and his colleagues (Roach 2000; Greasley et al. 2000), used emotional episodes from radio broadcasts (such as the commentary on the wreck of the Hindenberg) and covered a range of emotional states, many full-blown. Campbell's CREST database (Campbell 2002; see also Douglas-Cowie et al. 2003) is attempting to acquire truly natural data with a wide range of everyday emotions on an enormous scale by getting volunteers to record themselves on a daily basis.

Also fairly natural, but narrower in emotional range, are datasets that use material recorded during specific types of event, such as game shows, emergency flight situations for pilots, affectively loaded therapy sessions, journalists' reports of emotion-eliciting events (e.g. France et al. 2000 in Table 2 but also Johannes et al. 2000; Frolov et al. 1999; Huttar 1987; Kuroda et al. 1976; Roessler and Lester 1976, 1979; Sulc 1977; Williams and Stevens 1969, 1972).

Another type of natural data is that from databases of real human-machine dialogue, often in call center situations. Table 2 shows some examples, SYMPAFLY (Batliner et al. 2004b), and the DARPA Communicator Corpus used by Ang et al. 2002 (see Walker et al. 2001). The amount of emotional data elicited however may be limited as may the range of emotions (see below). There is also some recent work on human-human call center data (Devillers et al, 2002, 2003, 2004).

There is also data that is induced through more directed elicitation. Techniques are designed to induce states that are both genuinely emotional and likely to involve speech. Early examples include a task where subjects are introduced to unpleasant images (Tollkmitt & Scherer 1986) and some parts of the SUSAS database (Speech under Simulation and Actual Stress database) which use speech elicited in a range of stressful situations (Hansen & Bou-Ghazale 1997, <http://wave ldc.upenn.edu/catalog/>). More recent examples include simulations of call centres designed to elicit irritation (Batliner et al. 2003; Mitchell et al. 2000); a stressful driving task (Fernandez & Picard 2003); a spelling task designed to elicit embarrassment (Bachorowski & Owren 1995). The Erlangen AIBO database (Batliner et al 2004a) uses children's interactions with a robot and produces a wider range of emotional states such as neutral (default), joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding.

In summary it is fair to say that there has been a lot of work over the last decade or so on the development of speech databases of emotion, and that recently there has been a strong movement away from acted to more naturalistic data. Researchers have been quite creative in

this area in experimenting with different methods of collecting data, ranging from opportunistic use of pre-recorded naturalistic emotional situations to laboratory-based induction techniques. There is also a growth in the range of language and cultures covered, with work on Western European languages but also on Hebrew (Amir et al 2000) and on Japanese and Chinese (Campbell 2002). But there are some core problems. Clearly the community itself is recognising that acted material on which it has relied heavily does not do a great deal to illuminate the way speech expresses emotion in natural settings (Batliner et al. 2003; Stibbard, 2000, 2001). And it is also recognised that data protection law creates real problems with the distribution of some of the most emotional and naturalistic data: it has proved impossible to release the Reading-Leeds database for legal reasons (and distribution of the Belfast Naturalistic Database is similarly severely limited).

Accordingly there has been a very strong focus in the last five years or so on human-machine dialogue systems, often in call centre data which is more easily accessible and truly natural. These databases have several attractions. First, the emotion is presumably genuine, not acted. Second, they deal with dialogue, which exposes issues missing from the monologue type data often produced in acted or elicited emotion. Third, they are very directly related to a foreseeable application of emotion recognition. But with these advantages come limitations. The frequency with which emotion is expressed is low. To illustrate the scale of the problem, Ang et al (2002) used material totalling 14 h 36 min of speech. The commonest strong emotion was frustration, of which he obtained 42 unequivocal instances. The nature of the interaction imposes constraints of the forms of utterances and probably on the way emotion may be expressed within those forms, raising major questions about generalisability. Not least, the emotions tend to be from a narrow range, generally negative. Recent studies illustrate the point. They include the study by Ang et al cited above using the DARPA Communicator Corpus: users called systems built by various sites and made air travel arrangements over the phone. Lee and Narayanan (2003) detected negative versus non negative emotion using a corpus of utterances obtained from a commercially deployed human-machine spoken dialogue application; most dialogue turns had one utterance. Boozer et al. (2003) are working on detecting neutral, frustrated and happy states using human-computer dialogues generated by Mercury – a phone based airline flight planning system. The SYMPAFLY system (see Table 2) looks more promising in terms of range of emotions (includes states like ‘helpless, panic, touchy) and data size.

Table 2: Speech databases

Identifier	Emotional content	Emotion elicitation methods	Size	Nature of material	Language
Reading-Leeds database (Greasley et al., 1995; Roach et al., 1998, Stibbard 2001)	Range of full blown emotions	Natural: Unscripted interviews on radio/television in which speakers are asked by interviewers to relive emotionally intense experiences	Around 4 ½ hours material	Interactive unscripted discourse	English
France et al. (France et al., 2000)	Depression, suicidal state, neutrality	Natural: therapy sessions & phone conversations. Post therapy evaluation sessions were also used to elicit speech for the control subjects	115 subjects: 48 females 67 males.  Female sample: 10 controls (therapists), 17 dysthymic, 21 major depressed  Male sample: 24 controls (therapists), 21 major depressed , 22 high risk suicidal	Interactive unscripted discourse	English
Campbell CREST database, ongoing (Campbell 2002; see also Douglas-Cowie et al. 2003)	Wide range of emotional states and emotion-related attitudes	Natural: volunteers record their domestic and social spoken interactions for extended periods throughout the day	Target - 1000 hrs over 5 years	Interactive unscripted discourse	English Japanese Chinese

Capital Bank Service and Stock Exchange Customer Service (as used by Devillers & Vasilescu 2004)	Mainly negative - fear, anger, stress	Natural: call center human-human interactions	Unspecified (still being labelled)	Interactive unscripted discourse	English
SYMPAFLY (as used by Batliner et al. 2004b)	Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic	Human machine dialogue system	110 dialogues, 11581 words	Naïve users book flights using machine dialogue system	German
DARPA Communicator corpus (as used by Ang et al. 2002)  See Walker et al. 2001	Frustration, annoyance	Human machine dialogue system	Extracts from recordings of simulated interactions with a call centre, average length about 2.75 words  13187 utterances in total of which 1750 are emotional: 35 unequivocally frustrated, 125 predominantly frustrated, 405 unequivocally frustrated or annoyed, 1185 predominantly frustrated or annoyed	Users called systems built by various sites and made air travel arrangements over the phone	English

AIBO (Erlangen database) (Batliner et al. 2004a)	Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral	Human machine: interaction with robot	51 german children, 1299 words	Task directions to robot	German
Fernandez et al. (Fernandez et al. 2000, 2003)	Stress	Induced: subjects give verbal responses to maths problems in simulated driving context	Data reported from 4 subjects	Unscripted numerical answers to mathematical questions	English
Tolkmitt and Scherer (Tolkmitt and Scherer, 1986)	Stress (both cognitive & emotional)	Induced: 2 types of stress (cognitive and emotional) were induced through slides. Cognitive stress induced through slides containing logical problems; emotional stress induced through slides of human bodies showing skin disease/accident injuries	60 (33 male, 27 female)	Partially scripted: subjects made 3 vocal responses to each slide within a 40sec presentation period - a numerical answer followed by 2 short statements. The start of each was scripted and subjects filled in the blank at the end, e.g. 'Die Antwort ist Alternative ...'	German
Iriondo et al. (Iriondo et al., 2000)	Desire, disgust, fury, fear, joy, surprise, sadness	Contextualised acting: subjects asked to read passages written with appropriate emotional content	8 subjects reading paragraph length passages (20-40mmsec each)	Non interactive and scripted	Spanish

Mozziconacci (Mozziconacci, 1998) Note: database recorded at IPO for SOBUproject 92EA.	Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality	Contextualised acting: actors asked to read semantically neutral sentences in range of emotions, but practised on emotionally loaded sentences beforehand to get in the right mood	3 subjects reading 8 semantically neutral sentences (each repeated 3 times)	Non interactive and scripted	Dutch
McGilloway (McGilloway, 1997; Cowie and Douglas-Cowie, 1996)	Anger, fear, happiness, sadness, neutrality	Contextualised acting: subjects asked to read passages written in appropriate emotional tone and content for each emotional state	40 subjects reading 5 passages each	Non interactive and scripted	English
Belfast structured Database An extension of McGilloway database above (Douglas-Cowie et al. 2000)	Anger, fear, happiness, sadness, neutrality	Contextualised acting: subjects read 10 McGilloway- style passages AND 10 other passages - scripted versions of naturally occurring emotion in the Belfast Naturalistic Database	50 subjects reading 20 passages	Non interactive and scripted	English
Danish Emotional Speech Database (Engberg et al., 1997)	Anger, happiness, sadness, surprise, neutrality	Acted	4 subjects read 2 words, 9 sentences & 2 passages in a range of emotions	Scripted (material not emotionally coloured)	Danish

Groningen ELRA corpus number S0020 (www.icp.inpg.fr/ELRA)	Database only partially oriented to emotion	Acted	238 subjects reading 2 short texts	Scripted	Dutch
Berlin database (Kienast & Sendlmeier 2000; Paeschke & Sendlmeier 2000)	Anger- hot, boredom, disgust, fear- panic, happiness, sadness-sorrow, neutrality	Acted	10 subjects (5 male, 5 female) reading 10 sentences each	Scripted (material selected to be semantically neutral)	German
Pereira (Pereira, 2000)	Anger (hot), anger (cold), happiness, sadness, neutrality	Acted	2 subjects reading 2 utterances each	Scripted (1 emotionally neutral sentence, 4 digit number) each repeated	English
van Bezooijen (van Bezooijen, 1984)	Anger, contempt disgust, fear, interest joy, sadness shame, surprise, neutrality	Acted	8 (4 male, 4 female) reading 4 phrases	Scripted (semantically neutral phrases)	Dutch

Abelin (Abelin 2000)	Anger, disgust, dominance, fear, joy, sadness, shyness, surprise	Acted	1 subject	Scripted (semantically neutral phrase)	Swedish
Yacoub et al (2003)  (data from LDC, <a href="http://www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28">www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28</a> )	15 emotions  Neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt	Acted	2433 utterances from 8 actors	Scripted	English

(iii) Face and emotion

Table 3 gives a representative selection of key databases of facial expressions. As can be seen from the descriptions in the last column, these show faces under systematically varied conditions of illumination, scale and head orientation. Rather few consider emotional variables systematically, and the range of emotional expressions considered is quite limited and tends to focus on the 'primary' emotions. The data is also generally acted or posed and consists of static images. The term 'staged' is perhaps appropriate. The seminal database of this type is the classic Ekman and Friesen classic collection of photographs showing facial emotion (Ekman & Friesen 1978): this can now be bought in electronic form. Others in the same mould are the Yale database which contains 11 images for each of 15 individuals, one per different facial expression or configuration – center-light, with glasses, happy, left-light, without glasses, normal, right light, sad, sleepy, surprised and wink, and the ORL database of faces which contains 10 different images for each of 40 subjects. the images vary the lighting and aspects of facial expression which are at least broadly relevant to emotion – open/closed eyes, smiling/not smiling.

Rather few databases contain samples of faces moving, and moving sequences which are emotionally characterised are even less common. The material that tends to be available consists of images produced by research software, e.g. for facial animation, rather than the original video sequences used for the analysis or training. Examples can be found at [www.cs.cmu.edu/~face/](http://www.cs.cmu.edu/~face/). Databases that combine speech and video are still rare and the few examples that there are have already been mentioned in Table 1, in particular SMARTKOM and the Belfast Naturalistic Database. The XM2VTSDB multi-modal face database is also audiovisual but does not contain emotion.

The cultural range is dominated by the West, although there is a database of Japanese faces (see Table 3). In summary the data is limited in emotional range and level of naturalness. However the field is developing and genuinely natural data (of moving faces) is emerging with a much wider range of emotional expression. However getting appropriate facial images is not straightforward, and researchers report practical problems along the way. Genuinely natural data involves quite a lot of jerky movement, frequent occlusion of the face, and particular angles that make its use for facial analysis limited. And audiovisual images create real problems in terms of the interference of speech with facial movement. Clearly appropriate balances need to be found.

Table 3: Databases of Faces

Identifier	Emotional content	Emotion elicitation methods	Size	Nature of material
The AR Face Database ( <a href="http://rv11.ecn.purdue.edu/~aleix/aleix_face_DB.html">http://rv11.ecn.purdue.edu/~aleix/aleix_face_DB.html</a> )	Smile, anger, scream neutral	Posed	154 subjects ( 82 male, 74 female)  26 pictures per person	1 : Neutral, 2 Smile, 3 : Anger, 4 : Scream , 5 : left light on, 6 : right light on, 7 : all side lights on, 8 : wearing sun glasses, 9 : wearing sun glasses and left light on, 10 : wearing sun glasses and right light on, 11 : wearing scarf, 12 : wearing scarf and left light on, 13 : wearing scarf and right light on, 14 to 26 : second session (same conditions as 1 to 13)
CVL Face Database ( <a href="http://www.lrv.fri.uni-lj.si/facedb.html">http://www.lrv.fri.uni-lj.si/facedb.html</a> )	Smile	Posed	114 subjects (108 male, 6 female)  7 pictures per person	Different angles, under uniform illumination, no flash and with projection screen in the background
The Psychological Image Collection at Stirling ( <a href="http://pics.psych.stir.ac.uk/">http://pics.psych.stir.ac.uk/</a> )	Smile, surprise, disgust	Posed	Aberdeen: 116 subjs  Nottingham scans: 100  Nott-faces-original: 100  Stirling faces:36	Contains 7 face databases of which 4 largest are: Aberdeen , Nottingham scans, Nott-faces-original, Stirling faces Mainly frontal views, some profile, some differences in lighting and expression variation

<p>The Japanese Female Facial Expression (JAFFE) Database (<a href="http://www.mis.atr.co.jp/~mlyons/jaffe.html">http://www.mis.atr.co.jp/~mlyons/jaffe.html</a>)</p>	<p>Sadness, happiness, surprise, anger, disgust, fear, neutral</p>	<p>Posed</p>	<p>10 subjects 7 pictures per subject</p>	<p>6 emotion expressions + 1 neutral posed by 10 Japanese female models</p>
<p>CMU PIE Database (CMU Pose, Illumination, and Expression (PIE) database) (<a href="http://www.ri.cmu.edu/projects/project_418.html">http://www.ri.cmu.edu/projects/project_418.html</a>)</p>	<p>Neutral, smile, blinking and talking</p>	<p>Posed for neutral, smile and blinking  2 secs video capture of talking per person</p>	<p>68 subjects</p>	<p>13 different poses, 43 different illumination conditions, and with 4 different expressions.</p>
<p>Indian Institute of Technology Kanpur Database (<a href="http://www.cse.iitk.ac.in/users/mayankv/face.htm">http://www.cse.iitk.ac.in/users/mayankv/face.htm</a>)</p>	<p>Sad, scream, anger, expanded cheeks and exclamation, eyes open-closed, wink</p>	<p>Posed</p>	<p>20 subjects</p>	<p>Varying facial expressions, orientation and occlusions; degree of orientation is from 00 to 200 in both right and left direction, the similar angle variation are considered in case of head tilting; and also head rotations both in top and bottom are taken into account. All of these images are taken with and without glasses in constant background; for occlusions some portion of face is kept hidden and lightning variations are considered.</p>

The Yale Face Database ( <a href="http://cvc.yale.edu/projects/yalefaces/yalefaces.html">http://cvc.yale.edu/projects/yalefaces/yalefaces.html</a> )	Sad, sleepy, surprised	Posed	15 subjects	One picture per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink
CMU Facial Expression Database (Cohn-Kanade) ( <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> )	Six of the displays were based on descriptions of prototypic emotions (i.e., joy, surprise, anger, fear, disgust, and sadness).	Posed	200 subjects	Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units (e.g., AU 12, or lip corners pulled obliquely) and combinations of action units (e.g., AU 1+2, or inner and outer brows raised). Subjects began and ended each display from a neutral face
Caltech Frontal Face DB ( <a href="http://www.vision.caltech.edu/html-files/archive.html">http://www.vision.caltech.edu/html-files/archive.html</a> )	Unclear		27 subjects 450 images in total	Different lighting, expressions, backgrounds
HumanScan BioID Face DB ( <a href="http://www.humanscan.de/support/downloads/facedb.php">http://www.humanscan.de/support/downloads/facedb.php</a> )	None	n/a	23 subjects	contains 19 manual markup points: 0 = right eye pupil 1 = left eye pupil 2 = right mouth corner 3 = left mouth corner 4 = outer end of right eye brow 5 = inner end of right eye brow 6 = inner end of left eye brow 7 = outer end of left eye brow 8 = right temple 9 = outer corner of right eye 10 = inner corner of right eye 11 = inner corner of left eye 12 = outer corner of left eye 13 = left temple 14 = tip of nose 15 = right nostril 16 = left nostril 17 = centre point on outer edge

				of upper lip 18 = centre point on outer edge of lower lip 19 = tip of chin
Oulu University Physics-Based Face Database ( <a href="http://www.ee.oulu.fi/research/imag/color/pbfd.html">www.ee.oulu.fi/research/imag/color/pbfd.html</a> )	None	n/a	125 subjects	All frontal images: 16 different camera calibration and illuminations
UMIST ( <a href="http://images.ee.umist.ac.uk/danny/database.html">http://images.ee.umist.ac.uk/danny/database.html</a> )	None	n/a	20 subjects, 19-36 pictures per person	Range of poses from profile to frontal views
Olivetti Research ( <a href="http://www.mambo.ucsc.edu/psl/olivetti.html">www.mambo.ucsc.edu/psl/olivetti.html</a> )	None	n/a	40 subjects, 10 pictures per person	All frontal and slight tilt of head
The Yale Face Database B ( <a href="http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html">http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html</a> )	None	n/a	10 subjects	9 poses x 64 illumination conditions
AT&T (formerly called ORL database)  ( <a href="http://www.uk.research.att.com/facedatabase.html">http://www.uk.research.att.com/facedatabase.html</a> )	Smiling / not smiling	Posed	40 subjects	10 images for each subject which vary lighting, glasses/no glasses, and aspects of facial expression broadly relevant to emotion – open/closed eyes, smiling/not smiling

## B. Labelling and Format

### (i) Encoding emotion

Two approaches have been used. The most common of these is categorical labelling of emotion. The second, which is broad dimensional labelling is beginning to be used. A third approach, using an appraisal-based approach to emotion, has strong theoretical attractions. Some elements of it were used in the Reading-Leeds database (Greasley et al. 2000), but it has not been used as a stand-alone method of labelling databases, and applying it as a stand-alone system would appear likely to be very labour-intensive.

In practice categorical labelling has been applied in a variety of ways. The simplest way, which is often used with acted data is to ask actors to produce a limited number of emotions (through speech or face) and then to ask listeners to assign the productions to the target emotions. A more flexible use of the paradigm is to give listeners an opt-out for productions that do not fit the targets, or to select from a wider range of categories (Belfast Naturalistic Database). Sometimes raters are also asked to give a rating of the intensity of the emotion on a scale.

A more recent development in the categorical labelling of acted databases has been to attach a label of naturalness/genuineness to the emotion produced. Listeners make a choice from one of a range of target emotion labels but also label how genuine it was. Douglas-Cowie and Cowie applied this approach to the Belfast Structured Database using a program called VALIDTRACE. This approach gives a way of ensuring that forced choice categorical labelling is more than a paper exercise where people can match labels to stereotypes.

Broad dimensional labelling has been applied by Douglas-Cowie and Cowie to the Belfast Naturalistic Database (see Douglas-Cowie et al. 2003) and is currently being used in the SALAS database. In practical terms, dimensional descriptions are assigned using the FEELTRACE program (Cowie et al. 2001, Cowie & Cornelius 2003). FEELTRACE allows a rater to listen to (and/or watch) a recording, and simultaneously to move a cursor to a point on the screen whose co-ordinates reflects the speaker's current state in terms of activation and evaluation. Activation/evaluation space is represented by a circle, following a strong psychological tradition that regards emotion as inherently circumplex (Plutchik & Conte 1997) – i.e. it has a natural centre (alert neutrality), and a natural boundary (fullblown emotion), which is the same distance from the centre in any direction. Hence it is as valid to describe emotion in polar co-ordinates (magnitude and direction of emotionality) as it is in Cartesian co-ordinates. FEELTRACE output provides both descriptions.

Categorical labelling and broad dimensional labelling both have their problems. Experimentation on categorical labelling in the naturalistic Reading-Leeds database shows some of the problems with categorical labelling in naturalistic data, particularly open-ended choice. An experiment conducted by Greasley et al. 2000 on the Reading-Leeds database shows the problems. Labellers were asked to label 89 samples of emotional speech in two ways – (i) to assign a label from the 'basic emotion' categories - anger, fear, sad, happy and (ii) to attach a free choice label. Although free choice labels were largely consistent with the 'basic emotion' labels — there was a huge variation of labels, and for 46% of the negative emotion samples there was no clear similarity of response across subjects as to the 'basic emotion' label most appropriate. In conclusion, it would seem that open choice categorical labelling may well be suited to fullblown emotions produced in acted contexts, but may give rise to an intractable number of labels in more everyday, naturalistic data. This is not

surprising given the fact that we need around 100-200 words to cover the territory of emotional states (Whissell 1989; see also overview by Cowie & Cornelius 2003).

In terms of dimensional labelling, raters need considerable training to use a tool like Feeltrace and apply it. A two dimensional representation does not address some of the emotional distinctions that one might want to capture such as the distinction between anger and fear. Experimentation with a three dimensional version of FEELTRACE shows that it can be used but many users find it a problem to cope with the movements needed to operate a mouse in a way that controls three dimensions simultaneously. Nevertheless dimensional labelling using Feeltrace gives numerically tractable results and allows broad-based emotional modelling. It also deals well with emotion fluctuation over time. The rater can trace the emotion in real time and the co-ordinates are recorded at intervals of a few milliseconds. This allows change in emotion to be matched against change in the relevant signs. Alternatively the rating can be averaged across a period of time.

(ii) Encoding of the signs of emotion

Labelling aimed at capturing signs of emotion raises difficult questions which were introduced in section 3F(2), and which are taken up again in section 7E.

## 5. Review of key problems in the thematic area

The root problem in the area is straightforward. It is clear what the ideal database resource would be, but it does not exist, and it is not likely that it will in the foreseeable future. More specific problems flow from that.

An ideal database resource would have the following properties

- It would be fully naturalistic (except insofar as some parts deliberately captured the behaviour of actors, newsreaders, etc)
- It would sample the whole domain of emotion and emotion-related states
- It would represent all the types of action through which emotion and emotion-related states can be expressed
- It would sample the whole range of cultural and individual differences that are important for the expression of emotion
- The recordings would be of high technical quality
- There would be recordings in all relevant modalities
- The data would be comprehensively labelled
- The labelling would follow a uniform standard pattern
- The labelling would include objective verification of all the emotional states involved
- The material would be statistically tractable (for instance it would include balanced samples, samples constructed to match in all respects but one key emotional contrasts, etc)
- The material would be freely available
- The process of obtaining, storing and distributing samples would be ethically sound

Clearly few databases satisfy more than one or two those ideals. Equally clearly, the ideal is not likely to be achieved in full in the foreseeable future. Pure scale is an issue, but there are also problems of principle at every turn. For instance,

- There are considerable problems in balancing the need for multimodality with naturalness. Truly multimodal data can probably never achieve anything approaching naturalness. It is difficult to see how we can have many more than two or possibly three modalities (recorded in ways that can usefully be analysed) without eliciting data that is considerably constrained.
- There are considerable problems in balancing demand for high quality recordings with naturalness. Genuinely natural data involves quite a lot of jerky movement, frequent occlusion of the face, and particular angles that make its use for facial analysis limited.

And audiovisual images create real problems in terms of the interference of speech with facial movement.

- There are considerable problems in balancing demand for objective verification with naturalness, since objective verification tends to demand either intrusive methods or tight control over the situation; and with the subtler emotional states that make up a large part of emotional life, it is difficult to see how objective verification could be achieved at all.
- There are considerable problems in balancing demand for quantity with comprehensiveness of labelling. The more detailed a labelling scheme is, the more labour intensive implementing it is likely to be, and the more skilled the people involved need to be.
- There are considerable problems in balancing demand for statistical tractability with comprehensiveness of labelling. Experience shows it can be all but impossible to extract meaningful relationships from highly detailed labelling schemes unless the quantity of data is very large and the statistical techniques are very powerful.
- There is an almost unlimited range of types of action through which emotion and emotion-related states can be expressed, and many of them are very unlikely to occur in settings similar enough to facilitate comparison (consider, for instance, emotional driving behaviour and expression of emotion through large amplitude hand gestures and balletic movements).

The list could easily be extended.

Two things are essential to deal with a situation that presents problems like these. One is a clear understanding of the need to reach intelligent compromises. It is a major barrier to progress if groups become wedded to one or two ideals, dismiss work that does not fulfil them completely, and embrace work that does, even though it is far short on other criteria. The other essential is development of theory that provides a sound motivation for sampling. The traditional emphasis on 'primary' emotions fulfilled that need in the sense of directing research to a few strategic forms of expression, but unfortunately, it seems clear at this stage that the points of reference it identified did not allow the rest of the domain to be reconstructed in any straightforward way. A more soundly based alternative is badly needed.

## 6. Assessment of the key development goals in the thematic area

Databases are part of a wider research effort, and the development goals have to be seen in that context. Since HUMAINE is a research effort with a long term outlook, the database effort should aim to facilitate work on challenging problems. Key examples are

- interacting with humans who are expressing a variety of emotion freely
- recognising emotion from a wide range rather than discriminating between a small number of prespecified alternatives
- recognising and synthesising the kind of emotional colouring that is typical of everyday interaction rather than rare episodes of intense emotion
- fusion of information from multiple sensory modalities
- analysing the temporal patterning of emotion signs in a way that supports integration of information over substantial time intervals or synthesis with natural timing

What follows is an attempt to identify key elements of a strategy aimed at making progress on these problems.

The unique contribution of the database strand to these areas is primary records. The term is used to cover

- Recordings in audio and visual modalities
- Sensor traces for GSR, ECG, and other monitoring systems
- Records of procedures and/or circumstances relevant to inducing emotion
- Records of actions likely to be affected by emotion (eg errors or erratic performance in a task)
- Self-assessments of the person's feelings during or after the episode recorded
- Assessments of the person's emotional state by other people, either as experts with a claim to make objective judgments or as non-experts representing the way others would tend to perceive the episode in question.

These are described as primary records because so long as they exist, annotations of various sorts can be added. Different annotations may well be added for different applications.

The basic obligation of database research is to provide valid primary records. In practice, that means developing techniques of two main types:

- Ethological: This seems an appropriate term for records of episodes that are observed rather than deliberately created or facilitated by the researcher. At present the main examples are material from call centres and broadcasts. Unbroadcast material from reality TV is another large resource.
- Induction procedures, which divide into two main groups
  - Individual
    - Displaying still pictures, movies clips or music excerpts selected for their emotional impact.
    - Imagination techniques. Including: (a) recall and "reactivate" past emotional experiences; or (b) read emotional scenarios or emotionally loaded sentences and "get into" the corresponding mood (Velten technique, Velten 1968)
    - Facial feedback paradigm
  - Interactive
    - interviews or discussions on emotive topics
    - computer games featuring preset events (success or failures, etc...);
    - Wizard of Oz experiments (the "computer" makes various decisions or announcements that will affect the participant in a way or another);
    - challenging (or sometimes "ego-threatening") tasks – such as giving a public speech, etc.

Different methods can (and should) be used in association. More naturalistic techniques provide benchmarks against which the results of more formal techniques can be evaluated.

In addition to validity, collections of primary records need to have a degree of structure. The following issues are central.

- Substantial sample sizes are needed to support relevant types of statistical analysis and particularly learning algorithms.
- Confounding occurs if samples of two emotions differ systematically not only in ways that are related to the emotional difference, but also in ways that are related to other factors (such as the task used to elicit them). Deliberate measures need to be taken to ensure that databases provide protection against it. These hinge on ensuring that records incorporate multiple methods of evoking an emotion, or open ended methods, or both.
- Factorial designs facilitate analysis, and it is to the good if collections allow the option of using factorial design. However, factorial design should not be allowed to impose procedures that are incompatible with validity.
- Repeated measures designs where a single subject produces data in several conditions also have advantages, and should be used where possible.

Descriptions of emotional states are included in primary records as defined above. Finding appropriate descriptive techniques is a major challenge. The conceptual alternatives have already been outlined (see section 3, labelling and format). A key task is to establish how useful and how economical in terms of effort various options can be made.

Cross-linkages between elements need to be thought through. For instance, attempts to use call centre data to train ‘irritation detectors’ directly would be expected to suffer from major confounding. It would be much more satisfying if training could be based on a set including varied emotions and forms of speech, and the results shown to transfer to the ethological data. Different types of training set might be expected to perform differently in that role, for instance data elicited using the facial feedback paradigm and data from interactive elicitation techniques.

Database development should be informed by that kind of possibility, so that datasets which are artificial but broad connect to datasets which are natural but narrow.

The outline in this section is at the level of principle. In practice, WP5 has limited resources, and needs to address more concrete and limited goals. Section 8 follows up on that level.

## 7. Relation to other workpackages

WP5 has strong links with four other WP's (3, 4, 6 and 10). There are issues that link it to each one individually, and there is also an issue which a joint concern for four of them (3,4,5 and 6). That issue (describing signs of emotion) is considered under a separate heading.

Relations with other workpackages are less clearcut. However, these deliverables are intended to serve as a starting point for discussion, and teams in WP5 are open to the possibility that connections may emerge in discussions following the current round of deliverables.

A general point is that database work is often regarded as ancillary – a straightforward matter of collecting the material that learning algorithms (and other routines) require to feed them. That is not a position WP5 can adopt, for reasons related to what is known in the area and what is to be found out. With regard to signs of emotion, very little is known about the fine grain, the timing, the interplay between signs in different modalities, and the influence of task and context. WP5 is about finding ways to accumulate good evidence on those issues. That is central to the science of the project, and responding effectively to the evidence that emerges is central to the technical challenge. Collecting evidence to order would be dereliction of duty.

Of course, it is equally true that WP5 is not a self-contained exercise in empirical science. It has to register the issues related to theory, technology, and applications that drive other workpackages. In that respect at least, Minsky's classic prescription for AI applies – 'heterarchy, not hierarchy'.

### WP1 Dissemination

The practicality of making data available is an important issue, and depends on the portal. Issues of storage space, access and safeguards all need to be resolved in consultation with WP1.

## A. WP3 Theories of emotion

Three key types of relationship with WP3 are discussed here. A fourth area of interaction is considered separately, because two other workpackages are equally involved.

### (i) Evaluation of theoretical proposals

Theory stands or falls on its ability to make sense of data. In that sense, the primary records collected in WP5 define the ground truth against which ideas proposed in WP3 need to be evaluated.

The relationship needs to be approached in a sophisticated way. Theory does not respond to any arbitrary set of records, nor should it. There need to be *prima facie* reasons for thinking that the phenomena belong in a domain that the theory should naturally deal with, and are not massively dependent on factors that are not the business of theory (such as arbitrary constraints on the form of interactions).

Historically, these issues have raised problems in the domain of emotion. Teams with a strong theoretical orientation have been unimpressed by data collected with a (relatively) naturalistic

ethos, because it is subject to potentially confounding factors; while their own data collection exercises, driven by theory, have been criticised as artificial and selective.

It is not a trivial task to achieve a better balance, but HUMAINE approaches the challenge with a clear sense of the issues.

### (ii) Selection of techniques

Eliciting emotional reactions in the lab is a major challenge for WP5. It is known to be difficult creating (and controlling) experimental conditions that will affect any participant (participants with different backgrounds and personal histories may or may not be affected by a given experimental setup). Furthermore, ethical constraints limit the possibility of inducing "real" (negative) emotions in research participants.

WP3 provides input on the methods that psychologists have developed to address these issues. For instance, Herrald & Tomaka (2002, p. 435) list some key methods: " Emotion elicitation methods have included asking participants to (a) recall or relive past emotional experiences [...] (b) read scenarios or vignettes [...] (c) view emotional film segments [...] (d) contort their faces in positions that match specific facial expressions associated with discrete emotions." More exhaustive reviews of methods used in this regard can be found in: Westerman, Spies, Stahl, & Hesse (1996) or Gerrards-Hesse, Spies, & Hesse (1994).

Because the objectives of HUMAINE are different from traditional laboratory research in psychology, it is to be expected that alternative techniques will have to be developed. A clear example involves the way emotion is expressed in interaction, which is central to HUMAINE, but absent from the Herrald & Tomaka list. In that context, WP3 has a role in analysing and refining new techniques.

More generally, a co-operative approach to selecting elicitation techniques is central to the issues raised in (i) above. Not all the elicitation techniques used in HUMAINE need to carry weight with theorists, but some do, and they need to be appropriately connected to other techniques (which might be chosen primarily for realism or to meet computational needs).

### (iii) Description of emotional content

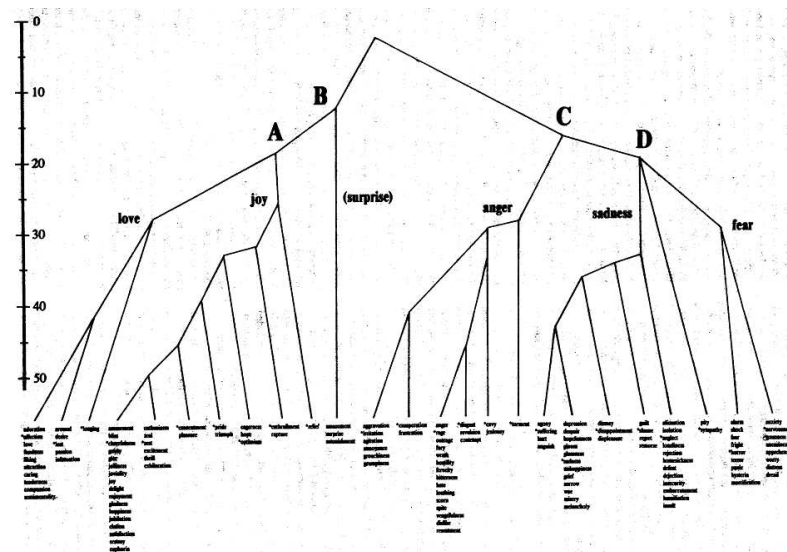
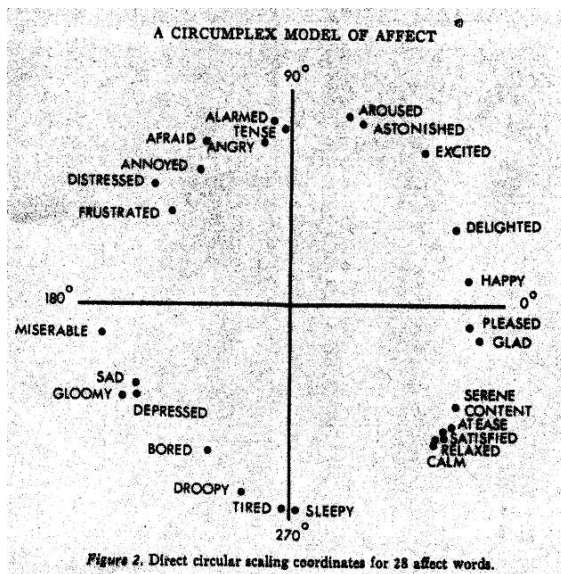
In section 6, assessments of the person's emotional state were included in the definition of primary records, and hence as a core part of the task for WP5. The problem of finding an appropriate format for these is directly related to the issue of "theories and models" that will be addressed in WP3

For some theorists/models, emotions are discrete reactions (categories), in limited number. For other approaches, emotions would be better described along various (underlying) dimensions: e.g. "introspective" (feeling-related) dimensions (such as perceived arousal, valence or control) or "cognitive-appraisal" dimensions (such as relevance or goal conduciveness of emotion-antecedent events).

In many research fields, there is a tendency to revert to so called "basic emotion" categories. Different authors have different views as to which categories should be considered "basic" (or fundamental). The table from Ortony & Turner (1990) shown in section 3 proposed a selection of emotions considered "basic" by different authors. From a theoretical point of view, the persistence of these categories is "unfortunate" in the sense that they are rather broad

categories that encompass a large variety of states that are different psychologically, and hence involve different expressions, different physiological reactions, etc.

A number of studies have addressed the question on the level of the lay representations of emotional words/labels ("semantic field studies"). In such studies, a number of emotional labels are presented to participants whose task is to rate the labels (similarity ratings or ratings on predefined dimensions). Based on this type of approach, different theorists/researchers have proposed different structures (or organisations) of emotional labels into categories or dimensions. Two such examples are given below: Russell's "circumplex" model (with underlying dimensions of valence and arousal) and Shaver's hierarchical organisation of labels into "trees".



Russell (1980, p. 1167): circumplex model (dimensional analysis)

Shaver, Schwartz, Kirson, & O'Connor (1987, p.1067) cluster analysis of similarity ratings (on 135 emotion terms)

**Figure 2. Different structures (or organisations) of emotional labels into categories or dimensions**

These approaches primarily give us indications about the way lay people categorise emotional feelings at a linguistic level. Hence they may relate quite directly to the task of capturing what a representative observer would make of the behaviour captured in a primary record. However, they are not directly relevant to describing emotion from the standpoint of an ideal objective analyst. For that role, the strongest candidate may well be the type of appraisal theory developed by Scherer and his co-workers (Scherer 1999, Scherer et al. 2001).

WP5 needs to work with WP3 to assimilate that kind of background research, and to develop schemes that are both theoretically well conceived and practically realistic.

**B. WP4: Signals to signs and vice versa**

It is not easy to define the relationship that should exist between the primary records in a database and current signal analysis techniques. It is certainly too restrictive to ask for records that allow verbal content and facial movements to be extracted using techniques that are

readily available at present – that would leave little or no scope for exploring genuinely emotional behaviour. On the other hand, it is asking too much of signal analysis to cope with records from concealed cameras and microphones that human beings struggle to decipher.

A core part of the issue is how to address problems that arise from allowing human beings to behave in the ways that are characteristic of spontaneous emotionally coloured behaviour. People who are in an emotional state put their hands over their mouths, turn their heads, allow their voices to trail off or rise to a shout, speak in whispery or creaky voices, intersperse their speech with giggles and moans, and so on. All of these make standard signal processing problems more difficult – finding face boundaries, locating features from an image, avoiding clipping or low signal to noise ratio, finding F0, distinguishing speech and non-speech intervals, and so on.

It is tempting to respond to these issues by asking WP3 to provide records where they do not arise. That makes little sense in the long run: they are part of the territory, and the ideal solution is to find ways of dealing with them within WP4. In practice, within the lifetime of HUMAINE, there is likely to be some mixture of the two.

Recordings for the ongoing ERMIS project represent a useful model. They reflect long interactions between teams with backgrounds in psychology, sociolinguistic fieldwork and signal processing. Similar discussions should probably be part of the preparation for new types of recording scenario.

### **C. WP6: Interaction**

Presumably teams concerned with emotion in interaction need primary records that show how signs of emotion appear in interactions. There should presumably be discussion about the kinds of interaction that are of particular significance to interaction modellers, and the methods available to provide appropriate records. On one hand, it does not make sense to invest large amounts of time and effort to record interactions with very specific characteristics that make them ideally suited to the task of developing a particular system, but largely irrelevant to any other type of development. On the other, it does not make sense to record interactions without reference to the kind of modelling that is being undertaken elsewhere in HUMAINE.

### **D. WP10: Ethics**

Both the collection and the storage of data are fraught with ethical issues. As a result, WP5 needs to work closely with WP10.

### **E. WP 3, 4, 6 Shared task to develop appropriate labellings**

One of the central challenges for HUMAINE is to develop an ‘interface language for signs of emotion’ (ILSE for short). The natural ideal is a single ILSE that would define the boundary between WP4 and WP6, in both directions. It would describe the signs that signal processing concluded were present as a result of analysis, and the signs that a theory of interaction indicated should be given to achieve a particular desired result. It would also define how the signs expressed in a primary record should be annotated.

Unfortunately, we do not have a satisfying ILSE, or any of its components. Systems like FACS in the facial domain and ToBI in the domain of prosody are sometimes treated as if they were, but that is to misunderstand their function. FACS is quite openly an approximate coding based partly on anatomy and partly on simplifications that are convenient for a human coder. To say that is not to detract from its success as an approximate coding, but a useful approximation should not be confused with the last word. In the domain of prosody, the dominant coarse coding is the ToBI system, which was designed to highlight information relevant to the linguistic content of speech; it is, to put it mildly, surprising that anyone ever thought a system designed with that purpose in mind would necessarily provide a good way to capture non-linguistic content. In the domain of gesture, there are notations which originate in schemes developed to convey one dancer's movement to another dancer. They could not be expected to define subtleties of movement which should be second nature to a dancer.

Research on emotion can and should take all of these existing coding schemes as useful inputs. However, it is bound to look beyond them to schemes that can capture in depth the variations that are relevant to conveying emotion. To do that, it needs to engage with empirical data without preconceptions. Superimposing preconceived sign systems on the data is not only wasted effort, it is counterproductive. It can only reinforce the grip of codings that we know are partial.

For that reason, we are not inclined to accept that WP5 should invest large efforts in applying current methods of encoding signs of emotion. It should provide the records that underpin attempts to devise richer codings, but it should invite other workpackages to take up the task of devising ways to structure the information that is in the records.

The other workpackages most obviously involved are WP3, WP4, and WP6. Our current position is that it is their role to develop schemes that code relevant features in an appropriate way. Until those schemes exist, it will not be clear how signs of emotion should be coded in a database. That will be the subject of discussion following the submission of this deliverable.

## 8. Preliminary ideas about an exemplar

WP5 will almost inevitably generate a single exemplar – that is, a library of data to be accessed by research on emotional states and their expression. The library is bound to include four broad types of archive.

**1. Historical:** Existing databases that are relevant to the work of HUMAINE should be assembled or otherwise made accessible to network members. Tables 1-3 represent an important step towards that. There will be an ongoing task of adding to the archive as new material is developed or discovered outside the network.

**2. Reactive:** Discussions have already been opened with other workpackages about needs arising from their proposals. Realistic assessment of the prospect of providing necessary data will necessarily be a factor in the way other WPs develop.

Some examples of the kinds of material that are likely to come under that heading are already clear. They include

- annotated sets of recordings designed to allow evaluation of signal processing routines for tasks like extracting pitch from emotional speech, or facial points from video
- gesture-rich samples for to be used in the development of coding systems for the interaction workpackage
- training material in both areas is a possibility, but that raises the issue of quantity: the cost of providing enough material for a particular purpose has to be weighed against other demands.

It should be noted that pre-existing material may meet needs in whole or in part. If so, it may be inefficient to generate completely novel material, though not necessarily – generating new material for one function may mean that other needs can be met in a way that is more integrated and satisfying than would otherwise be possible.

**3. Proactive:** WP5 has a long term remit of assembling material that will facilitate fundamental research in the area as a whole. At root, that means creating a situation where research is not prevented from addressing the most important and challenging questions because it lacks the data to do so. That aspect of the workpackage is discussed at more length below.

**4. Metaknowledge:** WP5 will necessarily assemble procedures, systems, summary analyses, and so on as well as records that make up a database in the standard sense.

It is in the spirit of an exemplar to provide these different elements as a cohesive resource: it acknowledges the real diversity of data-related needs that research in the area has.

The proactive archive is the most innovative component of the library, and the one where we propose most effort should be invested. There needs to be extensive consultation before plans for it are finalised, but the following outline provides a starting point that we will use for discussion.

## Emotional scope

We propose as a design principle that there should be a pre-stated framework defining the emotional and emotion-related states that should ideally be covered in a database designed to support research on emotion-sensitive interfaces. The point has been made repeatedly that standard lists of ‘primaries’ are not an appropriate framework. Cowie et al (1999) attempted to create a more appropriate list by asking naïve participants to identify words that they would include in a ‘basic English emotion vocabulary’. That exercise is a useful model, but it needs to be repeated with a better theoretical base and a clearer emphasis on the application (which will, for instance, make what Scherer called ‘interpersonal stances’ important in a way that was not reflected in the task set by Cowie et al). Table 4 below shows a list, based on that study and key interpersonal states, that one might reasonably feel an interface should be able to detect and perhaps simulate if it was to interact effectively with a person showing a natural range of emotional and emotion-related states.

Table 4. Proposed list of emotion and emotion-related states

admiration	cold anger	disagreeableness	fear	indifference	pride	shock
affection	coldness	disappointment	friendliness	interest	relaxation	stress
amusement	confidence	disapproval	greed	jealousy	relief	surprise
annoyance	contempt	disgust	guilt	mockery	resentment	sympathy
anxiety	contentment	distraction	hopeful	nervousness	sadness	wariness
		effervescent				
approval	cruelty	happiness	hot anger	neutrality	satisfaction	weariness
boredom	despair	embarrassment	hurt	panic	serenity	worry
calm	determination	excitement	impatience	pleasure	shame	

That list provides a first estimate of the emotional scope that a genuinely satisfying interaction-oriented database should aim at. One of the tasks of WP5 is to move from that list to one that can act as an agreed framework.

It is not suggested that WP5 should fully represent all of these states. The point is rather that the choice of states to represent should be informed by a well developed overview, and that gaps should be acknowledged explicitly.

## Levels of representativeness

As before, we distinguish three levels of representativeness – natural, induced, and acted. These are very broad categories, and the cutoffs between them are not always sharp.

We take it as axiomatic that an exemplary archive must include material that is natural or at least not elicited in a highly directed way. One of the touchstones of naturalism is that the people involved set their own emotional agenda – they are not being directed by an experimenter into states that conform to the experimenter’s images. Naturalistic material acts as the point of reference for material of any other sort.

To use naturalistic material as a point of reference, it is necessary to develop techniques for comparing material at different levels of representativeness. We have recently shown that simple approaches can be quite effective: raters can use a sliding scale to record how confident they are at any given time whether material is natural or acted. That kind of technique is much more searching than the traditional approach of asking whether people can recognise the relevant emotion.

At the other extreme, one would prefer in principle to avoid acted material. However, there is no doubt that many of the states in Table 4 are much easier to act than to sample in any other way. Hence, it seems reasonable to allow that acted material should be incorporated, with some provisos. In particular, the way acted material is collected should make it possible to draw comparisons between acted performance and more representative material for at least some emotional states.

The core effort that we propose to invest in selecting and developing appropriate induction techniques. The precise kinds of option to be considered we discuss later. Selection is related to the issue of emotional scope, considered above; but also to the issue of contexts, considered next.

## Contexts

We propose as a design principle that an exemplary database should reflect the likelihood that emotion will be expressed in different ways in different contexts. A football fan might experience pleasure at his team scoring a goal whether he was on the terraces or driving a car, but one would expect him to express it differently in the two cases. The example illustrates a point which is quite general, but rarely discussed. We distinguish six broad types of context that might be expected to influence the expression of emotion.

Individual Many standard induction techniques are applied to people on their own in a laboratory (see section 6). They may induce genuine emotion, but the way it is expressed (if it is expressed at all) is likely to be quite impoverished.

Sedentary human interaction A major context for the expression of emotion involves two people sitting talking to each other. The signs of emotion in that situation are likely to be very different from the signs given by an isolated individual, and so (perhaps less obviously) is the task of inducing emotion, because interactions between the parties tend to take control of the emotional trajectory very quickly.

Mobile human interaction This is distinguished from the previous context because it introduces the possibility of whole-body movements.

Social facilitation The example of the football fan on the terraces is a reminder of the well-known phenomenon of social facilitation – the presence of a group can amplify various kinds of action, and the expression of emotion certainly appears to be one of them.

Task-dominated The football fan driving is an example of this kind of context – the person's scope for expressing emotion is constrained by the fact that he or she is engaged in a task (or should be). Note, though, that a distinctive kind of sign may well appear, that is, change in performance of the task. Surface manifestations may differ enormously depending on the task, but there may also be effects which are general and important, such as interference with attention.

Machine interaction Interactions with relatively crude machines (such as automatic call centres) involve rather strange combinations of the types above.

Sedentary interaction is likely to be the main source of data, but it needs to be supplemented by other types, and it needs to be established whether there are substantial differences.

### **Modalities**

We propose as a design principle that records should include both audio and visual information with few exceptions. Audio-visual data should be supplemented by performance measures (error rate, reaction time, etc) where there is a task involved, and in some cases by standard visceral measures (heart rate, skin conductance, and temperature, possibly with the addition of chest capacity).

We do not propose to use brain scan techniques: they are likely to interfere with the expression of emotion, are costly in terms of effort and resources, and the return is not clear.

The effort to be invested in capturing gesture remains unclear, and discussions with other workpackages are needed.

### **Segmentation**

Practically, databases need to be divided into units of manageable size. We propose to standardise on beb (baseline-emotional-baseline) clips lasting of the order of a minute rather than on turns or phrase-like units (which tend to be shorter) or complete interactions (which tend to be longer).

### **Cultures and individuals**

HUMAINE is committed to sampling across genders and cultures. We also propose as a design principle that wherever possible, records should be made of a single individual undergoing multiple states in multiple contexts.

### **Scale**

Explicit provision has been made for reactive data collection, which may have to involve large sample sizes. For the proactive archive, the intention is not to generate large quantities of data per cell. Instead, the principle we propose is that there should be enough data per cell to convey the kind of expressive behaviour (etc) that a technique is likely to generate, but no more. Given that, people who need large amounts of data of a particular type can generate their own.

## Labelling

Two tasks fall under this heading – providing assessments of the person’s emotional state, and providing codings of the signs that reveal it.

Assessments of the person’s emotional state were identified as part of a primary record in section 6, and it is not in doubt that WP5 is obliged to ensure that records include that kind of information. Coding signs of emotion is a very different issue, and the general position proposed here is that it should not be regarded as a task that belongs in WP5.

Section 4B(i) reviewed the main approaches to describing emotion. As an opening position, we propose to extend the approach used in the Belfast Naturalistic Database. Three main types of coding would be provided

(a) each beb clip would have associated with it descriptions based on the terms listed in Table 4, with a qualifier specifying the level of emotionality (weak/moderate/strong).

(b) each beb clip would have associated with it a dimensional description based on FEELTRACE or a comparable system.

(c) where feasible, each beb clip would have associated with it a self-rating of emotion. It is not obvious how best to do that. In studies of boredom, we have used simple self-rating questionnaires that can be administered at the time of induction. In other contexts, it may be better to use ‘review’ techniques (where a person observes records of his/her own behaviour and rates his/her own states using a combination of observation and self-knowledge).

This is an area where it is clear that other options exist, and should be discussed with people based in other workpackages.

In contrast, our judgment is that WP5 should not attempt to provide codings of the signs that reveal emotion. The key reasons for that judgment were covered in Section 7. Essentially, it would be confusing the job of collecting specimens with the job of devising a taxonomy. It is most definitely part of the task in WP5 to assemble specimens on which taxonomists can work. It is a different matter to embed in the database descriptions based on a taxonomy that seems likely to be inherently limited.

## Selecting and developing appropriate induction techniques

The consortium has at its disposal a considerable range of techniques for inducing emotional and emotion-related states. They can be grouped for convenience into four broad types

Open ended techniques allow for emotion to fluctuate according to accidents of interaction and choices made by the person or people involved. These include the SALAS approach and non-directive techniques developed in Belfast and Bari, and interactions with the AIBO robot studied at Erlangen.

Application-oriented techniques simulate situations in which emotion-sensitive technology may be practically important. A major category here involves simulated interaction with automated services, an approach that has been developed at Erlangen and explored in Belfast. A second category is driving behaviour, which has also been studied in Belfast.

Targeting specific emotions The Geneva group has developed sophisticated methods of eliciting specific emotions through computer games. The team at Tel Aviv have used recall, and more recently have also developed computer game techniques. The Belfast team have focused on eliciting emotion in the context of tasks, particularly tasks which involve speech. These include tasks designed to elicit boredom and excitement, and the use of music to manipulate arousal level in the context of driving.

Gesture-oriented Few of the techniques listed above are likely to elicit large-scale gestures. Techniques which address that issue include the SENTOY project in which emotion is expressed by manipulating a puppet (Paiva et al 2003), and the ongoing ‘shit’ study devised in KTH.

Our current proposal is to concentrate on developing open ended techniques and on widening the range of methods available to target specific emotions, not only with respect to the emotions to be elicited (see Table 4), but also with respect to the contexts in which they are expressed.

## 9. Conclusions and Way Forward

One of the main functions of this report has been to make it clear that developing genuinely satisfying emotion databases is a very large task, well beyond the scope of HUMAINE. In fact, establishing a shared understanding of the scale of the task is one of the key contributions that HUMAINE can make.

Because the full task is beyond the scope of HUMAINE, compromises need to be identified and agreed. The previous section outlined the kinds of compromise that currently seem reasonable to us.

The function of that outline is to provide a basis for discussion over the coming months. Two reports are to follow before final agreement is reached, one at the end of 2004, the other in mid 2005. They will reflect the inputs of other partners, both within WP5 and in other workpackages. The negotiation has to be serious, providing leeway not only for resolution of detail, but also for strategic shifts if there are good reasons to make them.

## 10. References

### A. References cited in the text

Abelin, A., Allwood, J., 2000. Cross linguistic interpretation of emotional prosody. In: *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 110-113.

Amir, N., Ron, S., Laor, N. 2000. Analysis of an emotional speech corpus in Hebrew based on objective criteria. In: *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 29-33.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. ICSLP 2002*, Denver, Colorado, Sept. 2002.

AR Face Database ([http://rvl1.ecn.purdue.edu/~aleix/aleix\\_face\\_DB.html](http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html))

Bachorowski, J.A., Owren, M.J. 1995. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychol. Sci* 6(4), 219-224.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70 (3), 614-636.

Batliner A., Fischer K., Huber R., Spilker J. , Nöth E. 2003. How to find trouble in communication. *Speech Communication* 40, 117-143.

- Batliner, A., Hacker, C., Steidl, Noth, E., D'Arcy, S., Russell, M., Wong, M. 2004a. "You stupid tin box" - children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. *Proc. LREC 2004*.
- Batliner, A., Hacker, C., Steidi, S., Noth, E., Haas, J. 2004b. From emotion to interaction: lessons learned from real human-machine dialogues. *Proceedings of Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004*, Kloster Irsee, Germany, June 14-16, 2004, Springer-Verlag, pp. 1-12.
- Campbell, W. N. 2002. Recording Techniques for capturing natural everyday speech. *Proc. LREC-2002*, Las Palmas.
- Chung, S., 2000. Expression and Perception of Emotion extracted from the Spontaneous Speech in Korean and English (ILPGA, Sorbonne Nouvelle University, Paris, France). Thesis available online at <http://people.ne.mediaone.net/sangikoh/soojinchung.htm>
- Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Bunnell, T., Idsardi, W. (Eds.), *Proc. Fourth ICSLP*, 3-6 October 1996, Philadelphia, pp. 1989-1992.
- Cowie, R., Douglas-Cowie, E., Apolloni, B., 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 Press, 109-114.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine* 18 (1), 32-80.
- Cowie R., Cornelius R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40, 5-32.
- Cowie, R., McGuiggan, A., McMahon, E., Douglas-Cowie, E. 2003. Speech in the Process of Becoming Bored. *Proc. 15<sup>th</sup> ICPhS*, Barcelona.
- CVL Face Database (<http://www.lrv.fri.uni-lj.si/facedb.html>)
- Devillers, L., Vasilescu, I., Lamel, L. 2002: Annotations and Detection of Emotion in a Task-oriented Human-Human Dialog Corpus. *ISLE Workshop*, Edinburgh, December 2002.
- Devillers, L., Rosset, S., Maynard, H., Lamel, L. 2002. Annotations for Dynamic Diagnosis of the Dialog State, *LREC 2002*, Las Palmas, May 2002.
- Devillers, L., Vasilescu, I. 2003. Prosodic cues for emotion characterization in real-life spoken dialogs. *Proc Eurospeech*, Geneva, September 2003.
- Devillers, L., Vasilescu, I., Mathon, C. 2003. Acoustic cues for perceptual emotion detection in task oriented human-human corpus. *Proc ICPhS*, Barcelona, 2003.
- Devillers, L., Vasilescu, I., Lamel, L. 2003: Emotion Detection in a task-Oriented Dialog Corpus. *IEEE International Conference on Multimedia*, ICME, Baltimore, July 2003.
- Devillers, L., Vasilescu, I., Vidrascu, L. 2004. Anger versus Fear detection in recorded conversations. *Speech Prosody*, Japan, March 2004.

Devillers, L., Vasilescu I. 2004. Reliability of Lexical and Prosodic Cues in two Real-life Spoken Dialog Corpora *LREC 2004*.

Douglas-Cowie, E., Cowie, R., Schroeder, M., 2000. A new emotion database: considerations, sources and scope. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 39-44.

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

Ekman, P., Friesen, W., 1978. *The Facial Action Coding System*. Consulting Psychologists' Press, San Francisco, CA.

Engberg., I. S., Hansen, A. V., Andersen, O., Dalsgaard, P., 1997. Design, recording and verification of a Danish Emotional Speech Database. *Proc. Eurospeech '97*, Rhodes (Greece), 1997.

Fernandez, R., Picard, R., 2000. Modeling drivers speech under stress. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 219-224.

Fernandez, R., Picard, R. W. 2003. Modeling drivers' speech under stress. *Speech Communication* 40 (1-2), 145-159.

France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering* 47 (7), 829-837.

Frolov, M., Milovanova, G., Lazarev, N., Mekhedova, A., 1999. Speech as an indicator of the mental status of operators and depressed patients. *Human Physiology* 25 (1), 42-47.

Gerrards-Hesse, A. Spies, K., Hesse, F. W. 1994. Experimental induction of emotional states and their effectiveness: A review. *British Journal of Psychology*, 85(1), 55-78.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., Horton, D. 1995. Representation of prosodic and emotional features in a spoken language database. *Proc. XIIIth ICPhS*, Stockholm, vol. 1, pp. 242-245.

Greasley, P., Sherrard, C., Waterman, M., 2000. Emotion in Language and Speech: Methodological Issues in Naturalistic Approaches. *Language and Speech* 43, 355-375.

Groningen corpus S0020 ELRA ([www.icp.inpg.fr/ELRA](http://www.icp.inpg.fr/ELRA))

Hansen, J., Bou-Ghazale, S. 1997. Getting started with SUSAS: A Speech Under Simulated and Actual Stress Database. *Proc. Eurospeech 1997*, Rhodes, Greece, vol. 5, 2387-2390.

Herrald, M. M., & Tomaka, J. (2002). Patterns of emotion-specific appraisal, coping, and cardiovascular reactivity during an ongoing emotional episode. *Journal of Personality and Social Psychology*, 83(2), 434-45.

Huttar, G.L. 1968. Relations between prosodic variables and emotions in normal American English utterances. *J. Speech Hear. Res.* 11, 481-487.

Iriondo, I., Guaus, R., Rodriguez, A., Lazaro, P., Montoya, N., Blanco, J., Beradas, D., Oliver, J., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression

in Spanish using speech synthesis techniques. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 161-166.

ISLE Project ( <http://isle.nis.sdu.dk/>, IST project IST-1999-10647)

(JAFFE) The Japanese Female Facial Expression Database  
(<http://www.mis.atr.co.jp/~mlyons/jaffe.html>)

James, W., 1884. What is emotion? *Mind* 9, 188-205.

Johannes, B., Salnitski, V., Gunga, H-C., Kirsch, K. 2000. Voice stress monitoring in space – possibilities and limits. *Aviation, Space and Environmental Medicine* 71, 9, section II, A58-A65.

Johns-Lewis, C., 1986. Prosodic differentiation of discourse modes. In Johns-Lewis, C., (ed.), *Intonation in Discourse*. San Diego, College-Hill Press, pp. 199-220.

Juslin, P.N., Laukka, P. 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion* 1 (4), 381-412.

Kienast, M., Sendlmeier, W.F., 2000. Acoustical analysis of spectral and temporal changes in emotional speech. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 92-97.

Kuroda, I., Fujiwara, O., Okamura, N., Utusuki, N., 1979. Method for determining pilot stress through analysis of voice communication. *Aviat. Space Envir. Md.* 47, 528-533.

Lazarus, R. S. 1994. The stable and the unstable in emotion. In Ekman, P., Davidson, R. J. (eds.), *The nature of emotion. Fundamental questions*. Oxford University Press, Oxford, pp. 79-85.

Lee, C., Narayanan, S. 2003. Emotion recognition using a data-driven fuzzy inference system. *Proc. Eurospeech 2003*, Geneva.

Leinonen, L., Hiltunen, T. 1997. Expression of emotional-motivational connotations with a one-word utterance. *JASA* 102 (3), 1853-1863.

Levenson, R. W., 1992. Autonomic nervous system differences among emotions. *Psychological Science* 3, 23-27.

McGilloway, S., 1997. Negative symptoms and speech parameters in schizophrenia. PhD thesis, Queen's University, Belfast.

McMahon, E., Cowie, R., Kasderidis, S., Taylor, J., Kollias, S. 2003. What chance that a DC could recognise hazardous mental states from sensor outputs? *Tales of the Disappearing Computer*, Santorini.

Mitchell, C., Menezes, C., Williams, J., Pardo, B., Erickson, D., Fujimura, O. 2000. Changes in syllable and boundary strengths due to irritation. *Proc. ISCA ITRW on Speech and Emotion*, 5-7 September 2000, Textflow, Belfast, 98- 103.

Mozziconacci, S. 1998. Speech variability and emotion: Production and perception. Ph.D. thesis, Technical University Eindhoven, Eindhoven.

- Murphy, C. 2002. Automatic recognition of spoken emotion using audio signal processing. Unpublished undergraduate thesis, Dept of Electrical and Electronic Engineering, University College Dublin.
- Nakatsu, R., Tosa, N., Nicholson, J. 1999. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Proc. IEEE International Workshop on Multimedia Signal Processing*, 439-444.
- Nogueiras, A., Moreno, A., Bonafonte, A. & Mariño, J. 2001. Speech emotion recognition using hidden markov models. *Proc. Eurospeech 2001*, Aalborg, Denmark.
- Ortony, A., Clore, G., Collins, A., 1988. *The cognitive structure of emotions*. Cambridge University Press, Cambridge.
- Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, 97, 315-331.
- Oudeyer, P-Y. 2003. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2), 157--183.
- Paeschke, A., Sendlmeier, W.F., 2000. Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 75-80.
- Paiva, A., Costa, M., Chares, R., Piedade, M., Mourao, D., Sobral, D., Hook, K., Andersson, G., Bullock, A. 2003. SenToy: an Affective Sympathetic Interface. *Internat Journal of Human Comuter Srudies* 59 (1-2), 27-235.
- Pereira, C., 2000. Dimensions of emotional meaning in speech. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 25-28.
- Picard, R., Vyzas, E., Healey, J. 2001 Toward machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Trans Patt Analysis & Machine Intell* 23, 1175-1191.
- Plutchik, R., 1984. *Emotions: A general psychoevolutionary theory*. In Scherer, K. R., Ekman, P. (eds.), *Approaches to emotion*. Erlbaum, Hillsdale, NJ, pp. 197-219.
- Plutchik, R., Conte, H 1997. *Circumplex models of Personality and Emotions*. Washington: APA.
- Polzin, T. S., Waibel, A., 2000. Emotion-sensitive human-computer interfaces. *Proc ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 201-206.
- The Psychological Image Collection at Stirling (<http://pics.psych.stir.ac.uk/>)
- Roach, P., Stibbard, R., Osborne, J., Arnfield, S., Setter, J., 1998. Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association* 28, 83-94.
- Roach, P., 2000. Techniques for the phonetic description of emotional speech. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 53-59.

- Roessler, R., Lester, J., 1979. Vocal pattern in anxiety. In Fann, W., Pokorny, A., Koracau, I., Williams, R. (eds.), *Phenomenology and treatment of anxiety*. Spectrum, New York.
- Roseman, I. J., 1991. Appraisal determinants of discrete emotions. *Cognition and Emotion* 5, 161-200.
- Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.
- SALAS database: see ERMIS project <http://www.image.ntua.gr/ermis/> IST-2000-29319, deliverable D09 'Final version of non-verbal speech parameter extraction module'.
- Scherer, K. S. 1994. Toward a concept of modal emotions. In Ekman, P., Davidson, R. (eds.) *The nature of emotion: fundamental questions*. Oxford University Press, pp. 25-31.
- Scherer, K. R., 1999. Appraisal theory. In Dalglish, T., Power, M. (eds.), *Handbook of Cognition and Emotion*. John Wiley, New York, pp. 637-663.
- Scherer, K. R. 2000. Emotion effects on voice and speech: paradigms and approaches to evaluation. [www.qub.ac.uk/en/isca/proceedings/pdfs/scherer.pdf](http://www.qub.ac.uk/en/isca/proceedings/pdfs/scherer.pdf)
- Scherer, K., Schorr, A., Johnstone, T., (eds). 2001. *Appraisal Processes in Emotion*. New York, Oxford University Press.
- Scherer, K., Ceschi, G., 1997. Lost luggage emotion: a field study of emotion – antecedent appraisal. *Motivation and Emotion* 21, 211-235.
- Scherer, K., Ceschi, G., 2000. Studying affective communication in the airport: The case of lost baggage claims. *Personality and Social Psychological Bulletin* 26 (93), 327-339.
- Schiel, F., Steininger, S., Türk, U. 2002. The SmartKom Multimodal Corpus at BAS. *Proc LREC 2002*, Las Palmas, Gran Canaria, Spain, pp. 200-206.
- Schlosberg, H., 1954. A scale for judgment of facial expressions. *Journal of Experimental Psychology* 29, 497-510.
- Shaver, Ph., Schwartz, J., Kirson, D., O'Connor, C. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology* 52, 1061-1086.
- SMARTKOM: [www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom](http://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom)
- Stein, N., Oatley, K., 1992. Basic emotions: Theory and measurement. *Cognition and Emotion* 6, 161-168.
- Steininger, S., Schiel, F., Glesner, A. 2002. Labeling Procedures for the Multi-modal Data Collection of SmartKom. *Proc LREC 2002*, Las Palmas, Gran Canaria, Spain.
- Steininger, S., Schiel, F., Dioubina, O., Rabold, S. 2002. Development of User-State Conventions for the Multimodal Corpus in SmartKom. *Proc Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002*, Las Palmas, Gran Canaria, Spain, pp. 33-37.

- Stibbard, R.M. 2000. Automated extraction of ToBI annotation data from the Reading/Leeds Emotional Speech Corpus. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, 60-65.
- Stibbard, R. 2001. Vocal expression of emotions in non-laboratory speech. Unpublished PhD thesis, University of Reading, UK.
- Sulc, J.1977. To the problem of emotional changes in the human voice. *Activitas Nervosa Superior* 19, 215-216.
- SUSAS Database (Speech under simulated and actual stress) <http://wave ldc.upenn.edu/catalog/> LDC catalog no. LDC99S78.
- Tolkmitt, F., Scherer, K., 1986. Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance* 12 (3), 302-33.
- van Bezooijen, R., 1984. *Characteristics and recognizability of vocal expressions of emotion*. Foris Publications, Dordrecht.
- Velten, E. 1968. A laboratory task for induction of mood states. *Behaviour Research and Therapy* 6, 473-482.
- Yacoub, S., Simske, S., Lin, X., Burns, J. 2003. Recognition of emotions in interactive voice response systems. *Proc. Eurospeech 2003*, Geneva.
- Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. 2001 "DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection". *Proc. Eurospeech 2001*.
- Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. 1996. Relative effectiveness and validity of mood induction procedures: a metaanalysis. *European Journal of Social Psychology*, 26(4), 557-580
- 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 Press, New York.
- Williams, C.E. & Stevens, K.N. 1969. On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Md.* 40, 1369-1372.
- Williams, C.E. & Stevens, K.N., 1972. Emotions and speech: Some acoustical correlates. *J. Acoust. Soc. Amer.* 52, 1238-1250.
- Wundt, W., 1903. *Grundzuge der Physiologischen Psychologie vol 2*. Engelmann, Leipzig. (Original published 1874)

## B. General bibliography

Alter, K., Rank, E., Kotz, S., Toepel, U., Besson, M., Schirmer, A., Friederici, A. D. 2000. Accentuation and emotions two different systems? *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 138-42.

Andreasen, N., Alphert, M. & Merrill, J. 1981. Acoustic analysis: an objective measure of affective flattening. *Arch. Gen. Psychiatry* 38, 281-285.

Arnfield, S., Roach, P., Setter, J., Greasley, P., Horton, D., 1995. Emotional stress and speech tempo variation. *Proc. ESCA-NATO Tutorial and Research Workshop on Speech Under Stress*, Lisbon, pp. 13-15.

Aubergé, V., Cathiard, M. 2003. Can we hear the prosody of smile? *Speech Communication*, 40 (1-2), 87-97.

Bachorowski, J.A. 1999. Vocal expression and perception of emotion. *Current Directions in Psychol. Sci* 8(2), 53-57.

Bird, S., Harrington, J. (Eds.), 2001 Special Issue on Speech Annotation and Corpus Tools. *Speech Communication* 33.

Bonner, M. R., 1943. Changes in the speech pattern under emotional tension. *American Journal of Psychology* 56, 262-273.

Boozer, A., Seneff, S., Spina, M. 2003. Towards recognition of emotional speech in human-computer dialogues. Abstract. MIT Laboratory for Computer Science.  
[www.csail.mit.edu/research/abstracts/abstracts03/interfaces-applications/03boozer.pdf](http://www.csail.mit.edu/research/abstracts/abstracts03/interfaces-applications/03boozer.pdf)

Cahn, J. 1990. The generation of affect in synthesised speech. *J. Amer. Voice I/O Society* 8, 1-19.

Campbell, W. N., Black, A.W., 1996. CHATR – a multi-lingual speech re-sequencing synthesis system. Technical Report of IEICE SP96-7, pp. 45-52.

Campbell, W. N., Marumoto, T., 2000. Automatic labelling of voice quality in speech databases for synthesis. *Proc. ICSLP 2000*, Beijing., vol. IV, pp 468-471.

Campbell, N. 2000. Databases of emotional speech. *Proc. ISCA ITRW Speech and Emotion*, Newcastle, Ireland, Sept. 2000, 34-39.

Cauldwell, R. Where did the anger go? The role of context in interpreting emotion in speech. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 127-131.

Cowie, R., Douglas-Cowie, E., Sawey, M., 1995. A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures). *Proc. ICPhS 1995*, Stockholm, vol. 3, pp. 278-281.

Cowie, R., Douglas-Cowie, E., Wichmann, A., Hartley, P., Smith, C., 1999. The prosodic correlates of expressive reading. *Proc. 14th ICPhS*, San Francisco, 1-7 August 1999, Berkeley, University of California, pp. 2327-2330.

- Cowie R., Douglas-Cowie, E., Schroeder M., 2000. *Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework*, Newcastle, N. Ireland, 5-7 September 2000. Textflow, Belfast. Proceedings online at [www.qub.ac.uk/en/isca/index.htm](http://www.qub.ac.uk/en/isca/index.htm)
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. & Schroeder, M. 2000. Feeltrace: an instrument for recording perceived emotion in real time. *Proc. ISCA ITRW Speech and Emotion*, Newcastle, Ireland, Sept. 2000, 19-24.
- Cowie, R., Douglas-Cowie, E., Wichmann, A. 2002. Prosodic correlates of skilled reading: Fluency and expressiveness in 8- 10 year old readers. *Language and Speech* 45(1), 47-82.
- Cummings, K., Clements, M. 1995. Analysis of the glottal excitation of emotionally styled and stressed speech. *JASA* 98 (1), 88-98.
- de Gelder, B., Vroomen, J. 2000. The perception of emotions by ear and eye. *Cognition and Emotion* 14, 289-311.
- de Gelder, B., Vroomen, J. 2000. Bimodal emotion perception: integration across separate modalities, cross-modal perceptual grouping, or perception of multimodal events? *Cognition and Emotion* 14, 321-324.
- Douglas-Cowie, E., Cowie, R., Campbell, N. (Guest Editors). Special double issue of *Speech Communication* 'Speech and Emotion', volume 40, nos-1-2, (Elsevier), 257pp.
- Douglas-Cowie, E., Cowie, R., Schröder, M. 2003. The description of naturally occurring emotional speech. *15th ICPHS*, Barcelona, Spain, 2877-2880.
- Dybkjær, L., Berman, S., Kipp, M., Olsen, M. K., Pirrelli, V., Reithinger, N., Soria, C. 2001. ISLE Natural Interactivity and Multimodality Working Group Deliverable D11.1. Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data., <http://isle.nis.sdu.dk/reports/wp11/>
- Ekman, P. Friesen, W., 1969. The repertoire of non verbal behavior: categories, origins, usage and coding. *Semiotica*, 1, 49-98.
- Ekman, P. 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* 115, 268-287.
- Ekman, P. 1999. Basic Emotions . In Dalglish, T., Power, M. (eds), *Handbook of Cognition and Emotion*. New York, John Wiley pp. 45-60.
- Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40 (1-2), 189-212.
- Haddad, D., Ratley, R., Walter, S., Smith, M. 2002. Investigation and Evaluation of Voice Stress Analysis Technology. Final Report US Dept of Justice Report NCJ Number 193832.
- Hansen, J. H. L. & Womack, B. D., 1996. Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing* IV(4), 307- 313.
- Harré, R. (ed), 1986. *The Social Construction of Emotions*. Oxford, UK, Blackwell.

- Hargreaves, W., Starkweather, J., Blacker, K. 1965. Voice quality in depression. *Journal of Abnormal Psychology* 70, 218-220.
- Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 167 – 172.
- Johnstone, T. Banse R., Scherer, K. R. 1995 Acoustic profiles in prototypical vocal expressions of emotion. *Proc ICPHS*, Stockholm 1995; vol 4 pp 2-5.
- Johnstone, T., Scherer, K. R. 2000. Vocal communication of emotion. In Lewis, M. & Haviland, J. (eds.), *Handbook of emotion*, 2nd ed., Guilford, New York, 220-235.
- Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K., 1998. Within speaker variation due to induced stress. In Branderud, P., Traunmuller, H., (eds.), *Proc of Fonetik – 98*. The Swedish Phonetics Conference, Stockholm University, May 27-29, 1998, pp. 150-153.
- Klasmeyer, G. 2000. An automatic description tool for time-contours and long-term average voice features in large emotional speech databases. *Proc. ISCA Workshop Speech and Emotion*, Newcastle, Northern Ireland.
- Knudsen, M., Martin, J-C., Dybkjær, L., Berman, S., Bernsen, N., Choukri, K., Heid, U., Kita, S., Mapelli, V., Pelachaud, C., Poggi, I., van Elswijk, P., Wittenburg, P. 2002. Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. *ISLE Natural Interactivity and Multimodality. Working Group Deliverable D8.1* February 2002. <http://isle.nis.sdu.dk/reports/wp8/>
- Knudsen, M., Martin, J-C., Dybkjær, L., Ayuso, M., Bernsen, N., Carletta, J., Heid, U., Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., Wittenburg, P. 2002. Survey of Multimodal Annotation Schemes and Best Practice. *ISLE Natural Interactivity and Multimodality. Working Group Deliverable D9.1*. February 2002. <http://isle.nis.sdu.dk/reports/wp9/>
- Kwon, O., Chan, K., Hao, J., Lee, T-W. 2003. Emotion recognition by speech signals. *Proc. Eurospeech 2003*, Geneva, 125-128.
- Ladd, D.R., Silverman, K.E.A., Tolkmitt, F., Bergmann, G. & Scherer, K.R. 1985. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *J. Acoust. Soc. Amer.* 78, 435-444.
- Ladd, D. R., Scherer, K., Silverman, K. 1986. An integrated approach to studying intonation and attitude. In Johns-Lewis, C., (ed.), *Intonation in Discourse*. San Diego, College-Hill Press, 125-138.
- Lang, P. J. 1993. The three system approach to emotion. In Birbaumer, N., Ohman, A. (eds). *The structure of emotion Psychophysiological, cognitive and clinical aspects*. Hogrefe & Huber Publishers, Seattle, Toronto, Bern, Gottingen, 1993, pp 18-30.
- Lang P.J., Levin D. N., Miller G. A., Kozak M. J. 1983. Fear behavior, fear imagery, and psychophysiology of emotion: The problem of affective response integration. *Journal of Abnormal Psychology* 92: 276-306.
- Lazarus, R. S. 1991. *Emotion and adaptation*. New York: Oxford University Press.

- Martin, J.-C., den Os, E., Kuhnlein, P., Boves, L., Paggio, P., Catizone, R. 2004 Workshop "Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces". In *Association with the 4th International Conference On Language Resources And Evaluation LREC2004* <http://www.lrec-conf.org/lrec2004/index.php>, Centro Cultural de Belem, LISBON, Portugal, 25th may, <http://lubitsch.lili.uni-bielefeld.de/MMCORPORA/>
- Maybury, M., Martin, J.-C. 2002. Workshop on "Multimodal Resources and Multimodal Systems Evaluation", Conference On Language Resources And Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, <http://www.limsi.fr/Individu/martin/research/articles/ws14.pdf>
- Massaro, D., Cohen, M. M. 2000. Fuzzy logical model of bimodal emotion perception: Comment on 'The perception of emotions by ear and eye' by de Gelder and Vroomen. *Cognition and Emotion* 14, 313-320.
- McEnery, T., Wilson, A. 1996. *Corpus Linguistics*. Edinburgh, Edinburgh University Press.
- McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M. & Stroeve, S. 2000. Automatic recognition of emotion from voice: a rough benchmark. *Proc. ISCA ITRW on Speech and Emotion*, 5-7 September 2000, Textflow, Belfast, 207-212.
- MPEG4 SNHC: Face and Body Definition and Animation Parameters., ISO/IEC JTC1/SC29/WG11 MPEG96/N1365, 1996.
- Milroy, L., 1987. *Observing and analysing natural language*. Oxford, Blackwell.
- Murray, I. R., Arnott, J. L. 1993. Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Amer.* 93, 1097- 1108.
- NITE Project <http://nite.nis.sdu.dk/>
- ORESTEIA IST-2000-69091 New Deliverable 2.5. Collection of data for the multiple user case: experiment design, data description, sampling methods .
- Roach, P., Stibbard, R., Osborne, J., Arnfield, S., Setter, J. 1998. Transcription of prosodic and paralinguistic features of emotional speech. *Journal of IPA* 28, 83-94.
- Savvidou, S., Cowie, R., Douglas-Cowie, E. 2002. Contributions of visual and auditory channels to detection of emotion. *Proc. BPS*, vol. 10 (2), August 2002, 48.
- Scherer, K. R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99 (2), 143-165.
- Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227-256.
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S. 2001. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. *Proc. Eurospeech 2001*, Aalborg, vol. 1, 87-90.
- Schröder, M. 2003. Experimental study of affect bursts. *Speech Communication*, 40 (1-2), Special Issue following the ISCA Workshop on Speech and Emotion, 99-116.

- Stemmler, G., 1992. The vagueness of specificity: Models of peripheral physiological emotion specificity in emotion theories and their experimental discriminability. *Journal of Psychophysiology*, 6, 17-28.
- ten Bosch, L., Emotions: What is possible in the ASR framework? *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 189-194.
- Trainor, L., Austin, C., Desjardins., 2000. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science* 11 (3), 188-195.
- Trudgill, P., 1983 (revised edition). *Sociolinguistics. An Introduction to Language and Society*. Penguin, London.
- Waheed, K., Weaver, K., Salam, F. M. 2002. A robust algorithm for detecting speech segments using an entropic contrast. *45th IEEE International Midwest Symposium on Circuits and Systems*, Tulsa, Oklahoma August 4-7, 2002, vol 3, 328-331.
- Waterman, M., Greasley, P. 1996. Development of a qualitative instrument for coding cognitive antecedents of emotional responses. *Int J Psychol* 31 (3-4), 4761.
- Weizenbaum, J.1966. ELIZA- A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery* 9, 36-45.
- Wichmann, A., 2000. The attitudinal effects of prosody, and how they relate to emotion. *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 143-147.
- Zhou, G., Hansen, J. H. L. & Kaiser, J. F.1999. Methods for stress classification: Nonlinear TEO and linear speech based features. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. IV, 2087-2090.