Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches

L. Devillers, R. Cowie, J-C. Martin, E. Douglas-Cowie, S. Abrilian, M. McRorie

LIMSI-CNRS; France, Queen's University Belfast; UK
{devil, martin,abrilian}@limsi.fr, {r.cowie, e.douglas-Cowie, m.mcrorie}@qub.ac.uk

Abstract

A major barrier to the development of accurate and realistic models of human emotions is the absence of multi-cultural / multilingual databases of real-life behaviours and of a federative and reliable annotation protocol. QUB and LIMSI teams are working towards the definition of an integrated coding scheme combining their complementary approaches. This multilevel integrated scheme combines the dimensions that appear to be useful for the study of real-life emotions: verbal labels, abstract dimensions and contextual (appraisal-based) annotations. This paper describes this integrated coding scheme, a protocol that was set-up for annotating French and English video clips of emotional interviews and the results (e.g. inter-coder agreement measures and subjective evaluation of the scheme).

1. Introduction

Interest in emotion-oriented computing has grown over the past decade, and it has brought with it research on databases that are naturalistic in the sense that the material shows emotion as it occurs in everyday action and interaction rather than idealised archetypes (Douglas-Cowie et al 2003, Cowie, Douglas-Cowie & Cox 2005). Working with these databases quickly highlighted a descriptive challenge, because it is clear that the emotion they contain is not adequately described by assigning a few theoretically derived labels, such as Ekman's 'basic emotions' (Cowie & Cornelius 2003, Devillers, Vidrascu & Lamel 2005). The literature offers many ideas relevant to labelling naturalistic video material. This paper describes the first substantial effort to draw key types of descriptor together in an annotation scheme capable of capturing the emotional content of naturalistic databases. The scheme is first explained, and then its application to a challenging set of naturalistic clips is described.

2. Background

The groups at QUB and LIMSI have both made substantial collections of naturalistic emotional material from TV – the Belfast naturalistic database and the Castaway database (in English), and the EmoTV database (in French) – and have worked on the problem of annotating its emotional content.

The aim of the study was to consolidate their expertise by defining a multilevel integrated scheme and evaluating it empirically. The most promising descriptors were combined into a joint coding scheme developed jointly by the teams. It includes multiple types of descriptor – verbal labels, abstract dimensions and contextual annotations, at base and meta levels.

The groups have also developed different tools for assigning descriptors. QUB has developed continuous methods of annotating emotions, notably the Feeltrace tool (Cowie et al 2000), while the EmoTV corpus is annotated using Anvil tool to identify discrete segments and assign categorical labels (Abrilian et al, 2005). Both types of technique were used in this study. Anvil was used to annotate each clip globally (ie to decide whether a label applied to the clip as a whole), and continuous trace techniques were used to capture change within the duration of a clip. Work is in progress to compare discrete and trace approaches to describing within-clip variation.

The scheme was used to annotate videos in both languages. Clips for the study to were chosen to represent the range of material in these databases. The results provide a way to identify a core of mature methods of annotating naturalistic emotions, which are both reliable and conceptually satisfying, and others that are promising in some respects, but need development.

3. The descriptors

This section lists the descriptors that the teams judged, on the basis of practical experience and theory, ought to be evaluated empirically. The descriptors listed in sections 3.1 and 3.2 were applied globally, i.e. assigned to the clip as a whole, using Anvil to view the clip and record judgments. Section 3.3 explains how trace techniques were used to describe variation within each clip.

3.1. Everyday labels that apply to naturalistic samples

Emotions in a strict sense	Emotion-related states	
Anger (hot)	Affection	Interest
Anger (cold)	Amusement	Irritation
Contempt	Anxiety	Pleased
Despair	Boredom	Relief
Disgust	Courage	Relaxation
Elation/joy	Disappointment	Satisfaction
Fear	Doubt	Serenity
Guilt	Embarrassment	Shock
Happiness	Empathy	Stress
Pride	Excitement	Worry
Sadness	Friendliness	
Shame	Helplessness	
Surprise	Hope	

Table 1: everyday labels selected for the study

We use the term 'everyday labels' to describe words of the kind listed in Table 1 above. They are clearly important, but using them effectively requires a good

compromise between lists that distinguish too few states to be useful (e.g. the 'big six') and lists that distinguish too many to be manageable. The items also need to reflect everyday emotional states. Table 1 aggregates the labels that have proved valuable in work with naturalistic material at LIMSI and Belfast.

3.2. Theoretically derived descriptions

Theory suggests many ways of describing emotion other than everyday labels. The scheme used here incorporates what seem to be the most important options.

3.2.1. Temporal combination labels

Naturalistic data often show emotions blending into each other, either simultaneously or sequentially (Douglas-Cowie et al. 2005). To reflect that, we developed the following categories that describe emotion combination:

- *Single* (non-mixed): only one emotion is perceived.
- *Simultaneously mixed*: several emotions are merged, and occur at the same time,
- *Sequentially mixed*: several emotions, one occurring shortly after the other, in a single emotional clip,
- *Simultaneous and sequentially mixed* combinations present (necessary because labelling was global).

Each of the emotions that made up the combination (up to 5) was then annotated to describe it more precisely, with focus, control and time qualifiers and abstract dimensions.

3.2.2. Focus control and time qualifiers

According to theory, archetypal emotions have a distinctive structure. They arise in reaction to an external event; the reaction 'synchronises' several systems and briefly dominates mental life; and it has an object (what it is 'about'). Since emotion in naturalistic data often shows only some of those features, we devised labels to capture departures from the archetypal pattern that make logical sense, and seem to be common.

- *Episodic emotion* conforms to the archetype
- *Simmering emotion*: elements of the archetype are present, but not (yet) synchronised and dominant.
- *Flitting emotion*: instead of focusing on a single topic, emotion lights on one issue after another.
- *Mood*: involves feeling that is global and diffuse rather than being focused on a specific topic.

Cutting across these is a distinction related to time.

- *Reactive emotion* is focused on and triggered by immediate events (as in the archetypal pattern).
- *Established emotion* consists of long standing feelings about past events or ongoing situations (which become apparent when they surface for some reason).

3.2.3. Dimensional

Emotion is also often described in terms of dimensions relating to the person's state. Six were considered.

- *Intensity:* the perceived strength of the emotion.
- *Valence:* a global measure of the positive or negative feeling associated with the emotion.
- *Activation:* the degree to which the emotion inclines the person experiencing it to be active or inactive.
- *Approach/avoid:* the degree to which the emotion inclines the person experiencing it to engage with the events concerned or to withdraw from them
- *Acted-level:* the degree to which the person is felt to be simulating emotion that is not genuinely felt
- *Masked-level:* the degree to which the person is felt to be masking an emotion that is felt.

3.2.4. Appraisal

A strong tradition distinguishes emotions in terms of the 'appraisals' that they involve. Appraisals are perceptual evaluations of emotion-relevant aspects of the situation. on a set of dimensions due to Scherer and his colleagues (Sander et al 2005). The following set of labels was used to describe the protagonist's appraisal of the event or events at the focus of his/her emotional state (call these E). They are grouped under four broader headings.

- *Relevance* Suddenness of E; Familiarity of E; Predictability of E; Intrinsic pleasantness of E; Desirability of the consequences
- *Implications* Agency responsible (self / other / group / nature); Underlying motive (negligence / intent); Nature of likely consequences (negative to positive); Relation to expectation (consonant to dissonant); Conduciveness to goals (conducive to obstructive); Urgency (low to high)
- *Coping potential*; Controllability of primary event (low to high); Controllability of consequences (low to high); Power of person to alter events (low to high); Scope for person to adjust own goals (low to high);
- *Compatibility of the situation with standards* With external standards (norms or demands of a reference group); With internal standards (self ideal or internalized moral code)

These are brief descriptions – fuller explanations are given by Sander et al. (2005).

3.3. Trace descriptions

The Feeltrace tool developed at QUB rates emotion on the dimensions of activation and valence that were described above, but the strategy behind it is very general – let raters 'trace' the way aspects of emotionality appear to rise and fall by moving a pointer in real time between markers that correspond to extreme states. This is a variant on the magnitude estimation strategy that psychophysics has shown is surprisingly powerful. In this study, raters made six one-dimensional traces. All of the dimensions have already been introduced. They are:

- *Emotional status* is concerned with the archetypal pattern referred to in 3.2.2. Raters used a scale divided into three parts – episodic at one extreme, wholly emotionless at the other, and partial emotion in the centre.
- *Concealment* and *acting* look at temporal variation of the issues of simulation and concealment (see 3.2.3).
- *Activation* and *valence* (see 3.2.3)
- *Power* (listed under 3.2.4, but power is classically linked with valence and activation in dimensional analyses).
- *Everyday labels* are considered in a two-stage process. Raters decide after watching a clip which words from Table 1 are relevant. Then for each, they

use a trace technique to indicate how the strength of the state it describes varies from moment to moment. The result is a 'soft coding' showing how elements change over time.

4. The study

4.1. Method

12 clips (6 from each team) were chosen for annotation. 4 raters (3 French and 1 Irish) annotated each clip with the ANVIL tool and 5 annotators (3 French and 2 Irish) with the Trace tools. Each used the trace tools first, in a set order, then annotated with ANVIL.

The exercise was evaluated at two levels. First, measures of reliability were calculated. For the Anvil categorical labels, the basic measures were Kappa coefficients augmented to handle multiple labelled data points For the trace measures, the basic measures were correlations between arrays summarising traces. Summary arrays were formed by dividing raw traces into three-second windows and taking the average value within each window. Second, a user questionnaire was constructed. Each rater assessed in the light of the annotation experience how well formulated each item in the scheme was; how easy it was to annotate, and how informative they felt it could be if it were well formulated. The general aim was to group descriptors into the following categories:

a) Elements whose usefulness is confirmed
b) Intermediate (work needed)
c) Elements which seem unlikely to be useful

4.2. Emotion-related Words

For 75% of the clips (9/12), there was at least one everyday label that all the annotators applied to them. That indicates very substantial reliability.

Negative labels	Positive labels	Other
Anger (Hot) Anger (Cold) Irritation Contempt	**Elation/joy** Happiness Amusement Pleased Satisfaction	**Surprise**
Disgust	**Excitement**	
Fear Anxiety Doubt Shock Stress Worry	**Affection** Friendliness	**Empathy** Interest
Embarrassment Shame Guilt	**Relaxation** Serenity	
Sadness Despair Disappointment Helplessness	**Relief**	
Boredom	**Pride**	
	Courage	

Table 2: Macro-classes of everyday labels

Finding more formal measures is not straightforward when multiple labels can be assigned. Some proposed extensions of the familiar kappa statistic give low values when raters use multiple labels even if they agree perfectly (e.g. Rosenberg & Binkowski, 2005). We adapted Rosenberg & Binkowski's approach to give a kappa measure whose value is (as expected) 1 when raters choose the same multiple labels. It was calculated both for the 35 original labels and for the 15 macro-classes shown in Table 2. Table 3 shows the results.

Coders	35 labels	15 labels
1 & 2	0.37	0.61
1 & 3	0.42	0.66
1 & 4	0.55	0.58
2 & 3	0.47	0.64
2 & 4	0.54	0.65
3 & 4	0.43	0.62
Average Kappa	**0.46**	**0.63**

Table 3: Kappa coefficients for everyday labels, using both the 35 original labels and 15 macro-classes.

Everyday labels emerge as reliable with respect to clips of the kind we used. The main concern was that 9 of the 35 labels were never used: cold anger, contempt, interest, relaxation, friendliness, empathy, boredom, affection and guilt. That suggests the clip set should probably be broadened for future work.

4.3. Combination labels

Although clips were chosen to represent cohesive episodes, it is clear from the everyday labels that all but a few involved more than one emotion. Cases where raters assigned only a single label were rare (3 of the 48 annotations); the average was two or three labels (2.66) per clip. Clips were about equally likely to receive a mixture of negative labels (n=5), a mixture of positive labels (n=3), and both positive and negative labels (n=4).

The user questionnaire indicated that the labels designed to describe combination patterns explicitly were well formulated, easy to annotate, and often highly informative. However, they were not consistently assigned; Agreement was only 16.6 %. However, the problem is probably a simple one of phrasing to deal with the fact that a clip may contain periods where different emotions overlap and transitions from one to another.

4.4. Focus, control and time qualifiers

When clips were assigned the same everyday label, they were assigned same reactive/established label in 71% of cases, and the same focus/control label in 69%. Interesting patterns emerge. States were labelled as reactive more often than established (by a factor of 2.5). However, the established category was relatively common with a few labels, notably sadness, despair, serenity, happiness, and courage. User questionnaires rated the reactive/established distinction useful and informative.

Focus control qualifiers (episodic / simmering / flitting / mood) were also evaluated as being well formulated, but "simmering" and "flitting" were not always easy to annotate, and only occasionally highly informative. This may reflect the nature of the sample. As Figure 1 shows, most of the material was episodic, with much less material in the other categories. Since episodic emotion

forms a relatively small part of emotional life in general (Cowie, Douglas-Cowie & Cox 2005), the data expose easily overlooked questions about representativeness.

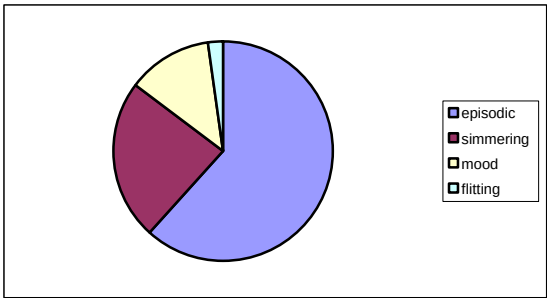


Figure 1: Division of 'focus and control' type labels.

4.5. Dimension labels

For dimension labels assigned to the clip as a whole, the computed measure of reliability was the percentage of pairs where ratings on these scales were the same or one step apart. The results are summarized in Table 4.

Acted	Masked	Valence	Intens	Activ	Approach
0.84	0.97	0.96	0.80	0.73	0.55

Table 4: % agreement on the global dimension labels

Results on classical dimensions show that the Valence and Intensity dimensions are more reliable than Activation. The Approach/Avoid dimension is an alternative to formulations involving power, which is taken up later. The Masked and Acted dimensions are innovations. Coding reliability is high and the user questionnaire indicates that the dimensions seem useful for describing complex real-life emotions. However, the material may give an artificially positive impression, because there was a very sharp contrast between one case where emotion was heavily disguised and most others. More work is needed to confirm this first experiment.

4.6. Appraisal labels

The classical kappa measure also raises problems with the appraisal variables. Table 5 gives coefficients across all the four individuals who carried out the global ratings.

1-Suddenness	0.1
2-Familiarity	0.19
3-Predictability	0.0
4-Intrinsic Pleasantness	**0.53**
5-Global Pleasantness	**0.4**
6-Cause Motive	0.13
7-Outcome Probability	0.01
8-Relation-Expectation	**0.27**
9-ConductivenessGoals	**0.7**
10-Urgency	0.22
11-Controllability-directEvt	0.19
12-Controllability-conseqEvt	**0.3**
13-power-of-person	0.15
14-adjustementGoals	0.05
15-Compatibility-External	0.05
16-Compatibility-Internal	0.03

Table 5: Global Kappa for appraisal items.

Figure 2 expands by showing how often at least 3 of the 4 annotators agreed. There was reasonably high agreement on the variables *Intrinsic pleasantness, Relation Expectation, Conductiveness to Goals, Controllability of direct Event* and *Controllability of consequences*. However, agreement elsewhere was low, and the kappa figures show that some agreements are empty (they arise because raters simply used one value almost all the time).

One of the main difficulties with the appraisal categories is that they are inherently linked to a focal event or situation. In fact, the naturalistic clips often imply that the emotional behaviour relates to multiple events (e.g. a trauma years ago, recent illness, the current interview). Parts of the appraisal framework cannot be applied without establishing which of these events is to be considered. That was recognised in the protocol, but the techniques designed to cope with the problem turned out to be problematic in themselves. The issue is taken up in section 4.8.

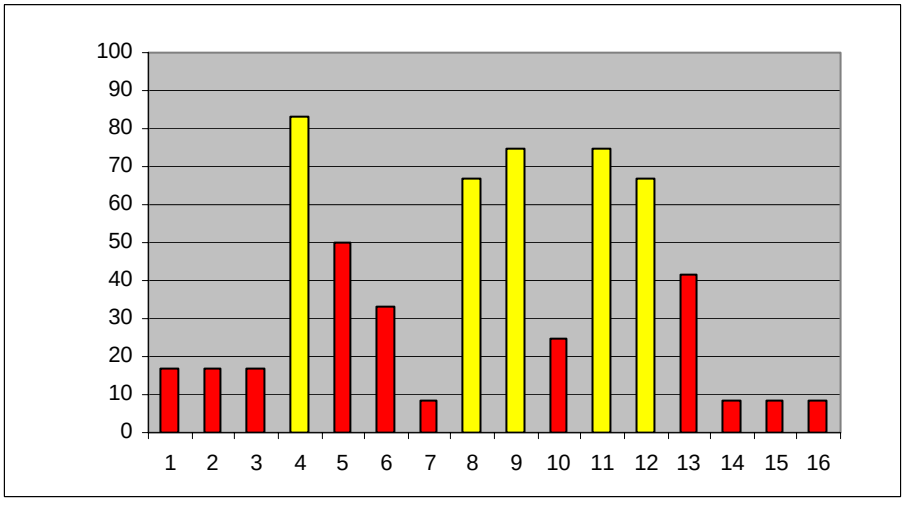
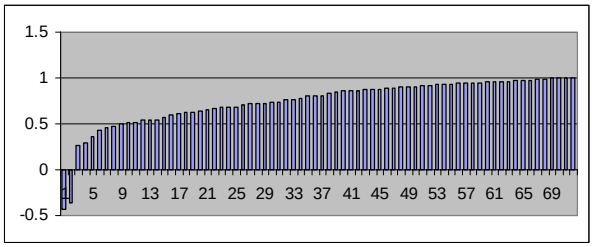


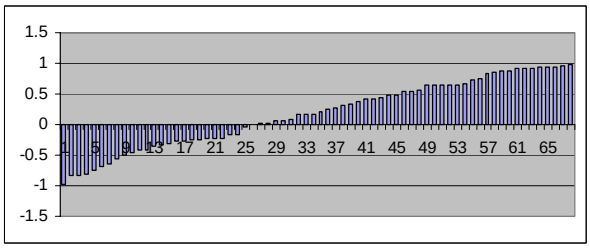
Figure 2: Agreement percentage obtained on the 12 clips for the 16 appraisal variables. The criterion of agreement is that at least 3 of the 4 annotators agree.

4.7. Trace measures

All of the trace measures were evaluated in the user questionnaire as being well formulated, easy to annotate and informative. The basic measure used to assess their reliability was correlation. Each trace was broken into 3 second blocks and an average rating calculated for each block: in that way each trace was reduced to an array of averages with a manageable number of points. Correlations were then calculated for every pair of arrays that involved both (a) the same passage, and (b) the same measure. These correlations indicate how strongly people agreed (or disagreed) on the way a particular attribute changed over the duration of a given passage.



(a) Intensity (generally strong agreement among raters)



(a) Power (generally weak agreement among raters)

Figure 3: Examples of how correlations between traces are distributed. Each bar represents the correlation coefficient between two traces of a single clip produced by different raters.

Figure 3 illustrates the way correlations associated with a single trace tend to be distributed, showing one of the dimensions where agreement is strongest, and one where it is weak. The pattern of the distributions invites a simple way of summarising them, which involves specifying the proportion of the correlations that are below -0.5; the proportion that are between 0 and -0.5; the proportion that are between 0 and 0.5; and the proportion that are above 0.5.

Table 6 summarises the correlations between traces in that format. The first part of the table considers the correlations associated with each of the scales that was used. It can be seen that for the dimensional scales, strong agreement was overwhelmingly the norm for the intensity scales (continuous and emotional status) and the valence scale. Agreement was intermediate for activation, and low for power, masking, and acted.
The findings for activation and power reinforce earlier findings, and suggest that moderate or low inter-rater agreement in these areas is to do with the underlying judgment rather than the particular technique that is used to assess it (note that the approach/avoid global dimension label, which also gave problems, expresses an essentially similar underlying concept). The findings for masked and acted are somewhat different: it may be easier to agree on global ratings than on when precisely emotion is being masked or simulated. But as noted earlier, the selection of clips may not have allowed this issue to be adequately explored.

For the words, there were far fewer pairs of arrays to consider. On the whole, when people used the same word, they agreed on the way its relevance changed over time. The figures conceal the fact that there were fewer very strong agreements in the word ratings than in the dimensions where agreement was the rule.

Scale	very -ve	weakly -ve	weakly +ve	very +ve
intensity	0.00	0.03	0.08	0.89
emotional status	0.01	0.04	0.16	0.79
valence	0.13	0.00	0.17	0.70
activation	0.11	0.15	0.19	0.55
power	0.12	0.25	0.29	0.34
mask	0.11	0.24	0.32	0.33
acted	0.15	0.26	0.35	0.24
anger	0.06	0.31	0.13	0.50
sadness	0.19	0.06	0.25	0.50
anxiety	0.00	0.25	0.25	0.50
shock	0.00	0.50	0.00	0.50
helplessness	0.13	0.13	0.38	0.38
serenity	0.38	0.25	0.13	0.25
stress				1
surprise				1
irritation				1
despair				1
embarrassment				1
amusement				1
elation			.33	.67
happiness			.5	.5
relief		.5		.5
excitement		.75		.25
pleasure		1		
disappointment		1		

Table 6: Correlations between traces involving the same dimension or label, showing proportions that are below -0.5 (very –ve); between 0 and -0.5 (weakly –ve); between 0 and 0.5 (weakly +ve); and above 0.5 (very +ve). Words are divided into an upper group where the number of paired traces is reasonably large; and a lower group, where it is small.

Trace techniques emerge from the study as a powerful way of measuring variation within a passage. They also raise a variety of interesting issues that cannot be pursued here, e.g. how individual raters differ (though differences are not extreme); and what is happening when they occasionally produce traces which are negatively correlated, as they quite systematically do.

4.8. Connections

An emotional episode is richly connected. It involves transient feelings about a particular topic (and in our material often several topics), expressed through various signs to an addressee, taking account of an audience. The transient feelings may be prompted by events or thoughts in recent past, but may arise from enduring feelings, about topics or issues of enduring significance to the person, which may be moulded by longer term factors (e.g. stress at work, bereavement).

We designed labels to capture key parts of this complex as simply as possible. The labels dealing with one part, the signs used to express the emotion, were well rated in the user questionnaire, but will not be discussed at length here. The other components were rated as potentially informative, but difficult to apply in practice in the existing format; or unsatisfyingly restricted to the particular context of TV interviews. Capturing the key connections that give an emotional episode its meaning is one of the largest challenges to be addressed in the immediate future.

5. Discussion & conclusion

The study has confirmed that a range of emotion-related descriptors can be applied with some confidence. They have already been identified, and will not be repeated here. Between them they support a very substantial description of perceived emotional content.

One of the main innovations was using appraisal-based labelling. It achieved some success. Several of the items appear to be reasonably successful as they stand, particularly items in the implication/consequences grouping (the outcome probability, the relation to expectations, the conduciveness to goals and the urgency were evaluated as well formulated, easy to annotate and informative). The study also clarified the difficulties involved in extending appraisal labelling. For many items, the existence of multiple relevant events is key (this affects agent responsible, motive, outcome probability, and all the items of the coping potential dimension except power - controllability of immediate event, controllability of consequence event, possible adjustment to person's own goals). The same problem affects the 'compatibility with standards' items, but in addition the basic issue was clearly not well explained in this protocol.

Recognising these problems with appraisal labelling is a major outcome. The lessons will be used to refine the approach. Trace methods have already been adapted to show where particular factors become relevant.

Another major issue for research is to integrate approaches to capturing variation within a clip. Trace techniques were used here, but there are alternatives, based on identifying segments to which raters judge a label can or cannot be applied. The relationships between the alternatives need to be studied.

Significant questions also arise about selection of test material. Clips need to cover an appropriate range if they are to be a good test of labelling techniques. The study helps to clarify what that entails. Most obviously, there should be material that exemplifies all the everyday labels being used. More subtly, criteria for selecting everyday labels need to be articulated; and episodic emotion should not be disproportionately predominant. These points highlight the need for better information about the prevalence of relevant states. In sum, creating a satisfyingly balanced body of test material emerges as a challenge that is tightly coupled to work on descriptors and prevalence.

More generally, there has been a tendency to regard different coding schemes (e.g. discrete and continuous, dimensional and appraisal-based) as competitors associated with competing theories. The approach here is to evaluate options on the basis of sustained engagement with real and diverse data, and to let that shape annotation rather than relying on a priori assumptions. It is recognised that two kinds of evidence are relevant – statistical consistency and raters' judgments. They are worth combining precisely because they do not always agree. That kind of evidence-based development is less exciting than theoretical battles, but engaging with it in a sustained way is absolutely fundamental.

Acknowledgement

This work was partly funded by the FP6 IST HUMAINE Network of Excellence (http://emotion-research.net).

6. References

Abrilian, S., Devillers, L., Buisine, S., Martin, J-C, (2005) "EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces", *HCI International*.

Cowie & Cornelius (2003). Describing the emotional states expressed in speech. *Speech Communication*, 40(1-2), pp. 5-32.

Cowie, Douglas-Cowie & Cox (2005). "Beyond emotion archetypes: databases for emotion modelling using neural networks" *Neural Networks*, 18, pp. 371-388.

Cowie, R. Douglas-Cowie, E, Savvidou, S., McMahon, E., Sawey, M. & Schroeder M. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. In R. Cowie, E. Douglas-Cowie & M. Schroeder (eds.), *Speech and Emotion*: *Proc. of the International Speech Communication Association Research Workshop*. Newcastle, Co. Down, September 2000, pp 19-24.

Devillers, Vidrascu & Lamel, (2005). "Challenges in real-life emotion annotation and machine learning based detection", *Neural Networks* 18, pp. 407-422.

Douglas-Cowie, E. Campbell, N. Cowie, R. & Roach, P. (2003) "Emotional speech; Towards a new generation of databases", *Speech Communication*, 40, pp. 33-60.

Douglas-Cowie, E., Devillers, L., Martin, J-C., Cowie, R., Savvidou, S. Abrilian, S., Cox, C. (2005) "Multimodal databases of everyday emotion: content and labelling", *Interspeech* 2005.

Rosenberg & Binkowski (2005). "Augmenting the kappa statistic to determine interannotator reliability for multiple labelled data points", *HLT-NAACL*.

Sander, D., Grandjean, D., & Scherer, K. (2005) A sytsems approach to appraisal mechanisms in emotion *Neural Networks* 18, pp 317-352.