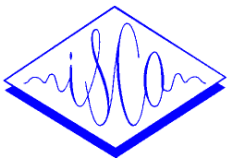

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

'FEELTRACE': AN INSTRUMENT FOR RECORDING PERCEIVED EMOTION IN REAL TIME

Roddy Cowie, Ellen Douglas-Cowie[*]*, Susie Savvidou, Edelle McMahon, Martin Sawey & Marc Schröder*

Schools of Psychology and English[*], Queens University Belfast

ABSTRACT

FEELTRACE is an instrument developed to let observers track the emotional content of a stimulus as they perceive it over time, allowing the emotional dynamics of speech episodes to be examined. It is based on activation-evaluation space, a representation derived from psychology. The activation dimension measures how dynamic the emotional state is; the evaluation dimension is a global measure of the positive or negative feeling associated with the state. Research suggests that the space is naturally circular, i.e. states which are at the limit of emotional intensity define a circle, with alert neutrality at the centre.

To turn those ideas into a recording tool, the space was represented by a circle on a computer screen, and observers described perceived emotional state by moving a pointer (in the form of a disc) to the appropriate point in the circle, using a mouse. Prototypes were tested, and in the light of results, refinements were made to ensure that outputs were as consistent and meaningful as possible. They include colour coding the pointer in a way that users readily associate with the relevant emotional state; presenting key emotion words as 'landmarks' at the strategic points in the space; and developing an induction procedure to introduce observers to the system.

An experiment assessed the reliability of the developed system. Stimuli were 16 clips from TV programs, two showing relatively strong emotions in each quadrant of activation-evaluation space, each paired with one of the same person in a relatively neural state. 24 raters took part.

Differences between clips chosen to contrast were statistically robust. Results were plotted in activation-evaluation space as ellipses, each with its centre at the mean co-ordinates for the clip, and its width proportional to standard deviation across raters. The size of the ellipses meant that about 25 could be fitted into the space, i.e. FEELTRACE has resolving power comparable to an emotion vocabulary of 20 non-overlapping words, with the advantage of allowing intermediate ratings, and above all, the ability to track impressions continuously.

1. INTRODUCTION

One of the major challenges facing research on speech and emotion is to describe and analyse the emotional content of everyday speech in the form of spontaneous monologues and interchanges. That means coming to terms with two key issues. One is gradation. Spontaneous speech is generally characterised by emotional shading rather than episodes of fullblown archetypal emotion. The other is variation over time. Intuitively it seems that the emotional shading of everyday speech is rarely constant. Instead it shifts, often gradually, but sometimes sharply – either because of a change in the speaker's state, or because something about the speaker's state suddenly becomes apparent.

Our team is part of a project known as PHYSTA, whose aim is to achieve automatic recognition of emotion. Part of the project is to assemble a database of spontaneously occurring speech with marked emotional content. To describe the emotional content of that material, we have developed a system called FEELTRACE. It is designed to let observers track the emotional content of a speech sample as they perceive it, taking full account of gradation and variation over time. FEELTRACE does not pretend to capture the emotional state that led the speaker to give the emotional signs in question: its function is strictly to record the way the signs are perceived by reasonably representative observers.

Underlying FEELTRACE is a representation called activation-evaluation space, which has a long history in psychology [1], [2]. Our contribution is to incorporate the space into a system that observers find it easy to use, and that gives reasonably reliable outcomes.

Activation-evaluation space represents emotional states in terms of two dimensions. Activation measures how dynamic the emotional state is. For instance, exhilaration involves a very high level of activation, boredom involves a very low one. Evaluation is a global measure of the positive or negative feeling associated with the emotional state. For instance, happiness involves a very positive evaluation, despair involves a very negative one. Many techniques converge on the conclusion that to a first approximation, emotion terms can be understood as referring to points in a space defined by those two axes. In addition, several techniques suggest that the space is naturally circular [2]. The circumference is defined by states that are at the limit of emotional intensity. These are equidistant from an emotionally neutral point, i.e. they define a circle, with alert neutrality at the centre.

We have gradually developed that basic concept into a recording tool. The essential idea is to present activation-evaluation space as a circle on a computer screen, and to have observers record their impression of emotional state by moving a cursor to the appropriate position in the space using a mouse. The development task is to ensure that users do that in a way that is consistent and whose meaning is reasonably clear. Figure

1 shows a screen captured from an up to date version of the system.

2. PROTOTYPES

Initial versions of Feeltrace incorporated several features designed to convey the idea of emotion as a point in a 2-D space. They can be seen (with some adjustments) in Figure 1.

The space was represented by a circle on a computer screen. Within it, the main axes were drawn and labelled - one (activation) running from very active to very passive; the other (evaluation) running from very positive to very negative. The basic response mode was to move a cursor within the space, so that at any given time its position signalled the levels of activation and evaluation perceived by the user. The cursor was controlled by a mouse.

The task in refining the approach was to provide feedback that ensured users were fully aware at any given instant what it meant to place the cursor in a given position. Two main devices were used for that purpose.

First, the cursor was colour coded using a scheme derived from Plutchik, which subjects find reasonably intuitive. The pointer took the form of a disc. It was coloured pure red when its position signified the most negative evaluation possible and neutrality with respect to activation, and pure green when its position signified the most positive evaluation neutrality with respect to activation. It was pure yellow when its position signified the most active state possible and neutrality with respect to evaluation, and pure blue when its position signified the least active state possible and neutrality with respect to evaluation. Colour at intermediate positions was set by a straightforward additive rule. It was white when the cursor was at the origin of the space.

To supplement the colour coding, we added verbal landmarks of two types. At the periphery we placed a small number of words designed to identify the strong, archetypal emotions associated with broad sectors of the circle. Representations provided by Plutchik [2] and Russell [3] provided prototypes.

Within the circle, less extreme emotion words were placed at the coordinates where empirical evidence indicated they belonged in the space. Initially, the positions were set on the basis of tables published by Whissell [4], which give activation-evaluation co-ordinates for common emotion words. Their function was to ensure that users could easily relate position in the space to everyday categorical descriptions of emotionality.

The dimension of time was represented indirectly, by keeping the circles associated with recent mouse positions on screen, but having them shrink gradually over time (as if the pointer left a trail of diminishing circles behind it). The effect was to provide a visual indication of the way ratings were being changed over time.

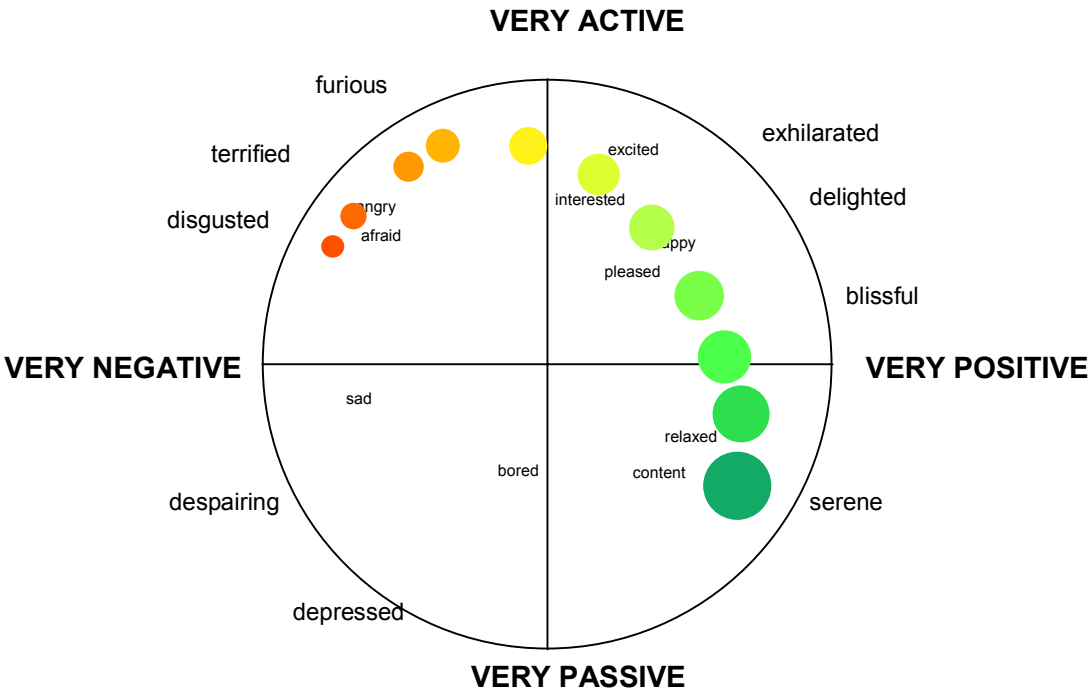

Figure 1: example of a FEELTRACE display during a tracking session. Cursor colour changes from red/orange at the left hand end of the arc, to yellow beside the active/passive axis, to bright green on the negative/positive axis, to blue-green at the right hand end of the arc

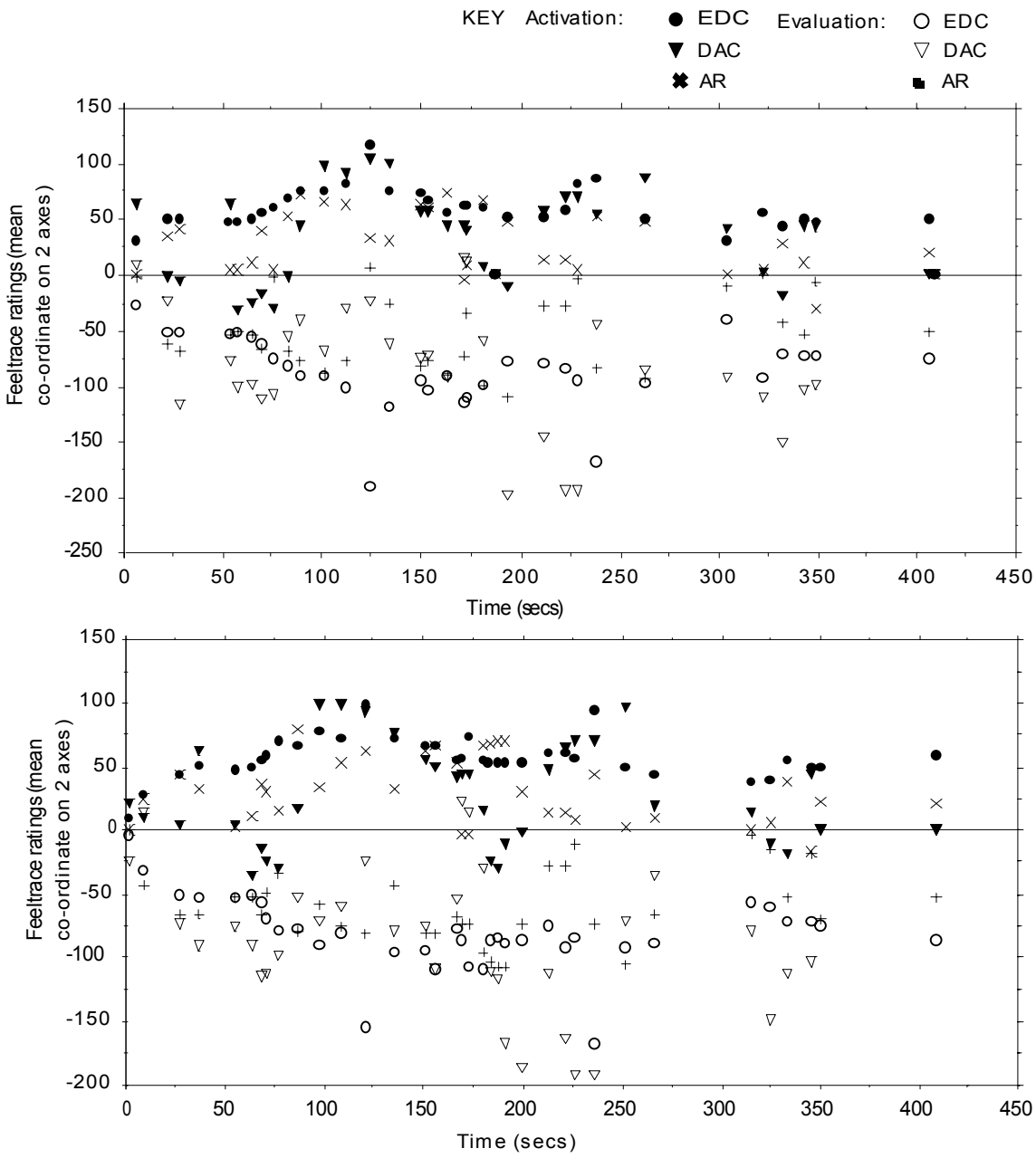

Figure 2: Feeltrace ratings by 3 raters for section of conversation involving 2 speakers, speaker R (top panel) and speaker J (bottom panel).

Two main studies were carried out with systems of that general type. In the first, the stimuli were tapes of discussions about emotive topics. In the second, they were samples of music. Both studies focussed on the ability of the system to detect changing shades of emotion over time.

Results of the discourse study have been reported elsewhere [5]. FEELTRACE ratings for each turn were summarised in terms of four measures, average level and spread with respect to both activation and evaluation. The pattern of results is illustrated in Figure 2. It was reasonably clear that the system captured broad rises and falls in the emotional temperature. Less predictably, there were associations between spread of FEELTRACE ratings (signifying change of emotional shade within a turn) and speech variables. However, differences between raters were large enough to cause concern.

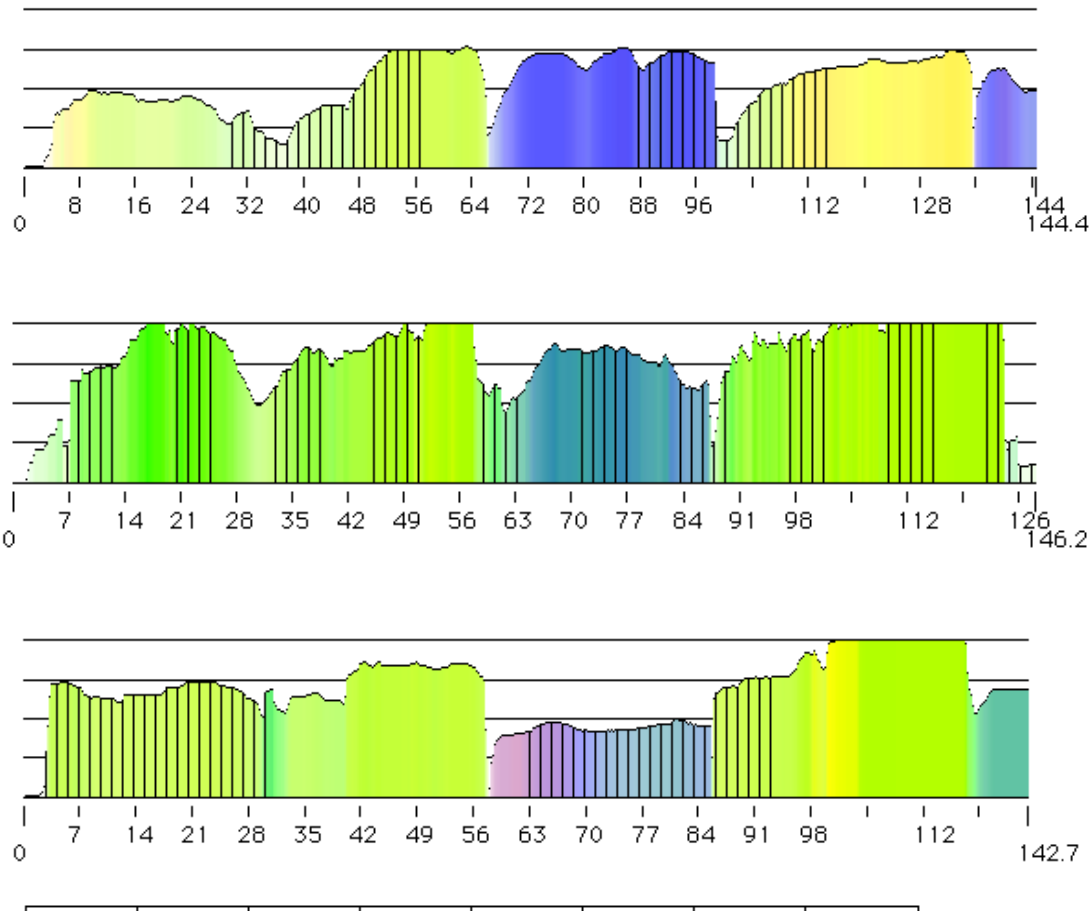

Figure 3: Feeltrace ratings by 3 raters for musical passage (the Great Gate of Kiev) that was expected to show three phases. The horizontal axis shows time (in seconds). The text explains what is meant by the height and colour of each vertical line.

The music study aimed examined the system's ability to capture variation on a finer temporal scale. Stimuli were musical extracts chosen for their ability to evoke emotions in different areas of the FEELTRACE circle. Some were intended to evoke a relatively consistent emotion, but others involved a change of mood, and were expected to show that ratings shifted from one part of the extract to the next. Ten subjects rated the passages.

Figure 3 shows examples of the ratings produced on one of the passages chosen to show internal change, the Great Gate of Kiev from Mussorgsky's 'Pictures at an Exhibition'. The format is designed to exploit Feeltrace colour coding. Each vertical line represents a cursor position. Its length is the distance from the centre of the activation-evaluation circle to the cursor position, and its colour is the colour that the cursor takes on at that position.

Intuitively it appears that the results do involve a shift from one emotional region into another, and back to the first (or something similar to it. Formalising that kind of intuition is a challenge, In the music study, a straightforward approach was adopted. Break points were identified a priori, and responses were averaged within the time intervals that they defined. For all of the relatively simple passages, the differences between blocks identified in that way were comfortably significant. In the case of the Great Gate of Kiev, the three blocks – corresponding to beginning, middle, and end – differed on MANOVA with $F(2,8)= 73.7$, $p<0.001$. The middle block, from about 60 to 100 seconds into the piece, was consistently below the midline in activation, whereas the beginning and end were above it.

3. SYSTEM DEVELOPMENT

The studies that we have described led us to adapt the system in several ways.

Landmark reselection

The first adaptation was to obtain landmark words based on subjects' responses to the particular representation of activation-evaluation space that we used. Data were collected as part of a larger procedure, which was called BEEVer because it was designed to obtain a 'Basic English Emotion Vocabulary' [6]. Subjects rated the emotional content of 40 words, in a procedure which included marking the position in an activation-

evaluation circle that they believed best represented the meaning of each word. They also selected from the initial list a subset of 16 that they judged would form an adequate basic emotion vocabulary.

Feeltrace landmarks were revised in the light of the BEEVer study. The words used within the circle were ten of the twelve items most often included in subjects' basic emotion vocabulary. They were placed at the average of the coordinates associated with them by the BEEVer subjects. The words around the periphery were revised to reinforce the layout set by the internal landmarks. The aim of these changes was to ensure that the verbal landmarks encouraged users to calibrating their responses in the way that that people on average find most natural.

Induction

The second modification was to develop an induction procedure, in which the properties of the system were introduced step by step, with illustrative examples. The examples were designed mainly to instruct, but also provide a mechanism for excluding individuals who used the system in a very eccentric way: the BEEVer study had indicated that a small minority of people do seem to do that.

Integrated presentation

The main medium for PHYSTA is video rather than pure audio. It is not technically trivial to present Feeltrace in a way that allows users to watch video input and use the visual feedback that Feeltrace is designed to provide. In response to that problem, we have developed an integrated package which presents Feeltrace and replays video on a single screen (from MPEG files). The program presents the activation-evaluation circle on the right of the screen, and brings up a window on the left of the screen in which video material can be played (using standard PC routines for video presentation). In the case where the task is to trace the emotional state of one participant in an interaction, the system presents a preliminary picture of the person in question so that the tracer can identify him or her.

Output and basic analysis

The immediate output of Feeltrace is a file in which each line begins with a time code, then gives the activation and evaluation co-ordinates of the mouse at that time. Originally time was counted from a manually signalled start. That led to appreciable differences in the timebases for different tracings of the same sample (the 'Great Gate of Kiev' examples illustrate the point). The current version takes time codes directly from the video source.

The format shown in figure 3 provides a satisfying way of displaying the contents of a tracing.

Programs have been developed to generate useful summaries from the immediate output files. For any given sequence, they generate the mean and standard deviation for cursor positions defined both in x and y co-ordinates (evaluation and activation respectively) and in polar co-ordinates (corresponding to emotional orientation and emotion strength). A facility used in the music study allows a file to be divided into constituent sequences defined by temporal boundaries. The standard measurement battery is then calculated for each sequence. The temporal boundaries can be specified by a file associated with the sample which raters Feeltraced. In PHYSTA, boundaries are being by automatic analysis of the speech sample which identifies 'tunes' (i.e. episodes bounded by substantial periods of silence).

4. SYSTEM VALIDATION

An experiment was carried out to assess the reliability of the system with these adaptations. Stimuli were clips extracted from TV programs involving real interactions (not acted). A clip showed a naturally bounded episode, and typically lasted 15 – 30 seconds. 16 were selected. According to the judgement of the experimenter, there were two in each quadrant of activation-evaluation space showing relatively strong emotions and two with the same people in relatively neural emotional states. 16 observers took part, each rating the clips in a different order in a counterbalanced design.

Statistical analyses of the results show that Feeltrace is a reliable measurement tool and give indications of its precision.

First, the difference in intensity (distance from the centre of the circle) between emotional and neutral passages was highly significant (repeated measures ANOVA, within subjects, $F(23,1) = 453$, $p < .001$). Hence the system is capable of capturing the kind of discrimination that the experimenter made in these cases.

Highly significant differences were also found when emotional passages were compared. Clips that differed in experimenter-judged activation were rated differently in activation ($F(23,1) = 239$, $p < .001$) and clips that differed in experimenter-judged evaluation were rated differently in evaluation ($F(23,1) = 847$, $p < .001$). This indicates that neutrality (the centre of the circle) and each of the four quadrants can be distinguished very reliably. In addition, paired t-tests showed significant differences between the two emotional passages in a quadrant for three out of four quadrants. This indicates that Feeltrace measurement is sufficiently fine-grained resolution to discriminate between episodes of moderately strong emotion within the same quadrant of activation/evaluation space.

Spatial presentation provides a clearer picture of Feeltrace resolution. Mean co-ordinates were calculated for each observer's response to each clip. The standard deviations of these measures were then calculated, one each for evaluation and activation for each clip. Graphically, the results can be pictured as elipses, each with its centre at the mean co-ordinates for the clip, and its radius in a particular direction equal to the standard deviation with respect to the appropriate axis of the space. Results are shown in Figure 4.

.

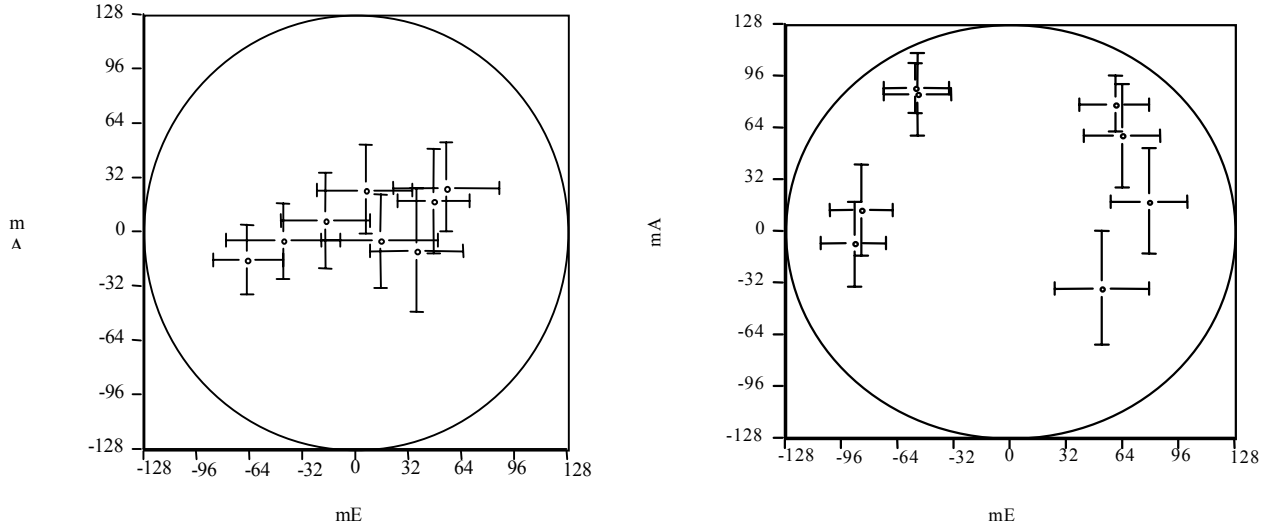

Figure 4: summaries of Feeltrace ratings in the validation experiment, showing clips judged neutral by the experimenter (left hand panel) and clips judged emotional (right hand panel). Ellipse height and width the standard deviations of average observer judgements in activation and evaluation respectively.

The diameter of an ellipse was generally about 1/5 of the diameter of the total space, implying that about 25 non-overlapping elipses could be accommodated in the whole space. As a rough guide, that suggests that Feeltrace has a resolution comparable to a vocabulary of around 20 non-overlapping emotion words.

5. DISCUSSION

Our results indicate that FEELTRACE in its current form has considerable resolving power. On the crudest summary, it is comparable to an emotion vocabulary of 20 words or more. Beyond that are distinctive advantages that mark it off from categorical approaches – the fact that intermediate states are handled smoothly; that there is no doubt about whether to treat every category as equally distinct from every other; the fact that numerical format makes statistical treatment straightforward; and above all, the ability to capture temporal variation, both in the relatively long term and from second to second.

It has to be stressed that FEELTRACE is not a perfect system. There are distinctions that it fails to capture, notably the distinction between fear and anger. That is because the space of possible emotions has more than two dimensions. As a result, projecting it onto two is bound to lose information. We have partially overcome the loss by supplementing pure FEELTRACE descriptions with verbal labels, though we have only attached them to whole clips, since there is no obvious way of giving them the kind of continuity that FEELTRACE provides. That illustrates the gain in tractability that the simplification offers.

We believe that for many emotional phenomena, particularly those where temporal issues are critical, FEELTRACE descriptions provide a useful starting point. Temporal issues certainly are critical in vocal expression of emotion, and capturing the vocal parameters associated with the changes FEELTRACE records is a challenge well worth taking up.

That kind of undertaking only makes sense if the instrument is standardised. Using ad hoc variants of the system is likely to produce noise within studies and systematic divergence between them. For that reason, our intention is to make available the version that we have developed and standardised.

6. REFERENCES

1. Schlosberg, H. (1954) A scale for judgement of facial expressions. *Journal of Experimental Psychology* 29, 497-510.

2. Plutchik, R. *The Psychology and Biology of `Emotion*. Harper Collinns, New York, 1994

3. Russell, J. A. (1997) How shall an emotion be called? In R. Plutchik & H. Conte (eds*) Circumplex Models of Personality and Emotions*. Washington: APA

4. Whissell, C. The dictionary of Affect in Language In R Plutchik & H Kellerman ed. *Emotion: Theory, research and experience vol 4* Academic Press: New York 1989

5. Cowie, R, Douglas-Cowie, and Romano, A Changing emotional tone in dialogue and its prosodic correlates Proc ESCA workshop on Dialogue and Prosody, De Koningshof, Sept 1999 pp 41- 46

6. Cowie, R, Douglas-Cowie, E, Apolloni, B, Taylor, J, Romano, A and Fellenz, W What a neural net needs to know about emotion words In N. Mastorakis *Computational Intelligence and Applications* World Scientific & Engineering Society Press, 109-114, 1999