GTrace - General trace program from Queen's, Belfast

Roddy Cowie & Martin Sawey

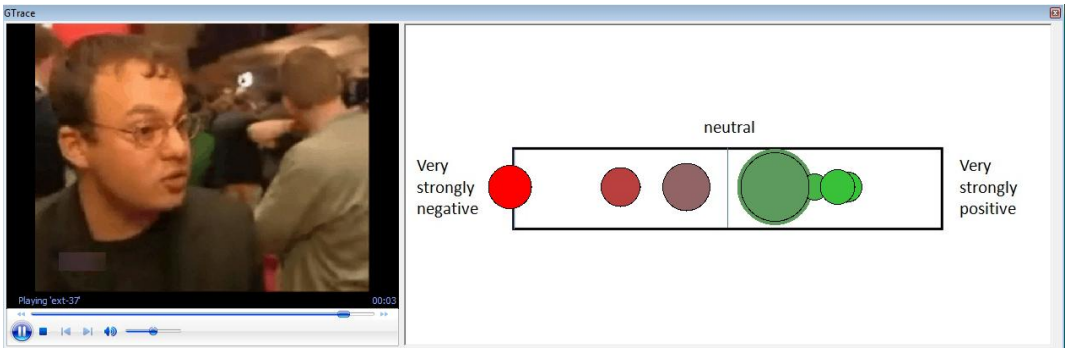
Gtrace is a program which allows a user to create a 'trace' that specifies how s/he sees emotion rising and falling over time. It can be downloaded from http://go.qub.ac.uk/GTrace . The most recent version of the program is compatible with the new standard for describing emotion on the web, embodied in EmotionML (http://www.w3.org/TR/emotionml/). These notes are not a complete manual – users should play with the system and get a feel for what it can do. The aim is to give users enough background to do that. We would appreciate feedback on obvious bugs in the system or the notes. If you use GTrace in an academic paper, please cite Cowie, McKeown and Douglas-Cowie (2012) – see the reference section for details. If you use it in business or the arts, we would appreciate an email to say so, to r.cowie@qub.ac.uk .

1. Overview

Gtrace is the successor to FEELtrace and related programs (Cowie, Douglas-Cowie et al 2000; Cowie, Cox et al 2011), which have been quite widely used. Its function is to let observers watch and/or listen to a designated person in a recording, and to indicate how particular features of the person's state appear to fluctuate with time. An observer records his/her impression of a given feature by moving a cursor back and forward along an appropriately designed scale as the recording plays. The result is a 'trace' – a stream of numbers showing how the state appears to change over time.

The original FEELtrace was written with a specific application in mind, that is, recording perceived emotion, in terms of standard dimensions – valence (positive to negative) and arousal/activation (dynamic to inert). However, it has become clear that the technique of tracing is much more general. Broadly speaking, it is comparable to the kind of n-point scale that questionnaires take for granted, but with a temporal dimension. Gtrace reflects that history. The package includes a set of scales that have been used to record emotion-related impressions, but it is designed to let people create their own scales with minimum effort, much as they could create their own n-point scales for a questionnaire.

A major new development has been the emergence of standards for the description of emotion on the web, embodied in EmotionML. Gtrace allows users to trace the emotion-related concepts incorporated in EmotionML, and provides outputs in an xml format that is compatible with EmotionML. Hence, for instance, if an avatar is controlled by EmotionML specifications, it should be possible to use Gtrace outputs to drive its expressions.



The basic form of Gtrace rating is illustrated above. On one side of the computer screen, raters see a recording that shows the person to be rated. On the other, they see a cursor that they can manipulate. It takes the form of a coloured disc, which they can move along a single dimension (left to right). The cursor moves within an area bounded by a rectangle, which has various markers

associated with it. There are usually light vertical lines within the rectangle, dividing it into equal parts – which may be halves, thirds, quarters, etc.. All the scales provided with Gtrace have text at each end of the rectangle, to indicate what the extremes of the scale mean. They usually also have a label associated with each dividing line. They may also have a caption that defines the attribute under consideration. The colour of the cursor usually changes as it moves along the scale, in a way that goes naturally with the meaning of the scale. For example, in the valence scale included with Gtrace, the cursor is pure red at the negative extreme, and pure green at the positive extreme. It leaves a 'tail' behind it, in the form of circles that show where it was recently, and shrink away over time. The point of all these devices is to help raters to understand the scale in the way that was intended.

Although tracing is quite general, there are some functions that it is naturally suited to, and other things that it should not be expected to do. At root, traces are records of the impressions that raters form, not records of the true state of the person (or people) seen and/or heard in the recording. For instance, the simplest description of what is described by a trace concerned with anger is 'apparent anger'. It is a separate question whether the apparent anger corresponds to actual anger, and answering that question requires techniques other than tracing. Similarly, tracing is suited to capturing impressions that people form and access easily and quickly, not considered judgments. An important implication is that tracing cannot necessarily be used to access any arbitrary attribute that an experimenter specifies. For instance, it is tempting to ask 'how intensely is the person feeling emotion in this particular theoretical sense?' However, there is no guarantee that a tracer can capture the movement of emotion in that particular theoretical sense, however carefully he/she is briefed. Tracing has to be guided by an impression that can be formed and accessed while the tracer is simultaneously carrying out two other tasks – watching a video and controlling a cursor. The trace will be governed by whatever sense of 'emotion' can be accessed more or less instantaneously under those circumstances. In one sense, that is a limitation. In another, it is an interesting property. After all, someone who is engaged in an interaction with another person is also likely to be dealing with demands that are comparably challenging, and needs to rely on impressions of the other person that are similarly accessible.

At a more specific level, a large amount of work is needed to separate scales that work well from scales which, for one reason or another, do not. Note that the same is true for questionnaires. Several quite different types of issue are involved. One is graphic design. Scales need to let raters 'get their bearings' at a glance. Gtrace scales have a variety of features which evolved for that reason – a prominent frame, dividing lines, informative cursor colour, and a 'tail' that shows where the cursor has been for the last few seconds. The design also needs to maximise the likelihood that individual raters will understand the scale in the same way. The verbal landmarks and cursor colour are designed to do that. Last but not least, the attributes to be traced have to be well chosen. Some attributes can be traced: others cannot. Exploratory research is the only way to establish which is which.

A key issue in evaluation is that tracers are different in non-trivial ways. Training is one factor. Tracing does not need the kind of in-depth training that, for instance, is needed for FACS coding. Nevertheless, tracers need a basic level of explanation and familiarisation: tracers who are confused about what they are doing simply generate noise. The explanations used in the SEMAINE project are given in appendices as examples. Even after training, there seem to be individual differences in tracing. In some tracing exercises – probably most – there is one dominant pattern, and people who depart from it can be classed as outliers. However, it is not uncommon to find more complex patterns. One which is well documented is that people from different cultures show the same trace shape, but it is shifted up or down depending on culture (Sneddon et al 2011). Less often, people

observing the same clip seem to gravitate towards two or more shapes of trace (Cowie, McKeown and Douglas-Cowie 2012). These issues are taken up in more detail later on.

The scales provided with Gtrace have been studied to at least some extent, and in particular something is known about the consistency with which they are used. The fullest information about them is in papers describing the SEMAINE project, which used tracing to annotate audio-visual recordings of emotionally coloured interactions (McKeown et al 2012; Schroeder et al 2012; Cowie, McKeown & Douglas-Cowie 2012). Many other scales have been explored to a lesser extent, and a brief overview is given later.

It is easier to understand the system by starting with a particular use than by describing it in an abstract way, and so the next three sections give basic technical information and then describe straightforward use of the scales that are provided with GTrace. After that, information is given about additional options, technical issues, etc..

2. Basic technical information

The program version distributed with these notes is called GTracev0.1. Later versions are likely to be different in various technical respects.

Requirements

Microsoft .Net Framework version 4
Screen resolution: 1280x1024 minimum

Installing the program

To run GTrace, you must first install Microsoft .Net Framework version 4. It can be downloaded from Microsoft.
Copy the GTrace distribution over to your computer and unzip if necessary. The folders can be located anywhere, but it is natural to put them on the C drive.
The program is run by clicking the GTrace.exe executable, which is found in the bin\Release folder of the distribution. (It is identified by the Gtrace icon  .) It must stay in that folder. It is probably easiest if a shortcut is made to this executable and then located in any convenient location (e.g. on the desktop). Clicking on the shortcut will then start the program.

Input and output

Gtrace reads clips stored in one specific subfolder and writes data to another. Both are in the folder 'Trace2011'.

- Trace2011 /media holds the clips that are to be rated, and a playlist that auto-fills with the contents of the Media subfolder - any AVI, MPG, WMA, MP4, MOV files in the Media folder will be loaded onto the list by default.
- Trace2011/data holds the output files that are produced when a rater traces a clip. An output file consists of text that gives details of the trial followed by numerical data which defines the trace. Later sections describe the data in the files.

3. The basic rating process

This section is designed to let users operate a version of the system that follows the core rating pattern for generating EmotionML files.

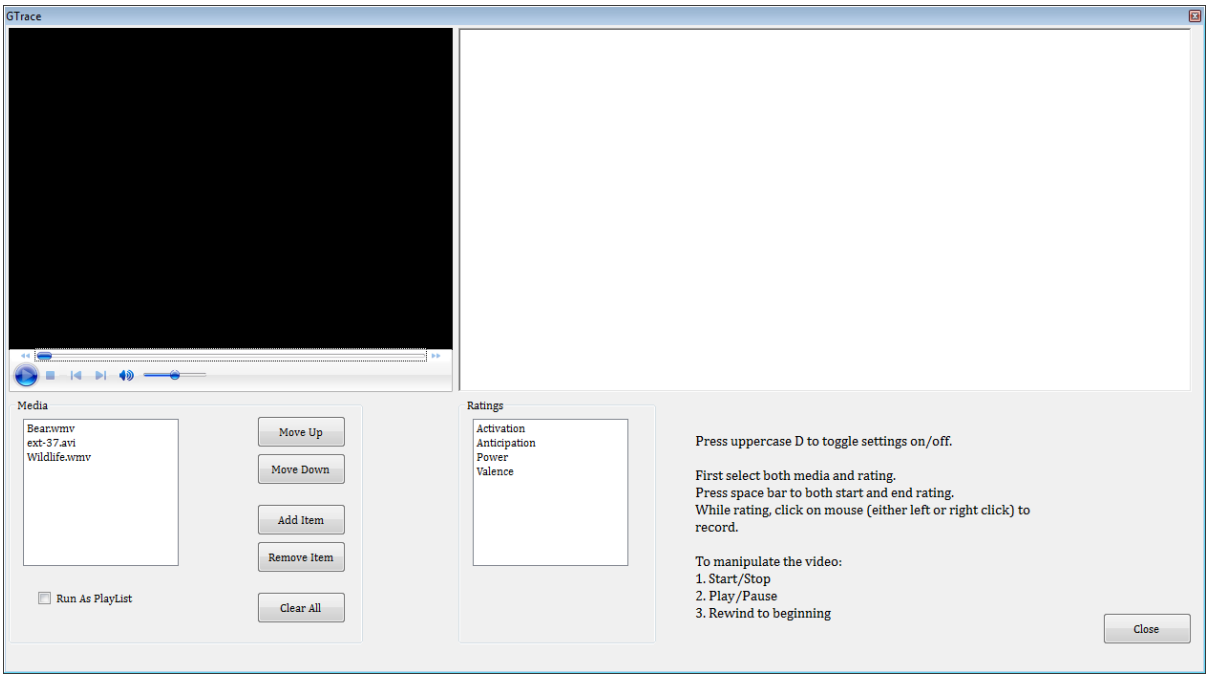
To start, click GTrace.exe or the shortcut to it. The first window provides several choices, but for this use of the system, only two things need to be done at this stage:

- in the box called 'identifier', enter a code that identifies the user (usually your initials),
- click in the two boxes above the identifier ('save data' and 'save xml')
- click 'continue'.

NB: The identifier code is used to name output files. Practically, it is important to make sure that each rater has a unique code, and that it is used consistently.

The next window asks the user to select one of the vocabularies that is available in EmotionML. Three are shown: "Big 6", "FRSE", and "Everyday". There is also an "other" option, and the user can set sampling frequency. These are explained later. For the time being, assume that "FRSE" is selected, and the user clicks "OK".

The main window now appears, as illustrated in the figure below. It has four main areas. At the top left is the 'screen' where the clip to be rated will show. At the top right is the panel where the rating scale will appear. Below the screen is the playlist from the media folder, which specifies the clip(s) available to be rated. Below the scale panel is a list of the scales that are available to use. The buttons beside the playlist allow items to be moved up and down in it.



Click to select one item from each of the lists - a clip and a scale. They will appear in the respective windows in the upper screen. Before going on, place the mouse pointer in a suitable position. A default might be at the centre or at the neutral end , depending on the scale. Ideally, the clip should be previewed using the numerical controls (see below), and the cursor put in a suitable position.

When both clip and scale have loaded, hold down the left mouse button and then press the space bar. The clip will begin to play, and the rating disc will appear at the position where the mouse pointer was left. It can be moved with the mouse (or mouse pad) in the normal way.

> *Note: The left mouse button is critical. So long as it is held down, the trace cursor will appear as a filled disc. The filled disc signifies that the mouse position is being written into the data file. When it is released, the trace cursor appears as a ring. The ring signifies that no data is*

being written to the data file. The option is useful because it provides a way of dealing times when the rater cannot make a judgment (e.g. the person being traced is out of view) – he/she simply releases the mouse button, and no data is entered until it is pressed again. The problem arises when a rater does not register that the button needs to be down in order to record data, and spends hours generating a series of empty files.

There are various ways to proceed after rating a clip. The default is described first, then the alternatives.

By default, the rater will continue tracing until the clip ends. Press the space bar to end rating of that clip (the program will wait until that happens). When the space bar is pressed, the same clip reloads, and the next scale in the series is chosen. Hold down the left mouse button and press the space bar to make a trace on that scale.

Note: The space bar has to be pressed to close the process of making a particular rating of a particular clip. An error message will appear if the user tries to move on to the next rating before pressing the space bar.
It may take some time for the clip to reload. Pressing 'space' to start the next rating before the clip has reloaded will produce an error message.

When the space bar has been pressed to end a clip, the rater always has the option of clicking on either or both of the lists on display – the list of clips and the list of scales – to select what will be rated next.

It is also possible at that stage to view the clip without tracing it. Instructions for doing that are below the scale. Press 1 to start the clip, 2 to stop it, and 3 to rewind to the beginning.

If an experimenter wants to restrict the options available to the user, pressing 'D' (uppercase) hides the lower part of the window, leaving only the screen and the scale on view. The program then follows the default sequence described above.

When all the relevant ratings are complete, pressing the 'exit' button at the bottom right hand corner ends the session.

4. The pre-formed scales

The scales supplied with Gtrace are in the subfolders Trace2011/Ratings and Trace2011/RatingsPlus. /Ratings provides scales included in EmotionML, and /RatingsPlus provides scales that were used in the SEMAINE project. The EmotionML scales generate xml files compatible with the w3C standard; the RatingsPlus scales provide output in an older format (though that may change).

EmotionML scales

The default process offers rating scales included in EmotionML. They are in four folders: "Big 6", "FRSE", "Everyday" and "Other". Relevant explanations and references are given in the EmotionML specification.

The folders "Big 6" and "Everyday" contain scales based on category terms – where the question is to what extent the person's emotions appear to involve an emotion type drawn from everyday language. The 'Big 6' are the so-called basic emotions which were proposed by Ekman. They are well known, but a large body of evidence suggests that they do not capture very much of the emotional

colouring that pervades everyday life. The 'Everyday' categories were chosen with precisely that in mind: they were empirically identified as the terms most likely to describe emotion in everyday life.

Dimensional descriptions have been used partly because they provide a way of describing the elusive states that typify everyday life. The FSRE scales are based on the work of Fontaine et al (2007), which showed that four dimensions were sufficient to discriminate the items in a theoretically interesting set of everyday emotion words. Three are variants on a set of dimensions that psychologists have used for a long time: valence, power, and arousal/activation. The fourth is less common, but reasonably intuitive. It is anticipation, i.e. the extent to which the person being traced appears to feel that events are proceeding as he/she would have anticipated.

The 'Other' folder contains a scale dealing one other dimension, intensity. It is available to add other scales derived from the EmotionML vocabularies.

Information on the reliability of all these dimensions is contained in McKeown et al (2012) and Cowie, McKeown & Douglas-Cowie (2012). Summarising, though, it seems fair to say that considering the inherent challenges of the area, these scales appear to be quite satisfactory instruments.

It should be noted that raters need prior explanation to use the scales effectively. Appendix 1 gives instructions to raters using the scales which are minimally adapted from the instructions which were used in SEMAINE, and which were reasonably effective.

The 'plus' option: SEMAINE scales

The 'plus' option lets raters use a procedure that was developed in the SEMAINE project. Its raison d'etre is that the category descriptors which feature in everyday language have an awkward property. Most of the time, they do not really apply at all: but there are episodes that they describe very well. Examples include most everyday words that are used to describe emotion, such as 'angry' and 'happy'. They could be studied by making a trace for each descriptor in each clip. However, that would be very frustrating, because for the vast majority of the time, the traces would only show that the descriptor in question had little or no relevance. The 'plus' option short-circuits that process by allowing the user to consider a wide range of descriptors, but to trace only those that are relevant to the clip in question.

The description below assumes that the scales provided are being used. The scales used in this option are stored in the subfolder Trace2011/ratingsplus. Their meanings are set out in appendix 2.

The process is entered by clicking the button 'Plus option' in the opening screen. As before, the operator also needs to enter a code that identifies the user in the box called 'identifier', and to click 'continue'.

The main window appears again. As installed, that brings up only one rating scale, intensity. Rating of intensity proceeds in the same way as before, but when the rater has traced intensity, a new window appears. It contains four panels, each of which shows a group of related scales. In the version provided, the groups are

1. Basic emotions (amusement, anger, contempt, disgust, fear, happiness)
2. Epistemic (agreeing, at ease, certain, concentrating, interested, thoughtful)

3. Validity (anomalous simulation, breakdown of engagement, sociable concealment, sociable simulation)
4. Interaction categories (Shows Solidarity, Shows Antagonism, Shows Tension, Releases Tension, Makes Suggestion, Asks for Suggestion, Gives Opinion, Asks for Opinion, Gives Information, Asks for Information).

In the procedure used by SEMAINE, the user chooses up to four of these items which seem relevant to describing the character of the clip being rated. Once the relevant words have been selected, the user clicks 'done'. The system then presents a scale for the first of the words that was selected. The user goes through the usual tracing procedure, tracing the salience of the attribute that the scale describes (for instance, if the word was 'anger', the trace records how the person's anger appears rise and fall over the clip). The system then presents a scale for the second word that was selected; and so on. The process repeats until all of the selected terms has been traced.

By comparison with the dimensional scales, evidence on the reliability of these scales is limited. Similar types of term have been widely used, but not usually to rate in real time, and data often use classes formed by combining the original terms post hoc.

5. Results files

By default, each rating session generates a data file, which is stored in Trace2011/data. The file name combines relevant identifiers, separated by underscores, in this order: clip name; scale; rater ID; date; starting time.

The format of the files depends on the source of the scale, Trace2011/Ratings or Trace2011/RatingsPlus

EmotionML output (from Trace2011/Ratings)

Below is an example of the xml output provided for scales in Trace2011/Ratings.

```
<emotionml version="1.0" xmlns="http://www.w3.org/2009/10/emotionml" xmlns:imdi="http://www.mpi.nl/IMDI/Schema/IMDI">

<info>
 <imdi:Actors>
   <imdi:Actor>
     <imdi:Role>Annotator</imdi:Role>
     <imdi:Name>John</imdi:Name>
   </imdi:Actor>
 </imdi:Actors>

 <imdi:Session_Type>
   <imdi:Date>250912</imdi:Date>
   <imdi:Time>15:45</imdi:Time>
   <imdi:Name>ext03.avi</imdi:Name>
 </imdi:Session_Type>

</info>

<emotion category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
                <category name="afraid">
                        <trace
                        freq="10Hz"
                        samples="0.315 0.535 0.513 0.309 0.531 0.464 0.568 0.622"/>
                </category>
        <reference uri="file:ext03.avi?t=3.24,15.4"/>
</emotion>

</emotionml>
```

The first line describes the (standard) web sources from which vocabulary is drawn – EmotionML and IMDI, which provides tools for describing annotations. The next block identifies the source and the annotator, using IMDI conventions. The last describes the trace itself, using EmotionML. The descriptions should be reasonably self-explanatory.

This format can only be generated for descriptions which use a vocabulary specified in EmotionML. The reason why it is not generated for scales in RatingsPlus is that not all of the SEMAINE terms are included in the EmotionML vocabularies.

SEMAINE output (from Trace2011/RatingsPlus)

This format is less sophisticated than the xml format described above, and it is always generated. The main body of each file contains two columns of numbers (separated by a space).

- Each number in the first column specifies a time
- Each number in the second specifies the position of the cursor at that time.

Call these the time co-ordinate and the scale co-ordinate.

The time co-ordinate takes the form of a time stamp taken from the clip. If the clip has reached the end, but the space bar has not been pressed, the time co-ordinate will be zero.

The scale co-ordinate is normalised using information in the scale definition file (see next section). The SEMAINE scales use two types of normalisation. In most cases, the values are normalised to lie between 0 and 1. However, for valence and some others, the extremes are -1 and +1 . Users can specify any extremes they choose – for instance, -100 to +100.

Additional information is given with these data columns, in two ways.

The file name takes the form C_S_I, where C is the name of the clip, S is the name of the scale, and I is the rater identifier entered in the initial window.

The first six lines in each file form a header. They specify

- Date and time of rating
- Media filename
- Scale image filename
- Scale extremes
- Name of the scale
- RGBs of the colours used in the rating

If a file with a particular name already exists, the new data will be **appended** to the existing data, starting with a new header section. The time stamp in the header can then be used to distinguish between different attempts to make the rating.

The files can be imported into EXCEL files, and (for instance) scattergrams can then be used to visualise the data. However, specialised analysis programs are needed to transform the files into a format that can be used statistically. Comments on those are made later.

6. Additional options.

Playlist

In the default version (but not the 'plus' version), there is an option that allows the same scale to be applied to all the clips in the 'media' file. In the main window, there is a box below the list of clips marked 'run as playlist'. When it is ticked, after one clip has been rated using a selected scale, the next clip is automatically brought up for tracing on the same scale.

Lists of scales

The 'plus' option allows a list of selected scales to be saved and brought back.

The standard way to make the list is from the screen that shows all the secondary scales. It offers the option of saving the selection that has been made.

If a list exists, it can be selected from the opening screen by selecting 'Load pre-selected' (the 'plus' option has to be selected first).

When a scale on the list has been used, the label 'rated' appears beside it. When the 'close' button is pressed, an window offers the option of saving an updated version of the list. The update indicates whether or not a rating has been completed for each scale in the list. As a result, if a user has to rate on a long series of scales, he/she can stop midway, save the list, upload it later, and see which scales still have to be completed.

Scale definition files

It is technically straightforward for users to create scales of their own.
Two files are needed to define a new scale. One is a .jpg file which includes the rectangle and any markings. The next section explains how to make the files. The other is of type .rtg. These are text files, with the following elements:

- The name of the scale as it appears in the main window ('valence', 'anger', etc)
- The name of the jpg file that contains the scale
- The numerical values of the scale extremes (e.g. 0,1 or -1, 1)
- The colour co-ordinates of the cursor at the bottom end of the scale
- The colour co-ordinates of the cursor at the top end of the scale

Optionally,

- The colour co-ordinates of the cursor at the midpoint of the scale (this is useful if, for instance, one wants the cursor to be a neutral gray at the midpoint of the scale).

Colour co-ordinates are of the standard form: red, green, blue where each has a value between 0 and 255.

For example, a file for a 'concerned' scale might be as follows:

Concerned
Concerned.jpg
0,100
255,255,0
0,255,255

Creating new scales

Technically, it is easy to create a new scale. It is another matter whether the scale makes sense: the next section considers the issues surrounding that.

Two files are needed to define a new scale: a .jpg file which shows the rectangle and any markings; and a .rtg file.

The easiest way to make a .rtg file is to make a copy of an existing file and modify it by editing in the normal way, using a basic editor (such as Notepad). It is assumed that users know how to use colour co-ordinates to define appropriate cursor colours. Websites can be used to preview the colour produced by particular sets of co-ordinates (e.g. http://mxipedia.com/rgb+color+picker), or it can be done via highlighting in Excel.

The main issue in making a .jpg file is to ensure that rectangle is correctly positioned relative to the cursor. That can be easily done by making a Powerpoint version, selecting the relevant part of it, and pasting it into Paint (or a comparable program) to make the .jpg.

The package includes a Powerpoint file called 'trace frames' which can be used for that purpose. The file begins with unlabelled frames divided into 2,3,4 and 5 parts. These are followed by the labelled SEMAINE scales. The scale and markings are enclosed within a box which is the right size and shape to fill the scale window in the program. To make a. jpg, select the box and everything in it; paste the selection into Paint; crop; and save the result as a .jpg (there are other ways, but this one is simple).

7. Issues relevant to new scales

Tracing can generate meaningless data (just as questionnaires can). This section identifies some of the key issues to be considered.

Tests of quality

The simplest test of quality was invoked earlier. It is reliability. Reliability is easy to measure, and if a scale is reliable, it can reasonably be assumed that it provides meaningful information.

A reliable scale is one for which different people tracing the same material generate similar profiles. That can be assessed in different ways. For example, SEMAINE data given above show that raters there agreed very well on the average valences of different clips; and also that for about 90% of clips, different raters tracing the same clips showed similar profiles. The two criteria diverge for anticipation: there was modest agreement on the average level of anticipation in different clips; but different raters still traced similar profiles for about 90% of the clips.

Several issues are relevant to maximising reliability.

- The text on the scale needs to be well chosen. In particular, it needs to convey broadly the same meaning to everyone (i.e. not to make it likely that some people will interpret the text in one way, others in another); and any landmarks need to mesh with the way people naturally use the scales (not, for instance, specifying that everything less than intense

emotion is to be located in the bottom third of the scale, and the top two thirds is to discriminate among degrees of intense emotion).
- An appropriate level of training needs to have been provided.
- Users who have unusual difficulty with tracing need to be indentified and filtered out (most obviously by asking potential raters to trace clips for which there is a large amount of data, so that norms can be identified with some confidence.

On a different level, but critical, reliability measures depend on the diversity of the data. It is very difficult to get high reliability scores if all the sources being considered are quite similar with respect to the attribute being considered. Hence sensible measurement depends on assembling a suitable diverse set of sources.

Reliability is an obvious issue to consider, but on the other hand, it should not be assumed that it is the only interesting criterion. Everyday experience suggests strongly that people do sometimes diverge in their reading of another person's emotions. If a scale effectively exposes the fact that people's reading of a particular display diverge, then standard measures will indicate that it is not reliable, but it will nevertheless be highly informative. To illustrate the point, a recent study suggested that ratings of the intensity of emotion followed two distinct patterns. It appeared that some raters assumed that intense emotion was present when there were visible or audible signs of it, and absent otherwise; others indicated that the underlying level of emotion changed much more slowly than the external signs. If that reflects a genuine divergence in people's reading of emotion, then the fact that the scale does not show high reliability is not a measure of its value. It exposes what is there, which is variability.

The literature on sources of variability is uneven. A large body of evidence indicates that the rater's own emotional state will affect his/her judgments of others', and there is good evidence that different cultures differ rather systematically in their reading of the same displays (Sneddon et al 2011). Some evidence suggests that different features of the information are salient for different individuals (Cowie & Douglas-Cowie 2009).

A final point on this topic is that one of the uses of tracing is to provide a model from which a machine can learn. Initially it was automatically assumed that the data should reflect some kind of ground truth, or at least an average response pattern. It seems increasingly likely that it may be more useful to model an individual. If not, there is a real risk of teaching the machine to respond in a way that human beings find anomalous, because it is modelled on a composite response that no one person would actually give.

For more information on these issues, see Cowie, Cox et al (2011), particularly section 3.1.7.

Scales that have been used

One way and another, quite a few scales have now been tested. The literature on them gives pointers to their properties (often including reliability).

Historically, several tracing tools emerged around the same time, notably *EmotionSpace Lab* (Schubert 1999, 2001) and FEELtrace (Cowie et al 2000, Cowie & Cornelius 2003). Both have interfaces that are significantly different from Gtrace, and so information about them does not transfer directly.

The design of Schubert's system reflects the fact that his primary interest was in music. Hence he developed an interface which assumed that raters could give their full visual attention to the tracing process. In contrast, FEELtrace was designed to let users capture the feedback that they needed from a glance while their main focus was on an audiovisual display of emotional behaviour. The devices that GTrace uses – a large coloured cursor, a clear frame, and landmarks – descend from that. However, the FEELtrace cursor could be moved in a two-dimensional space, so that raters could record impressions of both valence and activation/arousal simultaneously. That is quite difficult, and the format does not transfer to dimensions other than valence and activation.

A wider range of trace measurements was explored by Devillers et al (2006), using scales broadly like Gtrace. The main results are summarised in Table 3. The reporting format reflects the fact that agreement was distributed rather unexpectedly. For instance, in sadness, half of the pairs of traces correlated quite strongly (r>0.5), but there was also substantial proportion of cases where the traces tended to move in opposite directions (r>-0.5). Note, though, that the ratings were more consistent with the established dimensions (intensity, valence, and activation) than when people traced everyday emotion terms.

Scale	Proportion of trace pairs whose correlations were			
	very –ve (r < – 0.5)	weakly –ve (–0.5< r <0)	weakly +ve (0< r < 0.5)	very +ve (r > 0.5)
intensity	0.00	0.03	0.08	0.89
true emotion/not	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
masking	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

Table 3: Distribution of correlations between traces involving the same dimension form Devillers et al

The HUMAINE project (Douglas-Cowie et al 2011) built on Devillers et al, and used the traces at the top of table 3, plus the anticipation scale which is included in GTrace.

The SEMAINE database (McKeown et al 2011) used the scales included with GTrace, and information about them has already been given. It also included a measure of a speaker's apparent engagement in or disengagement from a conversation, Engage-trace, but did not measure its reliability.

Scale	Average inter-trace correlation
Intensity	0.23
Agree/disagree	0.35
Amused/not	0.23
Shows antagonism	0.3
Shows tension	0.43

Table 4: Average inter-trace correlations for scales from the GTrace Plus option.

Five of the scales provided with GTrace were compared In a recent study, with 7 raters applying them to the same clips. The average correlations between tracings are shown in Table 4 above. Tension is a dimension considered in various musical studies. It is natural to infer from Table 4 that it is a scale that elicits relatively high agreement. There is a caution, though. The recordings rated were a very particular type of interaction (a confrontation which was lively but generally good humoured). It may be that tension was simply the most straightforward feature of that particular situation.

The balance of evidence suggests that there is a moderate number of attributes which can be traced reliably, and a much larger number which cannot. However, it is still far from clear what belongs in which category, and even less clear what the absence of agreement means. It may be to do with the nature of tracing, or to do with the fact that people genuinely differ in their perceptions of the attribute. GTrace allows users to explore issues like that as well as to use scales that are known to be robust.

8. Analysis of trace data

GTrace samples cursor position as fast as it can, and for every sample, it writes the time and the cursor position into the output file. That produces files which contain all the detail that the machine's speed can provide; but they need to be processed before they can be analysed statistically. The procedures are straightforward for teams who are used to processing their own data. For others, programs developed in Belfast will be made available gradually.

Four basic steps are almost always needed

1. Identifying information about a trace needs to be extracted from the name and the header.
2. Irrelevant material needs to be removed. It arises in two main ways. When the same person has applied the same scale to the same clip more than once, data from all of the sessions will be in the same file, and the right one (usually the last) needs to be selected. Also, data may be entered when the space bar has been pressed, but the clip is not running. It is distinctive because the time co-ordinate is zero.
3. Missing data need to be handled. The usual reason for missing data is that the rater could not make a judgment during a particular interval, and released the mouse button. That is shown by gaps in the sequence of time stamps.
4. The raw data needs to be organised into appropriate time intervals. Usually, that means identifying an interval (say *i* seconds) and finding the average cursor position during each *i*-second interval. That raises issues which are taken up shortly.

It may or may not be useful to apply other transformations to the data. Questions about the best way to analyse the data also arise. Those are taken up shortly.

Several of the issues that arise from this outline are quite challenging. The aim here is simply to convey a broad sense of the issues. There are two extended discussions of them, one by Schubert (2010) and one by Cowie & McKeown (2010), which is available from the SEMAINE project website.

One clear question is to find suitable sampling intervals. The most basic form of a trace consists of points located at constant time intervals, but what should the interval be?

That question is often linked to the idea that there is a problem with applying correlation-based techniques directly to time sequence data, because it tends to produce inflated correlations. Instead, points which are clearly separated in time should be used for analysis. In fact the description of the problem is not generally true, as Figure 1 below shows. It takes traces of a single clip from a large dataset (described by Sneddon et al 2011) which is sampled at 10Hz, and shows the effect on average correlation of sampling at lower rates - once per 3 seconds, once per 5 seconds, and once per 6 seconds. Clearly it is far from the truth that using widely spaced samples is conservative.

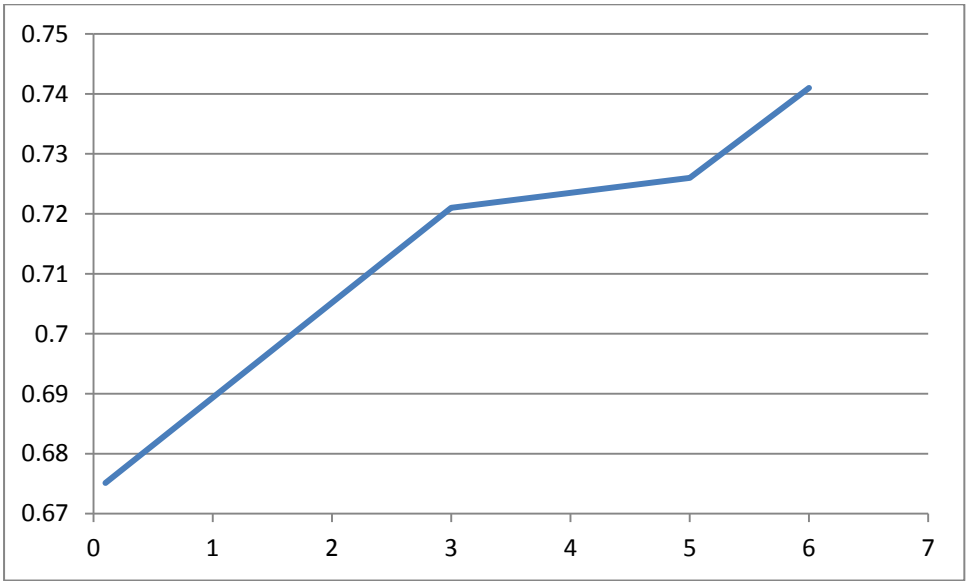


Figure 1 average inter-trace correlation as a function of the interval between samples (in seconds)

A little thought makes the reason obvious. Close sampling does give inflated correlations when traces are of a particular form – i.e. they can be modelled by a relatively small number of points connected by smooth functions. In that case, correlating the raw traces amounts to double counting a small number of data points. However, tracing data often does not take that form. On the contrary, ratings fluctuate quite sharply over the short term, and it is in the short term fluctuations that the disagreement lies. Taking widely spaced samples masks that.

It should be clear that this is an area where blanket prescriptions need to be treated with caution. The real issue is at what temporal scale information exists. Cowie & McKeown (2010) explored the question by averaging data over 'bins' of varying duration, and showed that the effect of bin size varied sharply according to the trace and the sample. However, they conclude that "overall, the safest choice of bin size seems to be in the region of 1-3 seconds. Increasing bin size beyond that often increases intercorrelations, but rarely by very much. Conversely, with bin sizes of that order, the level of agreement on internal structure (which the operation of binning throws away) is rarely very high; and it often increases quite substantially with larger bins".

A different way of organising trace data is stylisation, where raw traces are represented by a series of rises, falls, and level stretches. That makes it possible, for instance, to discriminate between traces that show a few large rises and falls and traces that show smaller movements – as in the two examples in figure 2. Cowie & McKeown (2010) describe techniques for obtaining stylisations, and show that they yield stable measures for some kinds of trace.

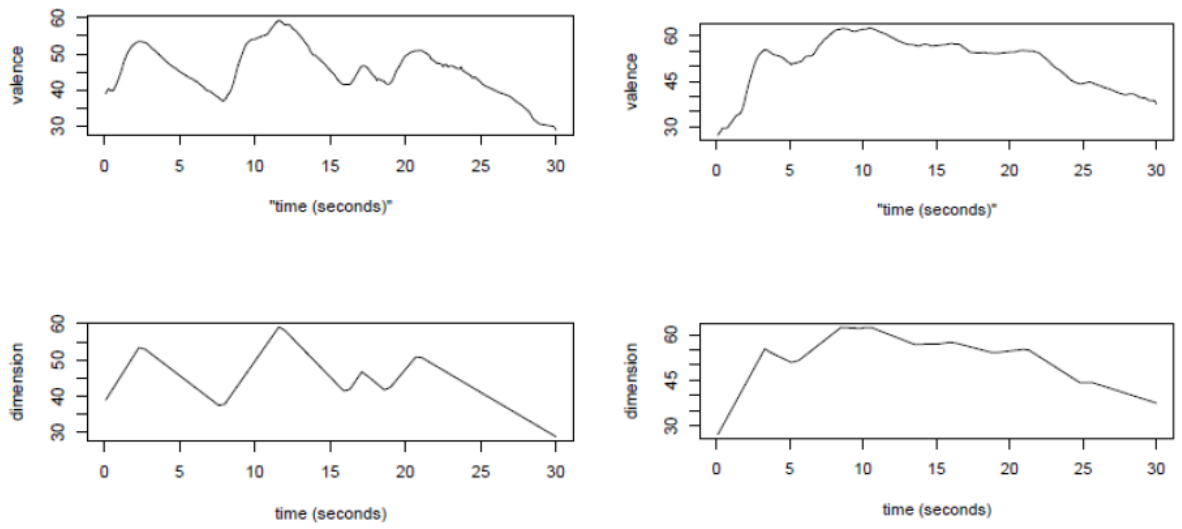


Figure 2: examples of stylisation (From Cowie & McKeown 2010).

Cowie & McKeown also describe an approach to analysing agreement between tracers which is quite different from correlation. It is called QA (for Qualitative Agreement). It considers each pair of points in a trace, and asks whether raters agree that one is higher or lower, or that they are about equal. That gives a measure which detects agreement even when raters do not use the scale in the same way. There are some scales where that kind of agreement does exist, but correlational measures of reliability are weak. That provides useful information about the scales.

Tools for these analyses are described in Cowie & McKeown, and will be made available over time to accompany GTrace.

9. Beyond reliability

It is frustrating to have written so much about measuring reliability when it is clear that reliability is far from an ideal measure of what a trace delivers. It would be much preferable to have better measures.

The QA technique described above is a step in that direction. It identifies relationships on which people agree, but where they do not, it is content to say that they disagree: it does not try to make an estimate of what is happening. That seems a rational way of acknowledging that a complex recording will provide some points that people are clear on, but others on which they are not. It would be useful to extend the strategy of homing in on points that are agreed.

A more radical solution is to exploit synthesis technology, and use traces to synthesise displays (most obviously a face) showing the sequence of emotions specified by the traces . The tracer can then indicate whether it appears to show the same emotions as the original. Similar techniques have been used for other purposes, but not for evaluation of traces.

10. References

Baron-Cohen S., Golan, O. Wheelwright, S. and Hill J., *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004

Cowie, R. & Cornelius, R. (2003) Describing the Emotional States that are Expressed in Speech Speech Communication vol 40, 5-32

Cowie, R., Cox, C, Martin, J-C, Batliner, A., Heylen, D. & Karpouzis, K, (2011) Issues in Data Labelling In Petta P., Pelachaud C. & Cowie R. (eds) Emotion-Oriented Systems: The Humaine Handbook Springer-Verlag Berlin Heidelberg pp 215- 244

R. Cowie and E. Douglas-Cowie, "Prosodic and related features that signify emotional colouring in conversational speech," in The Role of Prosody in Affective Speech Studies in Language and Communication, S. Hancil, Ed. Berne: Peter Lang, 2009, vol. 97, pp. 213–240

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schr¨oder, "'FEELTRACE': an instrument for recording perceived emotion in real time," in Proc. ISCA Workshop on Speech and Emotion, Northern Ireland, 2000, pp. 19–24.

Cowie, R. McKeown, G. & Douglas-Cowie, E. (2012) Tracing Emotion: An Overview International Journal of Synthetic Emotions 3, 1-17

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

Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, Anton Batliner, and Florian Hönig (2011) The HUMAINE Database In Petta P., Pelachaud C. & Cowie R. (eds) Emotion-Oriented Systems: The Humaine Handbook Springer-Verlag Berlin Heidelberg pp 245-288

J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not Two-Dimensional," Psychological Science, vol. 18, no. 12, pp. 1050–1057, 2007. [Online]. Available: doi:10.1111/j.1467-9280.2007.02024.x

McKeown, G, Valstar, M. Cowie, R., Pantic, M, & Schr¨oder, M (2012) The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent *IEEE Transactions of Affective Computing* 3, 165-183.

Sneddon I, McKeown G, McRorie M, Vukicevic T (2011) Cross-Cultural Patterns in Dynamic Ratings of Positive and Negative Natural Emotional Behaviour. PLoS ONE 6(2): e14679. doi:10.1371/journal.pone.0014679

Schubert, E. (1999) Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space *Australian Journal of Psychology* 51, 154-165

Schubert, E. (2001) Continuous measurement of self-report emotional response to music. In P.N.Juslin & J.Sloboda. *Music and emotion: Theory and research* 393-414

Schubert, E. (2010) Continuous self-report measures In P.N.Juslin & J.Sloboda. Handbook of *Music and emotion: Theory, research*, applications Oxford: Oxford University Press 223-254

Appendix 1 : A guide to using the dimensional scales

The aim of these scales is to let people record their impressions of emotionally coloured events. The descriptions that the scales use are designed to mesh with core features of the impressions that observers form naturally. That means that the instructions are about pointing people in the general direction of a kind of judgment that it is natural to make, and letting them register for themselves that their response includes something in that general area. That is very different from instructions designed to provide rules that people will follow to the letter.

The core reason for that is that tracers have to be responding in a fluent, intuitive way. They have to react in real time, and so they cannot always be referring back to rules and trying to ensure that they follow them. Tracing has to be like riding a bicycle, not like doing long division. The ideal is that the action of tracing should flow directly from a natural impression. If the process is dominated by attempts to follow written rules, it is not achieving that.

Nevertheless, people need help to get the hang of riding a bicycle. These instructions are meant to give that kind of guidance.

General

Tracing is usually concerned with the state of a particular person in a clip. If several people feature in a clip, it has to be made clear in advance which should traced.

In all cases, trace the individual's state as you perceive it – not, for instance, the evidence that you detect; or what you think you are justified in claiming; or change in the state.

If you are seriously uncertain, you can let go of the mouse button. When you do that, the cursor disc turns into an empty circle, and no data is recorded. Remember to press it down again when you feel you have a firmer impression and want to trace again.

The scales that the programs provide are marked out with labels at the ends and at some of the lines that subdivide them. The labels are chosen to reinforce intuitive ideas about the way possible states are distributed across the scale. In most cases, they are based on pilot studies which looked at the ways people found it natural to describe different levels of emotion using scales and words. The scales you are using put the two together in a way that is meant to confirm what people feel is a natural match between scales and words.

There are four individual scales in this package.

Valence

Valence is the individual's overall sense of 'weal or woe' – does he/she on balance feel positive or negative about the things, people, or situations at the focus of his/her emotional state?

It is a 'cover term'. At any given moment there may be some things that the individual feels positive about and some he/she feel negative about; there may be kinds of positiveness that he/she does feel and others that he/she doesn't. Tracing is not about making that kind of nuanced distinction. It is based on the assumption that observers do find it natural to say how positive or negative another person is overall and on balance. It is rather like judging how speedy a car is. Of course a car with great top speed does not necessarily have the best

acceleration from a standing start – but we do bundle all the separate measures into an overall sense that this is a fast car, and that is not. VALENCEtrace is about accessing a similar global impression of positiveness.

Activation

Activation is the individual's overall inclination to be active or inactive. Like Valence, activation is a cover term. Activation may include mental activity as well as physical, preparedness to act as well as overt activity, alertness as well as output, directed or undirected thought. Again, FEELtrace is not for making nuanced distinctions between those kinds of state. It is about giving what you think is an natural overall summary. The word arousal is sometimes used instead – the difference is relevant to some theoretical issues, but not generally to making a trace.

Power

The power rating scale deliberately mentions two related concepts, power and control. Obviously they are not the same conceptually – power is mainly about internal resources, control is about the balance between those resources and external factors. However, emotion is about people's sense of their own power, and that seems to be relative to what they are facing. The tracer's job is to rate his/her impression of that composite sense of being in a position to direct events rather than being at their mercy.

Anticipation

The last trace also uses a variety of related terms – expecting, anticipating, being taken unawares. Again, they are not the same conceptually, but they point to a dimension that seems to make sense – related to control in the domain of information. The tracer's job is to rate his/her impression of that composite sense of being ahead of events rather than lagging behind them.

You should not expect to read these notes and then immediately settle into tracing. Most people need to play with the tracing process for a while before they settle into it. If after reading the notes, and experimenting with the system, you don't find there is something that comes reasonably naturally, then stop. A good experimenter will check that you have settled by giving you examples that most tracers agree on, and making sure that your responses are not wildly out of line.

Appendix 2 : A guide to using the SEMAINE 'plus' scales

The basic ideas behind tracing have been described in a separate introduction (Appendix 1). This describes a particular type of rating used in SEMAINE. It has two levels. The first level involves the most global scale of all, the apparent intensity of the target person's emotion. The second level is selective follow-up. It involves picking out categorical descriptions that may identify the particular emotional or emotion-related state that is present, and tracing how each of the relevant states seems to come and go from moment to moment.

Level 1: **Intensity**
There are many ways of thinking about intensity. It could be argued, for instance, that it should be relative to the emotion in question, so that the most intense curiosity possible is regarded as equal to the most intense fury possible; or that intensities of different emotions are impossible to compare.

The intensity trace is based on the assumption that there is simpler kind of judgment that people do not find it difficult to make, however awkward it may be to explain what they are judging. It is roughly speaking about how far the person is from a state of pure, cool rationality, whatever the direction. A good deal of evidence suggests that it is a judgment most people can make quite reliably.

Level 2: Categorical descriptions
The second level develops a pattern that was used in work with the older FEELTRACE system. After the rating of intensity, the user is presented with an array of words, and asked to select those that apply particularly well to the clip that has just been rated for intensity.

Once the relevant words have been selected, the user clicks 'done'. The system then presents a scale for the first of the words that was selected. The user replays the clip, and traces the salience of the attribute that the scale describes (for instance, if the word was 'anger', the trace records how the person's anger appears rise and fall over the clip). The system then presents a scale for the second word that was selected; and so on. The process repeats until all of the selected terms has been traced.

Part of the rationale behind POLYtrace is economy. It would be not only extraordinarily time-consuming to trace all the descriptors that research might be interested in, but also very frustrating, because most of the traces most of the time would only show that the relevant descriptor did not apply in any useful way. POLYtrace short-circuits that process by allowing the user to trace only descriptors that are particularly relevant.

The purpose of the exercise is defeated if raters are too inclusive. It is probably true that almost every clip shows something that may be a mild flicker of annoyance. Hence, an overinclusive rater could end up making traces for every word in the list of alternatives, most of which never showed more than a small departure from zero. It is practically essential to avoid that. The problem is to set a sensible cut-off.

In this case, the primary criterion for choosing term x should be that the clip contains a relatively clear-cut example of x. The main function of the tracing is then to pinpoint where that example occurs. A secondary criterion is that it should be rare to choose more than four terms for a single

clip. A third is that if a substantial number of terms could be applied, and other things are equal, terms that are different in meaning from others that have already been rated should be preferred to terms that are similar.

Like everything else in the tracing system, selection has to be intuitive. One of the outcomes of preliminary rating is to establish how well the system seems to work. If it is unworkable, then it will have to be changed.

In the first program that used this approach (Intenstrace plus), the descriptive words were derived from earlier work on emotion per se. The SEMAINE list used here includes a few emotion words, but it also includes items related to cognitive or communicative functions that may have taken place during the clip. They fall into four subgroups, which are outlined below.

Basic emotions
There is a widespread belief that basic emotions are important points of reference even in material that involves emotional colouring rather than prototypical emotional episodes. Hence the commonest list of basic emotion terms, Ekman's 'big six', is included. To it is added a category that would not otherwise be represented, amusement.

It is important here to remember that the question is whether the clip contains a relatively clear-cut example of x. One of the uses of the material is to identify training examples. From that point of view, recording that there may be a subtle undertone of anger is actually counterproductive.

Validity
The fundamental reason for including this group is to weed out bad data. The recordings used in SEMAINE included sequences where an interaction designed to elicit emotion was simply not working. Those had to be excluded (or at least separated out).

The items were designed to exclude two main kinds of material. One is where one or more participants are not engaging with the interaction – they are thinking of other things, looking elsewhere, ignoring what the other party says, rejecting the fiction that they are speaking to a simulated character rather than to an actual experimenter, etc. The other is where there is a level of acting that suggests the material is likely to be structurally unlike anything that would happen in a social encounter. The main hallmark is that the expressive elements do not go together in a fluent or coherent way – they are protracted or separated or incongruous.

The options are summarised as follows
Breakdown of engagement
The labelling task is to identify periods where at least one of the participants is not engaging with the interaction. What that means is explained above.
Anomalous simulation (bad acting)
The labelling task is to identify periods where there is a level of acting that suggests the material is likely to be structurally unlike anything that would happen in a social encounter. What that means is explained above.
Marked sociable concealment

This is concerned with periods when it seems that a person is feeling a definite emotion, but is making an effort not to show it. In contrast to the two categories above, this is something that occurs in everyday interaction. It is an aspect of what Ekman calls display rules.

Marked sociable simulation

This is concerned with periods when it seems that a person is trying to convey a particular emotional or emotion-related state without really feeling it. Again, this is something that occurs in everyday interaction. People simulate interest or friendliness or even anger that they do not feel, not necessarily to deceive, but to facilitate interaction.

Epistemic states

These states were highlighted by Baron-Cohen et al (2004), and have a roused a lot of interest in the machine perception community. They involve a particular interaction of feelings and cognition. The items chosen are relatively self-explanatory:

Certain
Agreeing
Interested
At ease
Thoughtful
Concentrating

As before, the question is whether the clip contains a relatively clear-cut example of x. If there is an episode where the fact that someone is certain about something stands out, then raters should pick it out; but they should not select 'certain' if they simply feel that the person probably has no active doubts about something (e.g. that the sun will rise tomorrow).

IPA categories

The descriptors offered here are a subset of the system of categories used in Interaction Process Analysis [32]. The labels offered are:

Shows Solidarity
Shows Antagonism
Shows Tension
Releases Tension
Makes Suggestion
Asks for Suggestion
Gives Opinion
Asks for Opinion
Gives Information
Asks for Information

They are relatively self-explanatory.

Appendix 3 : Practical guidelines for experiments using Gtrace

1. Take care over your selection of scales. If you are not using well-proven scales, make sure that you have appropriate landmarks, and above all well-defined endpoints.

2. Make sure that the items you need are in the appropriate folders – clips in media, and scales are in Ratings or RatingsPlus depending on the form of the experiment.

3. Raters don't need elaborate training, but they do need to be briefed and to get a 'hang' of the system. Appendices 1 & 2 give briefings for the scales used in SEMAINE, and even if they are not directly relevant, they can be used as models. They make more sense if people can see the system alongside them.

4. Make sure that each rater has an identifier that is unique, and uses it consistently.

5. Make sure that raters know their data will only be recorded while they are holding the mouse button down.

6. A few people really don't get the hang of tracing, and it is wise to have ways of identifying them – you may or may not exclude them, but you should know who they are.

7. If there is more than one person in a particular clip, raters need to be told in advance who they are to trace.

8. There is a lot to be said for having seen a clip before you trace it. That can be done using the number controls (1 to run the clip, 2 to stop, 3 to go back to the beginning). If you are previewing, it makes sense to use the preview to place the cursor in the right starting position.

9. After tracing a clip, and pressing the space bar, allow time for the next clip to load. Trying to rush at that stage can generate an error message.

10. In principle, any length of clip can be traced. In practice, it is difficult to hold the length of attention that you need for more than five minutes without a break.

11. In the current version of the program, the media that GTrace can play are the video types supported by Windows Media Player – types AVI, MPG, WMA, MP4, MOV . Later versions will allow other formats, such as .WAV.

12. GTrace uses the codecs that are available to Windows MediaPlayer. If you can't play a video, you may not have the relevant codec. Installing VLCplayer from videoLAN.org may solve the problem.