The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data

Ellen Douglas-Cowie[1], Roddy Cowie[1], Ian Sneddon[1], Cate Cox[1], Orla Lowry[1], Margaret McRorie[1], Jean-Claude Martin[2], Laurence Devillers[2], Sarkis Abrilian[2], Anton Batliner[3], Noam Amir[4], and Kostas Karpouzis[5]

[1] Queen's University Belfast, Belfast, Northern Ireland, United Kingdom
e.douglas-cowie@qub.ac.uk
[2] LIMSI-CNRS, Spoken Language Processing Group, Orsay Cedex, France
[3] Lehrstuhl fur Mustererkennung, Universität Erlangen-Nurnberg, Germany
[4] Dept. of Communication Disorders, Tel Aviv University, Israel
[5] Institute of Communications & Computer Systems, National Technical University Athens

Abstract. The HUMAINE project is concerned with developing interfaces that will register and respond to emotion, particularly pervasive emotion (forms of feeling, expression and action that colour most of human life). The HUMAINE Database provides naturalistic clips which record that kind of material, in multiple modalities, and labelling techniques that are suited to describing it.

1 Introduction

A key goal of the HUMAINE project was to provide the community with examples of the diverse data types that are potentially relevant to affective computing, and the kinds of labelling scheme that address the data. Data has been collected to show emotion in a range of contexts. The database proper is a selected subset of the data with systematic labelling, mounted on the ANVIL platform [17]. It is designed to provide a concrete illustration of key principles rather than to be used as it stands in machine learning. Stage 1 (available via the HUMAINE portal at www.emotion-research.net) contains 50 'clips' from naturalistic and induced data, showing a range of modalities and emotions, and covering a balanced sample of emotional behaviour in a range of contexts. Emotional content is described by a structured set of labels attached to the clips both at a global level, and frame-by-frame, showing change over time. Labels for a range of signs of emotion have also been developed and applied to a subset of the clips: these include core signs in speech and language, and descriptors for gestures and facial features that draw on standard descriptive schemes.

2 Background

Even in the early part of this decade, most databases of emotion were unimodal, acted/posed, and focused on a few full blown emotions (e.g. [15], [16], [20], [29])

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 488–500, 2007.

Labelling was at a very basic level. However, researchers were increasingly experimenting with data from a range of induced and more naturalistic settings (e.g. [1], [2], [3], [6], [10], [12]).The move was undoubtedly related to the recognition that systems trained on acted stereotypical data do not transfer to more everyday situations [5].

However, the early work on naturalistic data exposed a number of problems. 'Truly' natural data tended to be noisy, making machine analysis difficult, and there were problems with copyright. It also became apparent that labelling naturalistic data was far from straightforward, and that time needed to be spent on developing appropriate labelling schemes [9]. The call center datasets which had become popular held several attractions (genuine data, dialogue material and clear applications). But with these advantages came limitations. The frequency with which emotion is expressed was low, the nature of the interaction imposed constraints on the forms of utterances, raising major questions about generalisability, and the emotions tended to be from a narrow range, generally negative.

Work on a HUMAINE database began in response to that situation. Three priorities were identified as key to theoretical progress.

1. Range of content. The data should reflect the range of ways in which emotion in the broad sense ('pervasive emotion') enters into everyday life. That involves showing a wide emotional range (negative to positive, active to inert) across a range of contexts, involving action and interaction across a range of contexts. This involves, for instance, moving from emotion in monologue to emotion in sedentary interaction to emotion in action, again starting with data that is reasonably tractable and moving forward to more complex data.
2. Multimodality. The data should be fully multimodal; at least audiovisual, but also involving some physiological recordings and performance data. Again multimodality is approached in gradations, starting from audiovisual recording and adding other modalities where it makes sense.
3. Labelling. Labelling schemes based on the psychological literature should be developed to capture the emotional content. These should span a range of resolutions in time (whole passage to moment by moment). Sound schemes should also be developed to describe signs of emotion, particularly vocal and gestural.

These priorities are linked to each other and to wider issues. Multimodality provides incentives to achieve range (e.g to find situations where both prosody and choice of action are informative); the adequacy of a labelling scheme depends on its ability to cover diverse types of data; and so on. It is implicit in these goals that the data should be available to the community in general. This entails ethical clearance and full consent of the participants. It is also implicit that the data are likely to be 'provocative' rather than 'supportive' – putting technology in a better position to identify significant challenges rather than providing material that is ready for practical use. That has led to internal debate. Some argue that only databases relevant to specific applications are worth developing, others that considering only short term goals carries a high risk of falling into local minima. It is an issue that will only be resolved in the long term.

3 Resources

3.1 Basic Resources

The database is underpinned by a large collection of recordings, only some of which are labelled. In the long term as many of the recordings as possible will be made available. In the short term a selection of episodes from these recordings form the HUMAINE Database. The sections below describe the core characteristics of the material. For convenience the recorded material is split into two categories, 'naturalistic' and 'induced'.

Naturalistic data
Belfast Naturalistic Database [10]

Nature of material: The database consists of audiovisual sedentary interactions from TV chat shows and religious programs, and discussions between old acquaintances.

Technical info & availability: 125 subjects (2 sequences of 10-60 secs each, 1 neutral 1 emotional); selection of 30 sequences with ethical and copyright clearance available.

EmoTV Database (in French) [9]

Nature of material: The EmoTV Database consists of audiovisual interactions from TV interviews - both sedentary interactions and interviews 'on the street' (with wide range of body postures)

Technical info & availability: 48 subjects (51 sequences of 4-43 secs per subject in emotional state); copyright restrictions prevent release.

Castaway Reality Television Database

Nature of material: This consists of audiovisual recordings of a group of 10 taking part competitively in a range of testing activities (feeling snakes, lighting outdoor fires) on a remote island. The recordings include single and collective recordings and post-activity interviews and diary type extracts.

Technical info & availability: 10 tapes of 30 minutes each; copyright clearance

Emotional content: All of these were chosen to show a range of positive and negative emotions. Intensity is mostly moderate, though EmoTV and Castaway contain more intense material.

Induced data
Sensitive Artificial Listener: (http://emotion-research.net/deliverables/D5e_final.pdf)

Nature of material & induction technique: The SAL data consists of audiovisual recordings of human-computer conversations elicited through a 'Sensitive Artificial Listener' interface designed to let users work through a range of emotional states (like an emotional gym). The interface is built around four personalities – Poppy (who is happy), Obadiah (who is gloomy), Spike (who is angry) and Prudence (who is pragmatic). The user chooses which he/she wants to talk to. Each has a set of stock responses which match the particular personality. The idea is that Poppy/ Spike/ Obadiah/ Prudence draws the user into their own emotional state.

Emotional content: There is a wide range of emotions but they are not very intense.

Technical info & availability: Data has been collected for 4 users with around 20 minutes of speech each. SAL has also been translated into Hebrew (at Tel Aviv University) and Greek (at National Technical University of Athens, ICCS) and adjusted to suit cultural norms and expectations, and some initial data has been collected. The data has ethical permission and is available to the research community.

Activity Data/Spaghetti Data

Nature of material & induction technique: Audiovisual recordings of emotion in action were collected using two induction techniques developed in Belfast. In the first, volunteers were recorded engaging in outdoor activities (e.g mountain bike racing). The second used a more controlled environment where certain kinds of 'ground truth' could be established. It is called the Spaghetti method, because participants are asked to feel in boxes in which there were objects including spaghetti and buzzers that went off as they felt around. They recorded what they felt emotionally during the activity.

Emotional content: Method 1 elicited both positive and negative emotions with a high level of activation. Method 2 elicited a range of brief, relatively intense emotions - surprise, anticipation, curiosity, shock, fear, disgust.

Technical info & availability: Method 1 produced 'provocative' data which was very fast moving and had a noisy sound track. Method 2 produced data where the participants were reasonably static and stayed within fixed camera range, making it easier to deal with face detection. The audio output consists mainly of exclamations. There are now recordings of some 60 subjects. The data has ethical permission and is available to the research community.

Belfast Driving simulator Data [22]

Nature of material & induction technique: The driving simulator procedure consists of inducing subjects into a range of emotional states and then getting them to drive a variety 'routes' designed to expose possible effects of emotion. Induction involves novel techniques designed to induce emotions robust enough to last through driving sessions lasting tens of minutes. Standard techniques are used to establish a basic mood, which is reinforced by discussions of topics that the participants have preidentified as emotive for them. The primary data is a record of the actions taken in the course of a driving session, coupled with physiological measures (ECG, GSR, skin temperature, breathing). It is supplemented by periodic self ratings of emotional state.

Emotional content: 3 emotion-related conditions, neutral, angry, and elated

Technical info & availability: 30 participants; will be available pending completion of PhD on the data

EmoTABOO: developed by LIMSI-CNRS and France Télécom R&D:[30], [31], [21]

Nature of material & induction technique: EmoTABOO records multimodal interactions between two people during a game called Taboo..One person has to explain to the other using gestures and body movement a 'taboo' concept or word.

Emotional content: range of emotions including embarrassment, amusement

Technical info & availability: By arrangement with the LIMSI team.

Green Persuasive Dataset

Nature of material & induction technique: The dataset consists of audiovisual recordings of interactions where one person tries to persuade another on a topic with multiple emotional overtones (adopting a 'green' lifestyle).

Emotional content: Complex emotions linked to varied cognitive states and interpersonal signals.

Technical info & availability: 8 interactions of about 30 mins each, and associated traces made by the interviewees to indicate how persuaded they felt from moment to moment. The data has ethical permission and is available to the research community.

DRIVAWORK (Driving under Varying Workload) corpus

Nature of material & induction technique The DRIVAWORK corpus has been collected at Erlangen, using a simulated driving task. There are three types of episode: participants are recorded relaxing, driving normally, or driving with an additional task (mental arithmetic). Recordings are video and physiological (ECG, GSR, skin temperature, breathing, EMG and BVP).

Emotional content: stress-related states rather than emotion per se.

Technical info & availability: Recordings are accompanied by self ratings and measures of reaction time. There are 24 participants (a total of 15 hours). Availability by arrangement with Erlangen team.

3.2 Procedures for Selecting the Clips Used in the HUMAINE Database

Selection involves non-trivial issues. Two levels were used.

The first was the selection of sections from within a whole recording. A selected section is referred to as a 'clip'. The basic criterion used to set the boundaries of clips is that 'the emotional ratings based on the clip alone should be as good as ratings based on the maximum recording available' (i.e. editing should not exclude information that is relevant to identifying the state involved). In the case of relatively intense emotional episodes the extraction of a section/clip includes build up to and movement away from an emotional nucleus/explosion – lead in and coda are part of identification of the state.

The second stage of selection was deciding which clips should form the HUMAINE database proper. It is very easy to drift into using a single type of material which conceals how diverse emotion actually is. To counter that, the 50 clips were deliberately selected to cover material showing emotion in action and interaction; in different contexts (static, dynamic, indoor, outdoor, monologue and dialogue); spanning a broad emotional space (positive and negative, active and passive) and all the major major types of combination (consistent emotion, co-existent emotion, emotional transition over time); with a range of intensities; showing cues from gesture, face, voice, movement, action, and words. and representing different genders and cultures.

4 Labelling

The labelling scheme emerged from sustained interaction with both theory and data. Early attempts (at QUB and LIMSI) dealt with the particular material for which they were developed, but failed to bring out issues that were salient in other sets, or to address theoretical issues that other HUMAINE partners considered important.

The resulting scheme provides ways of describing emotional content, the signs that convey it and relevant contextual factors, which can be applied to very diverse material, from induced to naturalistic emotion in action and interaction. It offers a range of labels, from relatively basic (and widely applicable) to more specialised (and probably application-specific). The emotion labels have been validated both theoretically and empirically (e.g. by measuring inter-labeller reliability). The labels for signs address multiple modalities – speech, language, gesture, face and physiological aspects (though some are only applied to a limited subset of clips). Information about many of the signs can be recovered automatically. Labels describing both signs and emotional content are designed to be time aligned rather than simply applied globally, since timing appears to be crucial in many areas.

The aim was to produce a system capable of dealing with most of the issues that an applied project might reasonably be expected to address. A scheme designed for a particular application will probably deal with a selected subset, but the HUMAINE scheme embodies a reasonable summary of the options that should be considered.

Specifications of database features are attached to the pilot database, which is available at http://emotion-research.net/deliverables/d5f-pilot-exemplar-database. What follows here picks out essentials and gives background information.

4.1 Development of Appropriate Emotion Descriptors and Labels

The emotion labelling scheme was developed mainly by QUB and LIMSI-CNRS. After much trial and error, two levels of description were included.

At the first level, global labels are applied to an emotion episode as a whole. Factors that do not vary rapidly (the person concerned, the context) are described here. It allows selection during the test process (e.g. identifying the emotion categories that are relevant). In the long term, it provides an index that can be used to identify clips that a particular user might want to consider. For instance, it will allow a user to find examples of the way anger is expressed in relatively formal interactions (which will not be the same as the way it is expressed on the football terraces).

Labelling at the second level is time-aligned. This is done using 'trace' type programs [8]. Unlike FEELtrace, from which they are derived, each of the current programs deals with a single aspect of emotion (e.g. its valence, its intensity, its genuineness). An observer traces his/her impression of that aspect continuously on a one dimensional scale while he or she watches the clip being rated. The data from these programs is imported into ANVIL as a series of continuous time-aligned traces.

The function of trace type labelling is to capture perceived flow of emotion. Focusing on that gives a rich picture in a reasonable time (roughly the duration of the clip times the number of traces). Different techniques are needed if it is critical to have fine timing or to know true rather than perceived emotion. They are usually much more time-consuming, and testing their validity raises difficult conceptual issues.

Emotion global descriptors
These cover eight main topics. Some simply summarise information associated with traces (e.g. range of intensities encountered, level of acting or masking encountered). Others are the outcome of long efforts to identify a minimal body of information that is needed to make sense of a clip that shows naturalistic material. That includes, for instance, understanding to what contexts the patterns observed might be generalized, and to which they should not. They are outlined here.

Emotion-related states: Full blown episodes of emotion make up very little of the emotion observed in naturalistic data. Descriptions of the types of state that do occur have been developed, taking Scherer's grid of design features [26] as a starting point. The categories included are Established (long term) emotion; Emergent emotion (full-blown); Emergent emotion (suppressed); Partial emotion; Mood; Stance towards person; Stance towards object/situation; Interpersonal bonds; Altered state of arousal; Altered state of control; Altered state of seriousness; Emotionless. Definitions are given in the guidelines for using the database on the HUMAINE portal.

Combination types: Again, 'pure' single emotions not the norm in naturalistic data [9], and so labels have been developed to describe the main types of combination that occur: unmixed, simultaneous combination (distinct emotions present at the same time), sequential combination (the person moves through a sequence of related emotions).

Context labels: Two broad types of context label are used. The first provides fairly factual data on the subject's personal characteristics; on technical aspects of recording; and on physical setting (degree of physical restriction, posture constriction, hand constriction and position of audience). The second deals with communicative context, including the purpose or goal of the communication (to persuade, to create rapport, to destroy rapport, or just pure expression of emotion); the social setting of the clip (for example, balanced interaction between two or more people, monologue directed to a passive listener); and social pressure (whether the person being rated is under pressure to be formal, as in a court, or to be freely expressive, as at a party).

Key Events: The labels in this set describe the key events with which the emotion is associated.. They identify both the focal key events that the person's feelings are about and other key types of event that contribute to the person feeling as they do. These include 'triggers' - events that prompt emotion about or towards another person/thing (e.g. OHP malfunction); 'causes' - long term influences on the person's state of mind (e.g crushing workload); and 'aspirations' – long term goal that shapes the way the person reacts (e.g retirement).

The global level also deals with two more standard types of descriptor.

Everyday emotion words: The sheer number of emotion words in everyday language [28] makes it a priority to find effective methods of selection. The HUMAINE Database selects at two levels. An outer set of 48 words is prespecified on the basis of studies [7], [4] that suggest they are important for labeling naturalistic material. From that set, individual raters then select those that are most relevant to the clip in question. That is a preliminary to tracing selected words to show the time-courses of the states that they describe.

Appraisal categories: Appraisal theory identifies emotional states with ways of evaluating significant events or people around. It provides an elegant framework, and pilot work tried to develop trace techniques based on it. Reliability was problematic for many of the categories [9]; not least because in naturalistic data, multiple aspects of the situation are likely to be appraised concurrently. However, appraisal-based descriptions are retained at global level, and those where reliability seems acceptable are also traced.

Emotion over time descriptors. Eight one-dimensional trace descriptors are used in the database. Many others have been explored, but were judged to be of less general interest or found not to give reliable ratings [9]. A key goal for analysis of the database is to establish whether the number can be reduced without losing information. The programs are as follows.

IntensTrace: This program deals with the intensity of the emotion raters believe the specified person is experiencing. Raters move a cursor on a scale that is displayed beside the clip as they watch it. Here and in the other programs, the layout of the scale is based on preliminary experiments. Definitions of the end points are displayed beside them ("zero emotion – totally emotionless" and "emotion at maximum intensity"). Intermediate descriptions are also displayed: 'distinctly unemotional', 'mild social emotion', and 'emotion in the full sense'. The markers are placed where average observers believe they naturally belong: the point of these is to minimize idiosyncratic departure from the average.

ActTrace: This deals with the extent to which the specified person is trying to give an impression of emotion that he/she does not actually feel. (i.e., there is an element of pretence or acting). The range is from "no attempt to simulate unfelt emotions" to "extreme attempt to simulate unfelt emotions".

MaskTrace: This is a converse of ActTrace: raters judge whether the specified person is trying to avoid showing emotions that they actually do feel. The idea that raters can judge how much emotion is being concealed can sound slightly paradoxical. In fact is is possible, because some signs of underlying emotion 'leak', and the effort of masking gives rise to its own signals (e.g. rigid posture).

ActivTrace: This deals with a quality that is sometimes called activation, sometimes arousal. It is basically how strongly the relevant person is inclined to take action, and it corresponds to a subjective sense of energy.

ValenceTrace: This asks tracers how positive or negative the specified person feels about the events or people at the focus of his or her emotional state.

PowerTrace: This asks tracers to rate how powerful the specified person feels, on a scale from 'absolutely no control over events' to 'completely in control of events'.

ExpectTrace: Tracers rate the extent to which the specified person has been taken unawares by the events at the focus of their emotional state. The range runs from "anticipated the events completely" to "taken completely unawares by the events".

WordTrace: In this case, raters choose an emotion word from the list of 48 (see above), and trace how the intensity of the chosen emotion varies through the clip. If

the chosen word is ‘fear’, for instance, the end points of the scale are “absolutely no fear” and “pure uncontrolled fear” , and there are intermediate markers for “slight fear” and “strong fear”.

The screen shot that follows conveys the net effect of putting traces together. The clip shows a participant feeling in a box, and suddenly triggering a buzzer. She gives a gasp, then a linguistic exclamation. The top trace, intensity, rises abruptly after the gasp. The rater does not judge that the response is acted, but there is a degree of masking at the beginning which breaks down abruptly at the unexpected event. Activation rises abruptly after a delay (during which the participant might be described as frozen).

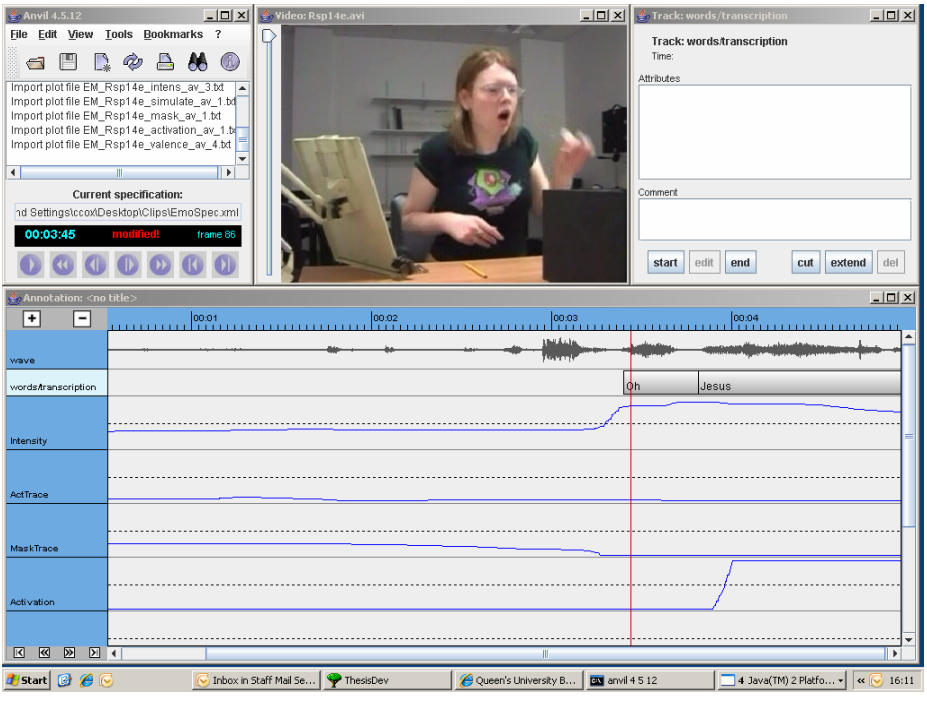

Fig. 1. Selected traces for a clip showing reactions to a sudden, surprising event

4.2 Labelling Signs of Emotion

Work on this part of the database has been highly interactive. The signal processing strand of HUMAINE has explored automatic extraction of parameters describing face and gesture, and identified the acoustic parameters that are most significant for recognition; and the teams building ECAs have helped to identify the parameters that needed to synthesise perceptually convincing gestures. In that sense, database research has a natural integrative function.

Speech and language descriptors: Three types of descriptor are used:

Transliteration. The core label here is the words spoken (see Figure 1). The data used in HUMAINE is largely interactive so both the words of the person observed and the words of any other interactants are transcribed. These are clearly differentiated. The words are time-aligned. Overlap is indicated. Future development will include the marking of pauses.

Largely automatically derived labels: The time waveform is displayed (see figure 1). It is first edited to suppress extraneous noises and the voices of other speakers (including sections of overlap). The edited .wav file is then used to derive a pitch contour for the person observed. This is derived using standard phonetic software, but hand edited to remove octave jumps and other errors (these are a substantial issue in emotional speech).

Work in the signal processing strand of HUMAINE will inform selection of a series of other automatically-derived labels, to be included in later versions of the database.

Auditory-based labels: These were developed by Douglas-Cowie on the basis of the Belfast Naturalistic Database [11] and then tested by another expert phonetician in further pilot work on a selection of clips from across the HUMAINE datasets. The original system contained many labels but for use in the HUMAINE database, these have been reduced to a core set of items which the tests indicate strongly characteristic of emotion and can be applied reliably. The labels address four descriptive categories and raters can assign a number of labels within these levels - Paralanguage (laughter, sobbing, break in voice, tremulous voice, gasp, sigh, exhalation, scream); Voice Quality (creak, whisper, breathy, tension, laxness); Timing (disruptive pausing, too long pauses, too frequent pauses, short pause + juncture, slow rate); Volume (raised volume, too soft, excessive stressing).

Gesture descriptors. The gesture coding scheme was developed at LIMSI-CNRS. It is a manual annotation scheme for coding expressive gestures in emotional corpora [21]. The following dimensions have been defined in the scheme:

- classical dimensions of gesture annotation ([23] [19] [18] [24]), to allow exploratory study of the impact of emotion on these dimensions:
- gesture units (e.g. to study how much gesture there is in an emotional corpus)
- phases (e.g. to study if the duration of these phases is emotion dependent)
- phrases / categories (e.g. to study the frequency of adaptators and compare it to other less emotionally rich corpora)
- lemmas adapted from a gesture lexicon [18] (e.g. Doubt=Shrug)

Face descriptors. Research in other parts of HUMAINE has developed procedures for extracting FAPs in a realistic timescale. They generate 17 FAPs which are suitable for automatic extraction, and which support reasonable levels of recognition. Fuller descriptions are given in [14]. The integration of this kind of data into Anvil has been carried out by the CNRS-LIMSI team using samples from EmTabou. Applying image processing techniques on video data like EmoTabou requires thorough comparison of manual annotation and image processing results. Some parts of the video cannot be processed because of factors like to hand-head occlusion or head movement and rotation, and even when there are no such obvious problems, mistakes can still arise.

Nevertheless, it is important that the database should acknowledge the role that automatic extraction can be expected to play in this area as techniques improve.

Physiological descriptors. Work on physiological descriptors is ongoing and will be expanded in later versions of the database. The physiological descriptors which will be incorporated in the database are those that apply to the physiological data from the driving experiments. Four basic channels are recorded: ECG (from which heart rate is derived); skin conductance; respiration; and skin temperature. There are many standard ways of deriving measures from these basic signals, involving differences, standard deviations, filtering, and various other operations. The approach in the database is generally to store the basic signals and leave users to derive other measures as and when they want to. Code that can be used to carry out standard transformations can be accessed via the portal from the Augsburg Biosignal Toolbox (AuBT), which was developed within the HUMAINE signal processing strand.

5 Conclusion

The HUMAINE Database workpackage set out to achieve a coherent set of responses to multiple challenges, involving both collection and annotation of diverse types of emotional material. This effort has included definition and testing of several coding schemes, and has influenced work in other areas of research (such as a W3C Incubator group on an Emotion Mark-Up language: Schröder et al., this conference).

It is a curious feature of the domain that the words commonly used to talk about it – affect, emotion, mood, anger, and so on – pull attention towards idealized landmarks and away from the everyday mixed cases between them. The obvious remedy is a collection of material that demonstrates by example what the everyday mixed cases look and sound like. The HUMAINE database makes a serious attempt to move in that direction. It is to be hoped that the effort will gather pace.

Acknowledgement. This research was supported by the EC FP6 Network of Excellence HUMAINE.

References

1. Abrilian, S., Devillers, L., Buisine, S., Martin, J.-C.: EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. 11th Int. Conf. Human-Computer Interaction (HCII'2005), Las Vegas, USA. Electronic proceedings, LEA (2005)
2. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proceedings ICSLP, Denver, Colorado (2002)
3. Batliner, A., Hacker, C., Steidl, S., Noth, E., Haas, J.: From emotion to interaction: Lessons learned from real human–machine dialogues. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 1–12. Springer, Heidelberg (2004)
4. Bänziger, T., Tran, V., Scherer, K.R.: The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions. Bari, Italy (2005)

5. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find trouble in communication. Speech Communication 40, 117–143 (2003)
6. Campbell, N.: Recording and storing of speech data. In: Proceedings LREC (2002)
7. Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., Fellenz, W.: What a neural net needs to know about emotion words. In: Mastorakis, N. (ed.) Computational Intelligence and Applications. World Scientific Engineering Society, pp. 109–114 (1999)
8. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: 'Feeltrace': an instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 19–24 (2000)
9. Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S., McRorie, M.: Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (2006)
10. Douglas-Cowie, E., Campbell, N., Cowie, R.P.: Emotional speech: Towards a new generation of databases. Speech Communication 40(1–2), 33–60 (2003)
11. Douglas-Cowie, E., et al.: The description of naturally occurring emotional speech. In: Proceedings of 15th International Congress of Phonetic Sciences, Barcelona (2003)
12. France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering 47(7) (2000)
13. Greasley, P., Sherrard, C., Waterman, M.: Emotion in language and speech: Methodological issues in naturalistic approaches. Language and Speech 43, 355–375 (2000)
14. Ioannou, S V., Raouzaiou, A T., Tzouvaras, V A., Mailis, T P., Karpouzis, K C., Kollias, S D: Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Networks 18, 423–435 (2005)
15. Juslin, P., Laukka, P.: Communication of emotions in vocal expression and music performance. Psychological Bulletin 129(5), 770–814 (2002)
16. Kienast, M., Sendlmeier, W.F.: Acoustical analysis of spectral and temporal changes in emotional speech. In: Cowie, R., Douglas, E., Schroeder, M. (eds.) Speech and emotion: Proc ISCA workshop. Newcastle, Co. Down, pp. 92–97 (September 2000)
17. Kipp, M,: Anvil - A Generic Annotation Tool for Multimodal Dialogue. 7th European Conference on Speech Communication and Technology (Eurospeech'2001), Aalborg, Danemark (2001), http://www.dfki.uni-sb.de/~kipp/research/index.html
18. Kipp, M.: Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Florida, Boca Raton (2004), http://www.dfki.de/~kipp/dissertation.html
19. Kita, S., van Gijn, I., van der Hulst, H.: Movement phases in signs and co-speech gestures, and their transcription by human coders. In: Wachsmuth, I., Fröhlich, M. (eds.) Gesture and Sign Language in Human-Computer Interaction. LNCS (LNAI), vol. 1371, Springer, Heidelberg (1998)
20. Leinonen, L., Hiltunen, T.: Expression of emotional-motivational connotations with a one-word utterance. Journ Acoustical Society of America 102(3), 1853–1863 (1997)
21. Martin, J.-C., Abrilian, S., Devillers, L.: Annotating Multimodal Behaviors Occurring during Non Basic Emotions. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, Springer, Heidelberg (2005), http://www.affectivecomputing.org/2005
22. McMahon, E., Cowie, R., Kasderidis, S., Taylor, J., Kollias, S.: What chance that a DC could recognise hazardous mental states from sensor outputs? In: Proc, DC Tales conference, Sanotrini (June 2003)
23. McNeill, D.: Hand and mind - what gestures reveal about thoughts. University of Chicago Press, IL (1992)

24. McNeill, D.: Gesture and Thought. The University of Chicago Press, Chicago (2005)
25. Sander, D., Grandjean, D., Scherer, K.: A systems approach to appraisal mechanisms in emotion. Neural Networks 18, 317–352 (2005)
26. Scherer, K R, et al.: Preliminary plans for exemplars: Theory HUMAINE deliverable D3c (2004), http://emotion-research.net/deliverables/D3c.pdf
27. Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I.: What should a generic emotion markup language be able to represent? In: Paiva, A., Prada, R., Picard, R.W (eds.) ACII 2007. LNCS, vol. 4738, pp. 440–451. Springer, Heidelberg (2007)
28. Whissell, C.: The dictionary of affect in language. In: Plutchnik, R. (ed.) Emotion: Theory and research, pp. 113–131. Harcourt Brace, New York (1989)
29. Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of emotions in interactive voice response systems Proceedings of the Eurospeech, Geneva (2003)
30. Zara, A., Maffiolo, V., Martin, J.C., Devillers, L.: (submitted). Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics. ACII (2007)
31. Zara, A.: Modélisation des Interactions Multimodales Emotionnelles entre Utilisateurs et Agents Animés.Rapport de stage de Master. Ecole Doctorale Paris XI. LIMSI-CNRS (8 September, 2006)