Piecing Together the Emotion Jigsaw

Roddy Cowie[1] and Marc Schröder[2]

[1] Psychology, Queen's University,
Belfast BT7 1NN, Northern Ireland
r.cowie@qub.ac.uk
http://www.psych.qub.ac.uk/
[2] DFKI GmbH,
Language Technology Lab,
Stuhlsatzenhausweg 3,
D-66123 Saarbrücken, Germany
schroed@dfki.de
http://www.dfki.de/lt/index.html

Abstract. People are emotional, and machines are not. That constrains their communication, and defines a key challenge for the information sciences. Different groups have addressed it from different angles, trying to develop methods of detecting emotion, agents that convey emotion, systems that predict behaviour in emotional circumstances, and so on. Progress has been limited. The new network of excellence HUMAINE explores the idea that progress depends on addressing the problem as a whole, not in isolated fragments.

1 Introduction

There is a growing sense that Information Technology should be addressing various issues linked to human emotion [1]. The EC has backed that view by funding a Network of Excellence called HUMAINE, which began in January 2004 and runs until December 2007. HUMAINE is charged with laying sound foundations for emotion-sensitive computing in Europe. It is governed by a 'technical annex' agreed with the EC, and the academic sections of it are available on the HUMAINE portal [2].

One of the ideas behind HUMAINE is that the area needs to develop strategies. It is often fair to take a purely tactical approach, based on seeing opportunities to make a small advance and seizing them. Experience suggests that that approach may not work well in the area of emotion-sensitive computing. One of the reasons is the sheer range of issues that need to be handled in a co-ordinated way in order to reach first base. Another may be that people come to the field with strong preconceptions about emotion, which direct their attention away from issues that are actually vital to progress. Part of the role of strategy is to offer protection against the pull of ideas that are deeply ingrained, but misleading.

The aim of this paper is to sketch an overview of the strategic issues. It draws on the framework of the HUMAINE proposal and discussions within the project, but it is not a statement of HUMAINE's position. In effect, it is an input to the debate, within HUMAINE and beyond, that will be needed to define a strategy capable of commanding wide acceptance.

S. Bengio and H. Bourlard (Eds.): MLMI 2004, LNCS 3361, pp. 305 – 317, 2005.

2 Clarifying the Subject Matter

One of the strategic challenges to HUMAINE is to provide clear ways of understanding the domain to be addressed, and talking about it. The issue arises because the word ‘emotion’ has many senses, corresponding to a variety of concepts that people intuitively use in thinking about the domain of emotion-related phenomena [3],[4],[5]. In that kind of area, relying uncritically on everyday terms and concepts is an invitation to confusion and/or misdirection.

The network adopted a general position which is stated early in the HUMAINE Technical Annex [2]:

> ‘We [will] consider emotion in an inclusive sense rather than in the narrow sense of episodes where a strong rush of feeling briefly dominates a person’s awareness. … emotion in the broad sense pervades human communication and cognition. Human beings have positive or negative feelings about most things, people, events and symbols. These feelings strongly influence the way they attend, behave, plan, learn and select.’

It is worth stressing the distinction because there are pressures that encourage research to focus on emotion in the narrow sense; and yet in the long term, it is the broad sense that is more likely to matter for technology.

Part of the response is to establish a vocabulary that makes the key distinctions easy to keep in view. Various options have been discussed in HUMAINE meetings, and the process is ongoing. Among the obvious options at this stage are terms derived directly from the passage cited above – ‘episodic emotion’ and ‘pervasive emotion’.

‘Episodic emotions’ seems an appropriate way to describe states where emotion dominates people’s awareness, usually for a relatively short time, by affecting their perceptions, feelings, and inclinations. The emotion may not determine the action that the person takes, but it requires effort to prevent it from doing so. Clear-cut episodic emotional states are not only relatively brief, but also relatively rare – at least among average people under average circumstances.

In contrast, ‘pervasive emotion’ refers to something that is an integral part of most mental states, if not all, including states where rationality seems subjectively to be firmly in control. It is integral because of the nature of a human being’s subjective world: emotional overtones are part and parcel of the way people experience not only the total situation that they are in (or are imagining), but also individual agents and things in it, and the courses of action that they take or contemplate. These overtones colour experience, and incline people towards some courses of action more than others, but they are part of the background unless something triggers a kind of phase shift and propels people into a state of episodic emotion.

Practically, there would be reasonably obvious advantages if machines could register the pervasive emotion that is involved when users feel positive or negative about people or situations that they are facing or talking about, alienated or engaged by the way a person or a system is communicating, and so on; and could generate appropriately coloured responses. It is less clear how much call there would be for information technology to deal with episodic bursts of emotion – they are quite rare, and they are a sensitive area that humans might well want machines to stay out of.

In fact, one of the obvious functions of sensitivity to pervasive emotion would be to make sure that systems did not propel users into fullblown emotional episodes.

These points bear emphasis because it is easy to drift into assuming that episodic emotion has a natural priority. Some of the reasons are grounded in everyday ways of speaking and thinking. The plural form, ‘emotions’, almost always refers to episodic emotions; and the archetypal emotion words (fear, anger, happiness, etc) apply primarily to episodic emotions. Psychological language tends to reinforce that outlook. For instance, Scherer has proposed a definition of emotions which begins by describing them as ‘episodes of massive, synchronized recruitment of mental and somatic resources ..’ [6]. The description draws on good scientific evidence to sum up a particular kind of emotional phenomenon. However, it is another matter entirely whether people who are interested in what everyday language calls ‘emotion’ should automatically concentrate on that kind of phenomenon.

A different kind of pressure comes from the idea that being emotional consists of flicking in and out of states that are at least close to the archetypes of episodic emotion - anger, happiness, and so on. Well-known views of emotion suggest that nothing else is possible, because ‘all emotions are basic emotions’ [7]. People who have worked with naturalistic data have repeatedly found that it does not take that form [5], [8], [9]. It is a measure of the strength of people’s preconceptions that they often respond by dismissing the data – if the data do not conform to their expectations, they conclude that the data have been poorly collected. That makes for a curious kind of empirical science.

A third kind of pressure may be inferred if not proven. Pervasive emotion, like other fundamentals of mental life such as consciousness, is frustratingly hard to pin down. It is much easier to describe episodic emotions, which contrast with other states and have a discernable structure. On the other hand, the very elusiveness of pervasive emotion may suggest that understanding it is the deeper problem.

Issues like these need to be addressed systematically. HUMAINE does so through two workpackages.

One deals with the linked tasks of conceptualizing emotion in the broad sense, and establishing terms that describe it without covertly misdirecting us. A taxonomy of ‘emotion related states’ is in progress: it gives recognition to multiple types of state other than the traditional archetypes, including

- Moods (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant)
- Interpersonal stances (e.g., distant, cold, warm, supportive, contemptuous)
- Attitudes (e.g., liking, loving, hating, valuing, desiring)
- Affect dispositions (e.g., nervous, anxious, reckless, morose, hostile) [6]

The second strand involves the collection of data that shows how emotion features in everyday settings. More detail is given in section 5 below.

Definition, taxonomy, and the collection of specimens are not glamorous activities in a technological field. Nevertheless, to ignore them is to invite chaos or superficiality. HUMAINE tries to avoid both.

3 Foreseeing Application Areas

One of the keys to strategy is a structured view of potential applications. This section tries to provide that. It distinguishes two broad groups of three, corresponding to relatively short and relatively long term types of goal. The cutoff is associated with a watershed, which is both the single most important application and a key to other possibilities.

3.1 Trouble Shooting

Probably the most active area at present is detecting troublesome emotions - in callers using automatic exchanges, pilots, drivers, etc. [10],[11],[12],[13]. Detection of stress [14] and lying [15] are closely related. Many extensions could be imagined, e.g. detecting staff who fail to display sufficient warmth towards customers. Trouble-spotting applications as an obvious starting point, partly because it looks as if rather simple systems could achieve useful results, and partly because it fits long-established stereotypes of what emotions are (trouble!). However, the area is fraught with problems, some ethical, some to do with doubts about the level of performance that simple systems can actually achieve.

3.2 Affective Selection

Affective selection involves detecting emotion-related responses to various stimuli, and using the responses to make choices that reflect the user's preferences. For instance, our group [16] proposed to use signs of emotion to locate the types of holiday that a user might respond well to, and Aizawa [17] proposed to archive video footage of times when brain signals indicate strong emotions. There are many natural extensions, such as adjusting ambient music or colouring to suit a user's mood.

3.3 Affective Loops

Höök [18] coined this term for a type of technology that goes back at least as far as the drum. The user acts; the system reacts; the result affects the user emotionally, and encourages the user to restart the cycle. Computing offers the prospect of designing a much wider range of systems that behave in that way, most obviously for enjoyment, relaxation, or therapy (the lines are blurred). The systems need not necessarily know any more about the user's emotions than a drum does, but it opens up new possibilities if they do.

3.4 Really Natural Language Processing

This is the watershed application mentioned above. Computers would become vastly more accessible if people could use really natural language to communicate with them – that is, talk to them as they would talk to another person, and have responses of the kind another person would give. Emotion is one of the areas where progress is needed to achieve that. Emotional colouring is an integral part of person-to-person exchanges, and it is clear that people react badly to discourse that follows other rules of conversation but ignores the emotional ones. There is no generally accepted description of the emotion-related rules that people expect an interlocutor to follow, but it is easy to suggest candidates, such as

match the general emotional tone of the other speaker (emotional convergence)
pick up topics that interest the other speaker
avoid topics or styles of speech that cause the other distress or boredom
repair ill feeling

Intuitively, it seems unlikely that speech interfaces incapable of observing rules like these will be widely accepted – and hence, achieving a level of emotion-sensitivity that allows them to be observed is integral to achieving really natural language communication.

Speech interfaces that allow really natural language processing are both the single most important application of emotion sensitivity, and a prerequisite for the longer term applications.

3.5 Uncovering Feelings

This term refers to uncovering the systems of values and dispositions that surround some kind of person or object or event in a user's mind. That is very different from attaching a label to a brief emotional state, and to do it well needs sophistication about language as well as about emotion. It is an essential key to a range of services – market or political research, careers advice, politics, non-directive counselling, personalised entertainment, etc..

3.6 Facilitating Learning

This is what good teachers do – not just presenting information to a learner, but taking account of the emotional issues – boredom, excitement, pride, humiliation – that make learning likely to succeed or fail. It probably depends on a fair degree of ability to elicit feelings. It is not confined to the classroom – it applies to manuals as much as multiplication tables.

3.7 Moulding

The term is meant to convey that the system sets out to change users' outlook or values or priorities rather than simply to extend their knowledge. It is separated from facilitating learning mainly because the two are very different ethically: moulding has obvious attractions for sales or politics, but it raises major ethical concerns. Emotionally sophisticated persuaders with no conscience, but infinite patience, are a nightmare.

Two implications of this overview should be noted. First, there are rather few areas that emotion oriented technology might see as its own particular preserve (specifically the first two). More often, techniques that are specific to emotion will enhance systems with many other elements. Second, the applications that most obviously involve episodic emotions are the short term ones: emotional colouring is the key issue in the longer term.

Generally speaking, the point of introducing emotion is to rehumanise functions that for various reasons we might want machines to carry out. The level of emotionality needed to do that is not dramatic, and it probably does not need to be present all the time. Nevertheless, it seems a safe bet that systems which are able to use relevant emotional colouring will obliterate systems that are not.

4 Ethics

Ethics has already been mentioned. Whether we like it or not, it simply is a subject that arises – most obviously because people are intensely sensitive about things that touch their emotions. As a result, if research gives a hint of irresponsible invasion or manipulation, it is likely to run into serious difficulties. The risk can probably be kept low if people working in the area take the trouble to be clear and moderately sophisticated in the area of ethics – much better that than to have horror stories emerge, and severe restrictions imposed by lawyers, protestors, or both.

5 The Empirical Base: Samples of Emotional Behaviour

Samples of emotional behaviour, spontaneous or simulated, are the empirical base of the area. They provide the raw material for developing sophisticated rules, either scientifically or by automatic methods; and for templates in the various techniques that use them. There are areas where intuition and informal experience are sufficient to produce first order approximations, such as the design of basic conversational agents [19]; but long term applications clearly need a stronger empirical base. Assembling an appropriate set of samples is a major challenge.

Traditionally, the area has relied heavily on acted samples. There is growing unease with that approach [8]. Actors reproduce people's stereotypes, not the behaviour patterns that ordinary people exhibit in everyday life. One marker of the difference is that training on acted speech does not lead to good performance on spontaneous speech [10]. Another marker is the time it takes to tell whether speech is acted or spontaneous. In a recent study we found statistically significant discrimination within the shortest interval we tested, 3 seconds.

Call centre data is attracting attention as an alternative, but it has its own problems. Its main relevance is to episodic emotions that signal trouble. For that purpose, the data rate is very low: Ang studied a sample of 13,187 utterances, and found 42 that clearly qualified as irritated [11]. The data is almost always in a single modality, speech. It is also constrained in form and emotional range, making it all too likely that solutions will be tied to those features and lack generality. To illustrate the problem of emotional narrowness, Yacoub et al [20] developed a system that discriminated anger from neutral speech. However, when happy speech was introduced, it too was classified as angry – ie the discrimination produced by training on a narrow base was not actually anger/neutrality at all. To illustrate the problem of task constraints, we published evidence [21] that speakers who were deeply bored paused less (ie they ran words together). A follow-up study has since shown an exactly opposite effect. The reason is a minor change to the paradigm – the presentation created natural blocks in the first study but not the second. Experiments can expose that kind of effect: telecommunications companies do not have the same leeway.

Difficulties with other options focus attention on deliberate elicitation [5],[8]. That may involve inducing an emotional state in passive subjects or facilitating people who are actively trying to achieve it. At one extreme is subjecting people to films of surgery [22]; at the other is Picard's subject who developed mental routines for achieving specific, highly repeatable target states [23].

Developing elicitation techniques highlights the distinction between a state and the way it is expressed in a given context. The range of context effects is enormous. For instance:

- emotion driven by immediate surroundings probably differs from emotion that relates to situations that are remembered or anticipated [24];
- situations that elicit happiness do not necessarily lend themselves to establishing how it affects scripted speech, still less to establishing how it is expressed in spontaneous dialogue;
- signs may be heightened or suppressed in quite different ways according to social context [25].

Issues like these show why one of the strategic needs in the area is to stand back from the task of achieving specified emotional targets, and to define an appropriate set of targets. It is impossible to record the whole domain of emotion: research needs points of reference chosen so that given information about them, the rest of the domain can be reconstructed reasonably accurately. In the past it made sense to hope that so-called primary emotions would provide those points of reference. That carries over into a de facto tendency to concentrate on eliciting states which are at least very strongly coloured by emotion. But if the long term goal is to understand commonplace emotional colouring, those inherited tendencies need to be questioned and ideally replaced.

6 Modality and Recording

Emotion is profoundly multi-modal. It is reflected in facial expressions, gestures, body language, and actions; in the propositions expressed, the words and syntax chosen to express them, and the way they are spoken; in involuntary visceral changes, and in blood flow and electrical activity in the brain.

In an ideal world, researchers would be able to record all of the relevant modalities without compromising the essential nature of the situation. In reality, they have difficult choices to make. Introducing a camera constrains; attaching leads for psychophysiology constrains still more; brain scans can only be collected in environments that make natural expression of emotion all but impossible. That is all over and above the fact that emotion will be expressed through speech in some situations, gesture in others, and so on.

Some theoretical perspectives do suggest that certain modalities are privileged – they define 'ground truths', against which the validity of other measures has to be gauged. According to James [26], the essence of emotion was visceral response. According to Cannon [27], it was activity in specific brain centres. Recent work is more pluralist. It regards emotion as intrinsically multifaceted [28]. To attribute an emotional state is to summarise a range of objective variables. Hence, no one modality is indispensable. Equally important, there is no measure that defines unequivocally what a person's true emotional state is.

From that viewpoint, the issue is to weigh the costs and benefits of collecting particular types of measure at particular levels of resolution in a given situation. All parties need to understand how difficult the balances are. Nobody gains from studies that achieve the highest possible quality and resolution in multiple modalities at the cost of enforcing completely stilted renditions of emotion.

7 Identifying Emotional Content

One of the core tasks of research concerned with emotion is to associate emotion-related signs with labels that identify the associated emotion-related states. Databases need to present an association that can be regarded as valid; recognisers need to generate appropriate state descriptions given appropriate signs; synthesis needs to be capable of generating signs that are appropriate to a given state. Finding appropriate ways to identify emotion-related states is a substantial challenge in itself.

Central to the challenge are two groups of issues with wide-reaching implications. They involve the *form* of the description, and the *perspective* from which it is given.

The most familiar form of description is categorical. Emotions are identified by identifying verbal labels, which are either drawn directly from everyday language, or adapted from it. Typical adaptations involve distinguishing hot and cold anger, romantic and nurturant love, etc.

Episodic emotions may well have an inherently categorical structure, though note that there is little sign of convergence in attempts to define a satisfying set of categories [5]. Even if there were, it seems very unlikely that a categorical system could capture the shades of emotional colouring satisfactorily. It would involve a very large number of categories, and the categorisation would not reflect the fact that some categories are obviously very close and others very far apart.

At the opposite extreme are dimensional descriptions, which identify emotional states by associating them with points in a multidimensional space. The approach has a long history – Wundt [29] proposed it and Schlossberg [30] reintroduced in the modern era. Analyses agree that emotion concepts reflect two main dimensions, which we have called activation and evaluation [4] (though see [31] for a different approach); and a number of others which are less important, such as power or approach/avoidance.

A third option, pioneered by Ortony et al, [32], is a logical description, which identifies emotional states in relation to a series of alternatives – is the focus present or imagined? a person or a thing? and so on.

All of these options have been used in practice [8], [33],[34]. Results confirm the obvious expectations – categorical and logical descriptions raise difficult statistical problems when there is a substantial range of emotions to deal with, dimensional descriptions are more tractable but fail to make important distinctions.

The second group of issues related to identifying emotional content have been described as involving perspective. The question behind them is: whose view of an emotional episode is research concerned with?

The obvious assumption is that research should be concerned with an absolute perspective which reflects a person's state with as much scientific accuracy as possible. In some applications, that is quite reasonable: for instance, a lie detector should presumably establish whether a person is actually lying.

Equally, though, there are applications where the natural goal is to match the perspective of a representative observer. A system designed to hold relatively normal conversations does not need to penetrate deceptions that would pose problems for a person – in fact, it should be fallible in about the same way as a person would be. There are interesting questions about handling signs that different people perceive in noticeably different ways.

A third aim is to design systems that reach the same conclusions about a person's emotions as the person him- or herself. That would correspond to a kind of empathy. It is something that humans often find difficult, but it might, for instance, be important in the last three types of application outlined above.

The main point to be made here is that decisions on these issues – form of description and perspective – have a far-reaching effect on the shape of a research effort. Decisions with such strategic implications should not be made by default, on the grounds that some options come to mind more easily than others.

8 Signals to Emotion Labellings and Back

The obvious work for technologists is to construct processes that lead from an input signal to an emotion label; or from an emotion label to an output signal. The signals may be audio, visual, or physiological. The aim of this section is to illustrate the kind of challenge that arises in any of the streams. It uses speech as an example.

In speech, the signal is a fluctuating voltage. The natural image is that processing should work through a series of transformations to an identification of the emotional state in which the signal was produced – in fact two series, one dealing with the linguistic content of the signal ('what you say'), the other dealing with the paralinguistic ('how you say it'). Following through the transformations defines a useful framework.

The basics of linguistic processing are well known, and do not need to be reviewed here. However, it is worth noting two major challenges.

The first is simply to recognize words in spontaneous emotionally coloured speech. It has various properties that pose problems for standard recognition systems, including non-standard largyngeal behaviour [35], great variation in phoneme duration [36], reduced articulation, 'trailing off' rather than delivering sharp speech-silence boundaries, devoicing, intruding non-speech sounds (laughter, sniffs, sobs), or morphing speech into a cry ('noooo …') [11].

The second challenge is to move from words to assessments of a person's emotional state. There has been related work on text, but it is impossible to judge how it transfers because the necessary samples of spontaneous emotional speech are not available.

The paralinguistic stream is less familiar. An initial set of transformations create several more useful time series. The core time series define intensity, energy in certain frequency bands, and local voice pitch; and arguably points derived from LPC or cepstral transformations, perhaps in combination with the Teager Energy Operator. With the core series may be associated derivatives and measures of local steadiness. Some teams supply these time series directly to a recognition algorithm; most insert further transformations [37].

The natural next task is segmentation – defining significant markers and the segments that lie between them. Key markers include boundaries of pauses and phrase-like units (it is debatable how closely that two are related), local maxima and points of stress in the pitch contour, and arguably boundaries of phoneme types (particularly vowels and fricatives).

Given a segmentation, processing can extract descriptors of set segment properties. These include completely standard descriptive statistics (magnitude and duration, and if a segment contains several elements, the means, ranges, etc associated with them); but also increasingly properties we have described as configurational. These include 'crescendo' (buildup of intensity over a phrase), 'topline' (the trend of pitch peaks over a phrase), and parameters of the way a phrase begins and ends; pitch peak and and trend within a vowel; prosodic similarity or contrast between successive phrases; and others. These descriptors can be related directly to measures of emotionality: it is clear that many of them correlate strongly with measures such as activation evaluation, and simple departure from neutrality [37].

Some investigators argue that a further level is needed, in which linguistic and paralinguistic streams are recombined [38],[39]. Prosody may be redescribed in systems like ToBI that are oriented to describing its linguistic significance. That allows observed prosody to be compared with the default predicted by linguistic content, and key regions (where variation is emotionally significant) to be identified. It is clear that considering linguistic issues can augment emotion recognition, but attempts to replace direct descriptors with linguistic ones have been disappointing [34].

A key reason for outlining this kind of structure is to highlight component challenges that deserve attention in their own right. For instance, really natural language processing means that people will not be constrained to ensure 'good' signals. In that context, recovering the most basic contours – intensity and pitch – is difficult. Hence, it makes sense to invest effort in developing standard methods that will deliver them reliably. The same applies to segmentation, with a qualification – the more familiar needs of linguistic processing should not be allowed to dictate the standardisation that emerges. Given that basis, teams with linguistic and psychological skills could explore the higher order issues of relationships between recovered structures and emotional expression.

Thinking at higher order makes it easy to see that a dimension has been left out in the discussion so far. It is a fundamental feature of emotion that it extends and fluctuates in time. The fundamental task is therefore not to match a set of features onto a single label, but to map structured sequences of features onto a changing emotional profile. Conceptually, that task seems easier to address from the perspective of synthesis than in the context of analysis.

In terms of synthesis, linguistics offers a natural model. Generative phonology takes speech synthesis to the point of defining sequences of tokens to be realized: generative phonetics translates that sequence into the domain of signals. Contemporary work on markup languages provides a first approximation to the kind of string that a phonology-like component might generate. There are obvious reasons to aim at convergence between the symbols used in more developed versions of that component and the properties delivered by the paralinguistic analysis, and we have shown that emotional speech synthesis can be controlled by parameters of the kind that one paralinguistic analysis system identifies as correlates of emotion [33].

Speech has been used as an example to work through, but similar issues arise in facial and gestural modalities, and to some extent in the analysis of physiological signals. In each case, the first step is to recognize the sheer scale of the task, and to find rational ways of subdividing it. It makes no sense to require every group that works in any of the areas to attempt all the tasks involved in connecting signals to states – the result is almost bound to be a proliferation of systems on too small a scale to address the basic problems in a satisfying way.

It remains to be said that issues also arise specifically from the attempt to combine different modalities. For instance, it is not at all clear whether they integrate additively; whether some have priority for some decisions (as happens with audiovisual speech reception); whether attention can determine priority; over what interval of time information from transient signals is assumed to be relevant; and much more.

9 Modelling Emotional States

To this point, the description of emotion has been considered as a process of attaching labels – whether they consist of words or sets of co-ordinates. For some short term applications, that may be enough, but it is clearly not enough to support really natural language processing. That requires ability to register what it means to be in a particular emotion-related state, in the sense of being able to gauge how people perceive their current situation, including what their priorities might be; to anticipate what a person might do next, and how they might regard alternative conceivable responses from the system; and perhaps to intuit reasons for their current state. These requirements add up to forming an internal model of the user's emotional state.

Modern theory provides a rich source of ideas about emotion models. It suggests that emotion is rooted in representations with a distinctive structure – they are selective, evaluative, link features of the situation to potential actions, and are at least not wholly propositional. Ideas like these have led to several different strands of research – traditional AI emphasizing representational power; neural nets emphasizing the subsymbolic; artificial life emphasizing the link to action and survival-related evaluation [40].

Deep progress in these areas may take a long time, but it is not difficult to envisage conversation controllers developing gradually from simple rule sets about what to say when somebody is angry, towards more general and principled solutions. Even simple rule sets open the way to systems that could sustain affective loops, and provide a context in which it is possible to begin putting together the key pieces sketched here.

10 Conclusion

HUMAINE is based on explicit recognition that achieving emotion-sensitive computing requires major empirical, theoretical, and structural issues to be addressed in a co-ordinated way. The outline given here simplifies every topic that it deals with and ignores as many others, not because they are insignificant, but because selections have to be made. Nevertheless, it may convey the daunting scale and sheer excitement of the attempt to pull such a large structure into a viable shape.

Acknowledgement

Preparation of this paper was supported EC grant IST-2002-507422, and the content is indebted to discussions with too many HUMAINE members to list.

References

1. Picard, R. W. *Affective Computing*. MIT Press, Cambridge, MA (1997)
2. HUMAINE portal http://emotion-research.net/
3. Russell J. & Barrett-Feldman L Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *J. Pers & Soc Psychol* 76 (1999) 805-819,
4. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, & Taylor J. Emotion recognition in human-computer interaction. *IEEE Sig Proc Magaz* 18 (2001) 32-80
5. Cowie R & Cornelius R. Describing the Emotional States that are Expressed in Speech. *Speech Comm* 40 (2003) 5-32
6. Scherer, K et al HUMAINE Deliverable D3c: Preliminary plans for exemplars: theory http://emotion-research.net/
7. Ekman P. Basic emotions. In Dalgleish, T. and Power, M. J. eds., *Handbook of Cognition & Emotion*. John Wiley, New York (1999) 301–320.
8. Douglas-Cowie E, Campbell N, Cowie R, & Roach P Emotional Speech: towards a new generation of databases. *Speech Comm* 40 (2003) 33-60
9. Kwon O, Chan K, Hao J, & Lee T-W Emotion recognition by speech signals. *Proc. Eurospeech* (2003) 125-128
10. Batliner A, Fischer K, Huber R, Spilker J & Nöth E. How to find trouble in communication. *Speech Comm* 40 (2003) 117-143
11. Ang J, Dhillon R, Krupski A, Shriberg E, & Stolcke A Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. ICSLP*, Denver, Colorado (2002)
12. Hadfield P & Marks P This is your captain dozing … *New Scientist* 1682267, (2000) 21
13. McMahon E, Cowie R, Kasderidis S, Taylor J, & Kollias S What Chance that a DC Could Recognise Hazardous Mental States from Sensor Outputs? *Tales of the Disappearing Computer*, Santorini (2003)
14. Zhou G, Hansen JH, & Kaiser JF, Methods for stress classification: Nonlinear TEO and linear speech based features. *Proc. IEEE Int Conf on Acoustics, Speech,& Signal Processing*, vol. IV, (1999) 2087-2090
15. Haddad D, Ratley R , Walter S & Smith M *Investigation and Evaluation of Voice Stress Analysis Technology*. Final Report US Dept of Justice Report NCJ Number 193832 (2002)
16. ERMIS team *D03: System Architecture and Testbed Specifications* ERMIS project IST-2000-29319 (2002)
17. Aizawa K Position Statement *Proc VLBV01*, Athens (2001) 3
18. Höök K, Sengers P, & Andersson G Sense and Sensibility: Evaluation and Interactive Art *Computer Human Interaction*, Fort Lauderdale (2003)
19. Paiva A ed. *Affective Interactions: Towards a New Generation of Computer Interfaces.* Berlin: Springer-Verlag (2000)
20. Yacoub S, Simske S, Lin X, & Burns J, Recognition of emotions in interactive voice response systems. *Proc. Eurospeech*, Geneva (2003)
21. R Cowie, A McGuiggan, E McMahon, & E Douglas-Cowie Speech in the Process of Becoming Bored. *Proc. 15th ICPhS*, Barcelona (2003)
22. JJ Gross & RW Levenson. Emotion elicitation using films. *Cognition and Emotion* 9 (1995) 87-108

23. Picard RW, Vyzas E & Healey J Toward Machine Emotional Intelligence: Analysis of Affective Physiological State *IEEE Trans Patt Analysis & Machine Intell,* 23 (2001) 1175-1191
24. Stemmler G, Heldmann M, Pauls C, & Scherer T Constraints for emotion specificity in fear and anger: the context counts. *Psychophysiology* 69 (2001) 275-291
25. Parkinson B *Ideas and realities of emotion*. Routledge, New York (1995)
26. James W, What is emotion? *Mind* 9 (1884) 188-205
27. Cannon WB Against the James-Lange theory of emotion *Psych Rev* 38 (1931) 106-124
28. Cornelius R *The Science of Emotion*: *Research and tradition in the psychology of emotion*. Upper Saddle River: Prentice-Hall (1996)
29. Wundt W *Grundzuge der Physiologischen Psychologie* vol 2. Engelmann, Leipzig, 1903. (Original published 1874)
30. Schlosberg H, A scale for judgment of facial expressions. *Journal of Experimental Psychology* 29 (1954) 497-510
31. Watson D & Tellegen A, Toward a consensual structure of mood. *Psych Bull*, 98 (1985) 219-235
32. Ortony A, Clore G & Collins A, *The cognitive structure of emotions*. CUP, Cambridge, England (1988)
33. Schröder M Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, *PHONUS 7, Research Report of the Institute of Phonetics, Saarland University* (2004)
34. Stibbard R Vocal expression of emotions in non-laboratory speech. PhD thesis, University of Reading (2001)
35. Cummings K & Clements M Analysis of glottal excitation of emotionally styled and stressed speech. *JASA* 98 (1995) 88-98
36. Williams CE & Stevens KN Emotions and speech: Some acoustical correlates. *JASA*. 52 (1972) 1238-1250
37. Cowie R, Douglas-Cowie E, Cox C, & Cemegil A T *D09: Final Version Of Non-Verbal Speech Parameter Extraction Module* ERMIS project IST-2000-29319 (2004)
38. Ladd DR, Silverman K, Tolkmitt F, Bergmann G & Scherer K Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *JASA*. 78 (1985) 435-444
39. Mozziconacci S Speech variability & emotion: Production & perception. Ph. D. thesis, Technical University Eindhoven (1998)
40. Trappl R, Petta P and Payr S, eds., *Emotions in Humans and Artifacts*, Cambridge, MA: The MIT Press (2003)