

Perceiving emotion: towards a realistic understanding of the task

Journal:	<i>Philosophical Transactions B</i>
Manuscript ID:	RSTB-2009-0139
Article Type:	Research
Date Submitted by the Author:	30-Jun-2009
Complete List of Authors:	Cowie, Roddy; Queen's University Belfast, Psychology
Issue Code: Click http://rstb.royalsocietypublishing.org/site/misc/issue-codes.xhtml target=_new>here to find the code for your issue.:	COMPUTATION
Subject:	Behaviour < BIOLOGY, Cognition < BIOLOGY, Neuroscience < BIOLOGY
Keywords:	emotion, perception, expression, face, speech, naturalistic

Perceiving emotion: towards a realistic understanding of the task

Roddy Cowie
School of Psychology,
Queen's University Belfast
Belfast BT7 1NN
r.cowie@qub.ac.uk

ABSTRACT

A decade ago, perceiving emotion was generally equated with taking a sample (a still photograph or a few seconds of speech) that unquestionably signified an archetypal emotional state, and attaching the appropriate label. Computational research has shifted that paradigm in multiple ways. Concern with realism is key. Emotion generally colours ongoing action and interaction: describing that colouring is a different problem from categorising brief episodes of relatively pure emotion. Multiple challenges flow from that. Describing emotional colouring is a challenge in itself. One approach is to use everyday categories describing states that are partly emotional, partly cognitive. Another is to use dimensions. Both need ways to deal with gradual change over time and mixed emotion. Attaching target descriptions to a sample poses both problems of both procedure and validation. Cues are likely to be distributed both in time and across modalities, and key decisions may depend heavily on context. The usefulness of acted data is limited because it tends not to reproduce these features. By engaging with these challenging issues, research is not only achieving impressive results, but also offering a much deeper understanding of the problem.

Key phrases: perception of emotion, facial expression of emotion, speech and emotion, multimodality, naturalistic data

1. INTRODUCTION

A good deal is known about the perception of emotion, but is not generally presented as a distinct research topic. The relevant material is distributed across various disciplines, and the stated topic of the research is often the expression of emotion, rather than the perception of emotion. The research has also been directed towards various different practical goals, and it is not necessarily easy for people concerned with one goal to see the relevance of research concerned with another.

Recent developments help to bring the topic into sharper focus, and the aim of this paper is to digest their implications. Part of the task is to look back and reconsider practices and assumptions that were implicit in earlier approaches.

A useful starting point is to distinguish three styles of work. The oldest focused on intuitive descriptions that might allow people to recognize emotions better. The second aimed to meet more recognizably scientific standards in terms of signal processing and experimental techniques. The third aims to provide underpinnings for emotion-related technologies. For brevity, the three styles will be called impressionistic, experimental and technological.

1
2
3 The paper is particularly concerned to highlight the way technological research is
4 changing the field. Most fundamentally, by engaging with people's ordinary ability to
5 register other people's emotions, it reveals how extraordinary that ability is.
6
7

8 Reviewing the field in that way allows a wide range of research efforts to be given a
9 place. There are traditions that stay outside it, though. Broadly speaking, the research
10 that it includes is in the tradition of cognitive psychology, which regards perception
11 (of emotion among other things) as something that can be measured; analysed in ways
12 that have some generality; and modeled by artificial systems that attempt to match
13 human competence. Other traditions regard it as something that is irreducibly
14 subjective, able to be spoken about but not measured, and so intimately tied to
15 particular situations that no generalization is possible. So, for instance, Sengers et al
16 argue for "an enigmatics of affect, a critical technical practice that respects the rich
17 and undefinable complexities of human affective experience" (2002, p.87). It is
18 interesting to ask how links could be made with traditions like that, but beyond the
19 scope of this paper.
20
21
22
23

24 **2. RESEARCH IN THE IMPRESSIONISTIC STYLE**

25 Darwin's contemporaries launched a style of research that is still active. It focuses on
26 signs of emotion that people can detect once they are alerted to them, but may not
27 notice spontaneously. The descriptions of signs tend to be impressionistic in the sense
28 of the term that is used in phonetics: they draw attention to patterns that human
29 perceivers can recognize and identify consciously, given appropriate guidance.
30
31
32

33 Research in that impressionistic style often does not mention perception, but in effect,
34 it is bidirectional. It describes signs that people tend to give in various emotional and
35 emotion-related states, usually with the implication that they have the potential to be
36 used in perceiving emotion. Its natural application is training people to perceive
37 emotion more accurately – something that there are many reasons to want.
38
39

40 Research in that mould raises general issues that clearly should be part of a mature
41 science dealing with the perception of emotion. It implies that the perception of
42 emotion is an area where significant perceptual learning can take place. It also implies
43 that a particular form of perceptual learning occurs, where recognition is improved by
44 conscious identification of relevant signs. There is some evidence that these ideas are
45 at least sometimes true (e.g. Ekman & O'Sullivan, 1991; Lacava et al., 2007).
46
47

48 On a more detailed level, the research provides a rich resource of descriptive material,
49 from classics like Birdwhistell's 'kinesics' (1970) to contemporary developments
50 (e.g. Poggi, 2006).
51
52

53 The impressionistic tradition contains acute observations that should not be dismissed
54 simply because they are informal. On the other hand, its assumptions should not be
55 accepted uncritically. Refining people's ability to perceive emotion is not the only
56 motive for studying the subject; and there are delicate questions about the relationship
57 between patterns that we can recognise in their own right and the everyday business
58 of perception.
59
60

3. RESEARCH IN THE EXPERIMENTAL STYLE

More formal techniques entered the field gradually. They affected various issues: formalizing descriptions of stimuli and perceptual responses to them; control of presentation; studying observer characteristics; and so on. The combination produced research with a recognizably different emphasis.

Some of the earliest experimental work emerged from a controversy over the relationship between cognition and emotion (Lazarus, 1999). It indicated that perceptual processes can derive emotion-related information without creating a conscious impression of the stimulus. That seemed paradoxical initially, but it is now clear that perceptual processes do not necessarily involve conscious awareness (Milner & Goodale, 1995). In that context, similar effects in emotion perception are unsurprising. So-called mirroring of laughs, yawns, smiles, and so on occurs without deliberate intent (Hatfield et al., 1994), and facial musculature responds to stimuli that are not consciously perceived (Dimberg et al., 2000). That material reinforces doubts about the connection that the impressionistic style envisages between the ordinary business of perceiving emotion and conscious identification of potential cues.

Formal analyses of emotion-related stimuli were basic to the development. In the case of faces, the FACS scheme (Ekman & Friesen, 1976) became established early on as a 'gold standard'. Similar efforts in other modalities achieved less consensus. In the case of speech, variables such as intensity and pitch contour were used early on (Lieberman & Michaels, 1962), but there was long running debate over schemes concerned with their linguistic function (Mozziconacci, 1998). Voice quality was problematic: neither impressionistic schemes (Laver, 1980) nor spectrum-based measures (Hammarberg et al., 1980) proved fully satisfactory. Schemes for annotating movement were developed (e.g. Grammer et al., 1998), but not really standardised.

These analyses provided practical tools, but they also bear on questions about perceptually significant units. These were not generally explored in the way that related questions were elsewhere; for example, whether geons functioned as perceptual units (Biederman, 1987). A related question, whether Gestalt effects dominate the interpretation of individual features, was raised by Ekman (1982), but research is quite limited. In the speech domain, questions about the perceptual reality of descriptive schemes for voice quality (Bhuta et al., 2004) and intonation (Prom-on et al., 2009) have been actively pursued, but they have proved difficult to resolve.

A key way to test the perceptual relevance of analyses was by using them to synthesise stimuli capable of evoked a specified perceptual response. Ekman and Friesen's 'Pictures of Facial Affect' (1976) are effectively an early example, since the actors made them by generating expressions defined by FACS. The system gained credibility from its adoption into graphics technology in the early 1990s (Terzopoulos & Waters, 1993), and a modified version (FAPs) was incorporated into the MPEG 4 standard (Pandzic & Forscheimer, 2002). The early 1990s also saw speech synthesisers designed to convey well-defined emotion categories (Cahn, 1990; Murray & Arnott, 1995).

These syntheses confirm that the underlying analyses have some relevance to perception. However, they also expose a problem. The stimuli that they produce do

1
2
3 allow observers to distinguish the relevant categories; but on the other hand, they are
4 clearly not perceptually natural. The issue is taken up later.
5
6

7 Experimental research relied on these technical developments, but as Russell et al
8 (2003) observed, its characteristic direction came from the hypothesis that emotion
9 can be partitioned into discrete types, and that the partition is governed by evolution
10 rather than culture.
11

12 Standard instances of proposed categories were central to the exploration. The
13 archetypal example is the Ekman and Friesen (1976) collection of posed photographs.
14 These were known to be discriminable, and so they provided a basis for studying
15 mechanisms of discrimination. Research on emotion in speech followed a partly
16 similar pattern. Oster and Risberg (1986) recorded actors simulating six states: angry,
17 astonished, sad, afraid, happy and positive. Later studies analysed these to identify the
18 features that distinguished the recordings (Carlson et al, 1992). Similar databases
19 followed in other languages (Burkhardt & Sendlemeier, 2000). The outstanding work
20 in this mould (Banse and Scherer, 1996) used recordings of actors simulating 16 types
21 of emotional state, preselected to ensure that they were discriminable by human
22 beings. Instrumental analysis then identified speech variables associated with the
23 discriminations.
24
25
26
27

28 Alongside the work with static images of faces, there was a considerable body of
29 experimental research on the role of movement. Bassili (1978, 1979) demonstrated
30 that category judgements could be made on the basis of facial movements rather than
31 static configurations. Ekman and Friesen (1982) added the influential idea that timing
32 distinguished different types of smile. But although there was experimental support
33 (Frank, Ekman & Friesen, 1993), the overall pattern of findings tended to be equivocal
34 (Ambadar, Schooler, and Cohn 2005). It was demonstrated that various kinds of body
35 movement – including dance (Dittrich et al., 1996), knocking movement (Pollick et
36 al., 2001), and gait (Crane & Gross, 2007) – can be used to classify an agent's
37 emotional state.
38
39
40

41 The core stimuli provided a basis for constructing more complex material. The effect
42 of context was extensively studied. Early studies used pictures of real-life situations
43 (Munn, 1940) or film sequences (Goldberg, 1951), and reported strong context
44 effects. However, the paradigm that came to be most widely used combined posed
45 facial expressions showing extreme emotion with verbal descriptions of context; and
46 the task was essentially to judge whether the expression was to be believed. In that
47 paradigm, facial cues typically predominated (Fernandez-Dols et al., 1991).
48
49

50 A few teams also considered the effect of combining information from different
51 modalities, primarily by pairing a voice with a photograph of a face (de Gelder &
52 Vroomen, 2000). An influential explanation of the results (Massaro, 2004) proposed
53 that combination follows fuzzy logical rules, which are both rational and widely used
54 in perception.
55
56

57 Controlled pictures such as Ekman and Friesen's lend themselves to morphing, and
58 that was exploited to study category boundaries. Two notable types of finding
59 emerged. There is evidence of categorical perception, meaning that the perceptual
60 effect of objectively equal differences between stimuli are perceptually small if the

1
2
3 stimuli lie well within a category, large if they are close to a category boundary
4 (Young et al., 1997). There is also evidence that boundaries are labile; Niedenthal et
5 al (2000) showed that they shift with mood.
6
7

8 A more recent manipulation is the ‘bubble technique’, using stimuli where some
9 patches of an original stimulus are retained, others filtered out. It has been used, for
10 instance, to provide strong evidence for the hypothesis, intuitive but not easy to
11 confirm, that “the eyes and the mouth of faces are most useful to viewers in
12 discriminating the emotion” (Adolphs, 2006, p.224).
13
14

15 The bubble technique is linked to recurring efforts to establish the irreducible
16 minimum of information needed to achieve classification. The work on movement:
17 typically tried to show that it makes a distinct contribution by presenting stimuli, such
18 as point light displays, where static frames contain virtually no information. In the
19 context of speech, synthesis techniques allowed research to manipulate a single
20 variable at a time – pitch level, pitch rate, pitch contour, and speech rate have all been
21 shown to influence classification (Mozziconacci, 1998). It has also been shown that
22 very short extracts from human speech – as little as a single vowel – are sufficient to
23 allow some discriminations (Laukkanen et al., 1996).
24
25
26

27 What has been outlined above is a core of work broadly comparable to research in
28 other areas of perception. It has been interwoven with research on two other key
29 themes.
30
31

32 The first key theme is differences between cultures and individuals. Not many now
33 dispute that there are universals underpinning the perception of emotion (Schmidt &
34 Cohn, 2001). However, the process is clearly subject to very substantial variation.
35 Among the factors that affect recognition are mood, culture, gender, emotional
36 intelligence, and various disorders including schizophrenia and autism (see e.g.
37 Chakrabarti & Baron-Cohen, 2006).
38
39

40 The second key theme is identifying the brain structures involved in perceiving
41 emotion. The techniques used to trigger brain activity in healthy participants, and to
42 probe deficits in patients, are generally based on the work outlined above.
43
44

45 The literatures in both areas are large, but they do not generally have much effect on
46 the kind of argument that is being developed here. The converse is not true. If there
47 are problems with the standard types of experimental stimulus, or the analyses applied
48 to them, then they affect all of the literatures that make use of them.
49
50

51 In that context, it is a substantial concern that experimental research was so focused
52 on the task of deciding which of a few strong emotions a brief, archetypal stimulus
53 was conveying (or simulating). It is not obvious how effectively that kind of
54 experimental task much captures the everyday business of perceiving emotion.
55 Research in the technological style has brought that concern to the fore.
56
57
58

59 **4 RESEARCH IN THE TECHNOLOGICAL STYLE**

60 Picard’s book ‘Affective Computing’ (1997) signalled the arrival of research on
automatic techniques for detecting emotion-related states in human beings, and

1
2
3 responding to them appropriately. It has been influenced by experimental work on the
4 perception of emotion. But over time, technological research has increasingly been
5 drawn to contrasting conceptions of the problem, and different kinds of solution.
6
7

8 This section aims to convey the kind of understanding that is emerging from
9 technological research. It only considers research that uses the same modalities as
10 humans. There is interesting research using body-worn sensors (Kim & Andre, 2008),
11 but it has much less bearing on the perception of emotion by humans, and it is not
12 considered here.
13
14

15 **(a) Engagement with naturalistic material**

16 Around 2000, several groups became interested in samples of emotion that were
17 (broadly speaking) naturalistic. Computational research has a very particular reason to
18 deal with naturalistic material, since its applications are bound to be in the real world
19 rather than laboratory settings. Some psychologists moved in similar directions, to
20 some extent because technological developments made it feasible to work with
21 naturalistic material.
22
23

24 The concept of naturalistic material is not straightforward. For example, it is often
25 opposed to ‘acted material’. That can lead people to dismiss material that is taken
26 from real conversations, on the grounds that the participants are manipulating the
27 emotions that they show. In the worst case, they then revert to material which shows
28 emotion as they assume it would appear if there were no such manipulation – which is
29 usually produced by actors. Some studies avoid using actors by having random
30 members of the public simulate emotion.
31
32
33

34 A useful way of putting it is that material is naturalistic if it is of a kind that might
35 have to be dealt with in an application. The contrast is with idealised material, which
36 is generated to match someone’s conception of what emotion should be like.
37
38

39 In various ways, research using naturalistic material found itself facing issues that
40 work with idealised displays did not equip it to handle. Later sections consider the
41 details of solutions: this section concentrates on the signs that there were problems.
42
43

44 **(i) Speech**

45 In the context of speech, attempts to develop naturalistic databases provided early
46 indications that there were problems. A seminal conference (Cowie et al., 2000)
47 showed that several groups had encountered similar issues. They had turned to
48 existing recordings to look for clear examples of standard emotion categories
49 conveyed by voice, and found them much rarer than they expected. Roach et al
50 expected to find vocal expressions of emotion in clinical interviews, and found so
51 little that they turned to other sources. Douglas-Cowie et al (2003) turned to TV chat
52 shows, and while there was intense emotion, very little of it corresponded to a single
53 category. Although they selected samples to be as pure as possible, all but a very few
54 were given mixed labels by most raters. Campbell (2004) recorded phone
55 conversations over a long period, and concluded that the recordings did not contain
56 very much emotion as such. Later work identified some contexts where emotion did
57 occur, notably recordings from call centres, but even there, the frequency of clearly
58 emotional material was very low. For instance, Ang et al. (2002) used material
59
60

totalling 14 h 36 min. The commonest strong emotion was frustration, of which they obtained 42 unequivocal instances.

When technological research did use natural sources, results underscored the difference between it and acted material. Figure 1 conveys the point using a simplified classification into three levels of material. Fully stylized speech is produced by competent actors, often in a carefully structured format. The second level, mediated speech, includes emotion simulated by people without particular acting skill or direction; and samples selected from a naturalistic database as clear examples of the category being considered. The third level includes speech that arises spontaneously from the speaker's emotional state, and which includes naturally occurring shades, not only well-defined examples. As the figure shows, high recognition rates were restricted to material that was acted and/or carefully chosen. They also depended on reducing the task to a choice between a small number of alternatives. Dealing with naturalistic material, which might convey any emotion whatsoever, posed unsolved problems.

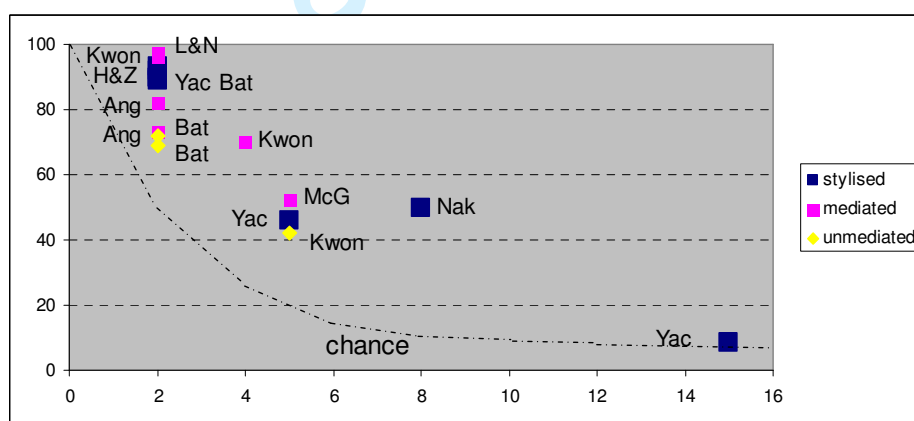


Figure 1: Plot of discrimination results from key early studies of emotion recognition – Lee and Narayan, 2003; Kwon et al., 2003; Zhou & Hansen, 1999; Ang, 2002; Yacoub et al., 2003; Batliner et al. 2003; McGilloway et al. 2000; and Nakatsu et al. 1999; against number of categories to be discriminated (horizontal axis)

Batliner et al (2003) added an important rider. One might assume that the simulations would present the same kinds of relationship as real emotionally coloured interactions, but in a cleaner form. The reality appears to be that there are significant ways in which they are simply different. When systems were trained on data generated by actors, they performed poorly in the application scenario. The same was true to a lesser extent when the training data came from simulated interactions.

Psychologists made a similar point around the same time. Bachorowski (1999) analysed speech produced by inducing emotion in realistic situations. Rather than sharp categorical distinctions, she argued that the speech tended to signal affective dimensions - activation level strongly, valence rather weakly. Hence different lines of research converged on the conclusion that extracting emotion-related information from speech in everyday contexts is not the same as categorising idealised samples.

(ii) Naturalism as a criterion in speech synthesis

1
2
3 The issue of naturalism took a different form in the context of speech synthesis,
4 reviewed by Schroeder (2001). Early research used formant synthesis techniques,
5 where rules (derived from experimental research) generate speech 'from scratch'.
6 Listeners could classify outputs produced by that approach in a forced choice task; but
7 they sounded too unnatural to convey emotion in a meaningful sense. The practical
8 consequence was that the field moved towards unit selection techniques, which splice
9 together samples taken from a human speaker. The implication is that creating a
10 convincing impression of emotion depends on details of the speech waveform that the
11 research underlying formant synthesis had discounted.
12
13
14

15 (iii) Vision

16 In one sense, research on facial expression addressed issues of natural and posed
17 expression long before research on speech, because of long standing interest in
18 sincerity and deception. However, the spirit of the research generally reflected the
19 impressionistic tradition. It focused on cues that a skilled observer could use to
20 distinguish posed from spontaneous smiles. The eye wrinkling associated with the
21 famous Duchenne smile is among the most widely cited, but actually occurs
22 frequently in posed smiles (Schmidt & Cohn, 2001). There is better evidence for
23 others, such as asymmetry (Frank, Ekman & Friesen, 1993) and amount of muscle
24 movement (Hess, Banse & Kappas, 1995). But although the cues exist, people tend
25 not to use them: even children can produce posed smiles that untrained observers fail
26 to discriminate from spontaneous smiles (Castanho & Otta, 1999).
27
28
29

30 More closely related to the work on speech is the kind of research described by
31 Carroll and Russell (1997), where the issue was not deception, but context. They
32 studied Hollywood movies which were regarded as well-acted, and extracted episodes
33 where there was strong agreement on a character's emotion. For happy episodes, the
34 corresponding facial expression usually involved the pattern of action units specified
35 by Ekman and Friesen's account. However, for other emotions, the expected pattern
36 of action units occurred in only 10% of cases.
37
38
39

40 The implication is that when emotions other than happiness occur in a complex,
41 ongoing situation, recognising them is not a matter of detecting highly specific
42 patterns of activity. This is a case where crude acted/naturalistic distinctions are
43 particularly unhelpful. There is a need to confirm that the finding is not simply
44 because of poor acting; but the point of the exercise is that it signals the need to
45 consider whether emotion is expressed in the course of action and interaction, or as an
46 end in itself. The perceptual problems that they pose may be very different.
47
48
49

50 (iv) Multimodality

51 Most of the work described above is unimodal. However, it includes some multimodal
52 sources, notably the Belfast Naturalistic Database. Comparing ratings of the different
53 modalities suggests that there are complex intermodal effects to be understood.
54
55
56
57
58
59
60

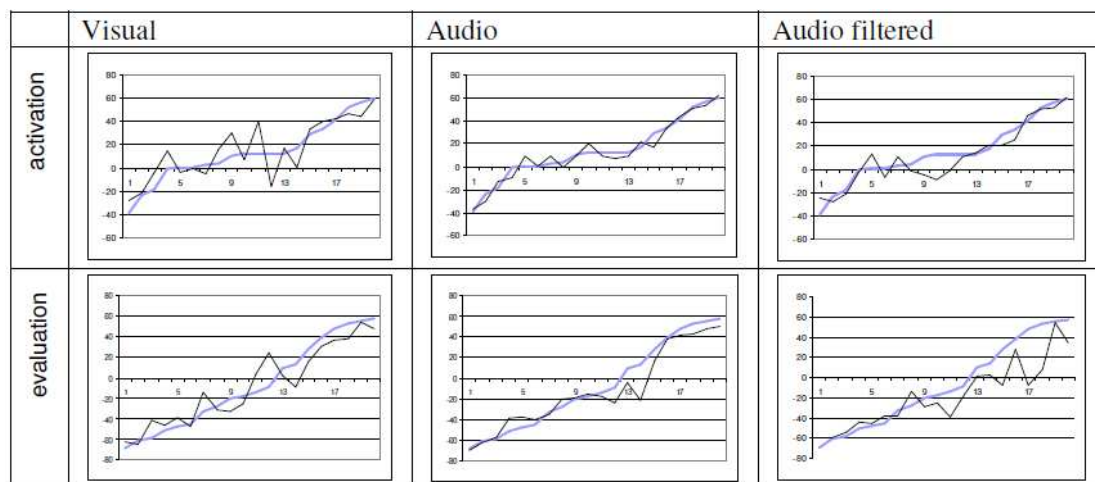


Figure 2: Multimodal effects in the Belfast Naturalistic database: Each panel shows average dimensional ratings for (a) the audio-visual versions of the clips (broad light line) and (b) one of the partial versions (thin dark line). Activation ratings are in the upper panels, evaluation in the lower.

Each panel in Figure 2 (from Douglas-Cowie et al., 2005) shows ratings of 20 clips on one of the standard emotion dimensions, juxtaposing ratings of the full audio-visual presentation with ratings of a single modality. The top left hand panel shows that removing audio modality produced erratic ratings of activation in clips where activation was judged to be moderate when full information was available: the bottom right shows that filtering the audio signal to remove linguistic information (leaving prosody relatively intact) led to substantial underestimates of valence in clips where valence was judged to be high when full information was available. These are prima facie evidence of rather complex interactions in which different modalities tend to make different contributions.

The point made in this section is a very general one: perceiving emotion in naturalistic contexts seems to be a substantially different task from perceiving it in set-piece or posed material. That provides a motivation to explore several more specific avenues.

(b) What is the output of emotion perception?

Research in the experimental style makes it natural to assume that the outcome of emotion perception is straightforward: it involves assigning a category label roughly corresponding to an everyday emotion term, such as 'angry' or 'happy'. There are multiple reasons to question that.

As a starting point, technological research takes its impetus from the belief that emotion colours very large parts of human life, and is practically important for that reason. However, as database research recognised early on, material that is well described by prototypical labels such as anger and happiness is rather rare. If there is a widespread phenomenon to study, it does not consist of assigning labels like that.

Language compounds the problem. Many psychologists reserve the term 'emotion' for phenomena that are at least close to prototypical emotions (e.g. Scherer, 2005). Adopting that convention would leave people without a convenient way to refer to the phenomena that simple category labels fail to capture. A convention designed to avoid

1
2
3 that difficulty (Cowie, in press) uses the term ‘emotional life’ to cover all the parts of
4 human life that distinguish it from the life of a being who, like Star Trek’s Mr Data, is
5 always unemotional; and ‘pervasive emotion’ (following Stocker and Hegman, 1992)
6 is used to describe what is present when a person is not truly unemotional
7
8

9
10 It is pervasive emotion, rather than prototypical emotional episodes, that technology
11 has obvious reasons to address. The corresponding perceptual task is not labelling
12 prototypical emotion episodes, but registering the emotional colouring that pervades
13 emotional life. Addressing that task poses conceptual challenges, the most basic of
14 which is to develop workable ideas about the kinds of representation that a perceptual
15 system concerned with emotional life might use to specify what it sees and hears.
16

17
18 An intuitive option is to use a description that is based on categories, but that extends
19 well beyond prototypical emotions. Several attempts have been made to develop
20 appropriate lists on pragmatic grounds, by cumulating categories that research has
21 consistently found useful. Examples are the ‘Basic English Emotion Vocabulary’
22 (Cowie et al 1999) and the derivative list used in the HUMAINE database (Douglas-
23 Cowie et al 2007). A more principled approach due to Baron-Cohen (2004) has
24 attracted considerable interest (El Kaliouby et al 2005). One of its notable features is
25 that it covers ‘affective epistemic’ states, such as ‘sure’ or ‘thoughtful’, as well as
26 states that are purely emotional.
27
28

29
30 There is a more generic typology to be considered beyond the distinctions discussed
31 so far. It involves distinguishing classes of phenomena like short-lived, intense
32 emotions; moods; long-lasting ‘established’ emotions (such as grief or shame);
33 stances; attitudes, and so on. These are practically important, and a competent
34 perceiver should be able to judge whether someone is angry because of a specific
35 event, or a long standing grievance, or simply in a bad mood. They are also interesting
36 because they relate to frequency of occurrence, and therefore to arguments about what
37 is worth perceiving: moods and stances make up a very large part of emotional life,
38 unbridled emotions rather little (Wilhelm & Schoebi 2007; Cowie, in press).
39
40

41
42 It is clearly that emotion perception is not simply deciding which category to apply.
43 For instance, particularly with subtle or complex emotions, it may take time and effort
44 to find a category that even approximately captures the state a person appears to be in.
45 That implies an underlying representation to which categories are fitted. At least four
46 possible ways of capturing that underlying representation are of interest.
47

48
49 The least radical option extends categorical description by using ‘soft vectors’
50 (Laukka, 2004). The vector that describes a state consists of multiple category labels,
51 each associated with a numerical estimate of confidence that the relevant state is
52 present (El Kaliouby & Robinson 2005; Douglas-Cowie et al 2005; Batliner et al
53 2006). The approach is limited by the lack of consensus on a set of categories that
54 could combine to capture the range of percepts that people clearly form.
55

56
57 Dimensional representations are a means of addressing essentially that problem which
58 have a long history in Psychology. The simplest version, which describes emotion in
59 terms of valence and arousal, was imported into technological research early (Cowie
60 et al., 2001). One of its attractions is that it provides a reasonable way of capturing the
colouring that people perceive in moderately emotional interactions (Cowie &

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Cornelius, 2003). Fuller dimensional schemes have emerged more recently, notably one due to Fontaine et al (2007), which adds a dimension related to power and one related to predictability.

There are some indications that dimensional schemes are more than a pragmatic way of summarising information that perceptual systems make available. In studies where people rate recordings of emotional displays in terms of either valence and arousal (Cowie & Cornelius 2003) or the fuller Fontaine set (Devillers et al., 2006), judges tend to assign dimensional descriptions more reliably than categorical, suggesting that the dimensional judgments are not derived from more basic categorical assignments.

Another option is to propose that perceiving another person's emotion amounts to perceiving the appraisals that he or she forms. The idea is attractive because of appraisal theory's logical elegance, and it would be more so if, as has been argued, the signs provided by facial expressions reflect elements of appraisal more directly than holistic emotion categories (Wehrle et al 2000). Considering its attractions, there is surprisingly little empirical work on the idea. But when observers have been asked to rate appraisal-related states in other people from audiovisual recordings, agreement was relatively low (Devillers et al., 2006).

A radically different option proposes that embodiment is fundamental to the perception of emotion: to perceive an emotion is to some extent to re-enact it (Niedenthal et al, 2005). There is certainly evidence of interactions between the perceiver's bodily state and the perception of emotion: the issue is whether bodily states affect the perception of emotion or constitute it. Similar issues have not been easy to resolve in related areas (Moore, 2007), and they are not likely to be easily resolved in the context of emotion. A related, but distinct point is that people may act in response to cues that they have not consciously registered. Hence guidance of action should be counted among the possible outputs of emotion-related perceptual systems.

Although the options outlined above are different in many ways, there is an important common thread. All of the representations involve multiple elements which vary over time. The output of emotion perception can therefore be visualised as a family of 'traces' which fluctuate over time. The HUMAINE database (Douglas-Cowie et al, 2007) provides a concrete illustration of the way a set of time-varying traces might capture the perceived emotional content of emotionally coloured situations.

Trace-like representations apply most naturally to feeling-like elements of emotion. It would be natural to call them affective if the term were not used in so many other ways. Elements with closer links to cognition are also important, though.

The most basic is what philosophical accounts call the object of emotion. Not all emotion-related states have objects (mood is generally thought not to); but it would seem eccentric to say that the perception of emotion includes registering (for instance) that a person is angry, but not whether the anger is directed at the perceiver or something else.

1
2
3 The concept of active perception is well established in other areas, and it seems highly
4 relevant to emotion. It is very common for the immediate outcome of perception to be
5 uncertainty that prompts a question designed to clarify how the person is.
6
7

8 Last and not least, choosing appropriate words to describe emotions is often an
9 important part of the process. It is a highly complex one, which is clearly dependent
10 on culture, and involves judgments about causes, perceptions, justifications,
11 entitlement, and so on (Cowie, 2005). It should no more be equated with the whole of
12 emotion perception than colour naming is equated with colour perception.
13
14

15 Many of the issues raised in this section have been addressed from a technological
16 standpoint by a W3C incubator group seeking to develop a standard motion markup
17 language, known as EmotionML (Schroeder, 2008). In effect, what the project offers
18 is formalism for capturing the output of a competent emotion perception system.
19
20

21 Work on the issues considered in this section is ongoing. However, it suggests an
22 image of emotion perception which is far removed from the image implicit in
23 classical experimental research. What emotion perception does in natural contexts is
24 to construct a multidimensional, time-varying stream that is attuned to events both
25 within and around the person perceived, and affects both awareness and action. The
26 means by which that is achieved clearly cannot be quite like those that were envisaged
27 by classical experimental research.
28
29
30

31 **(c) What are the cues?**

32 Questions about relevant cues are quite open in all modalities. Different groups favour
33 different sets, and consensus is slow to develop because comparison is difficult. The
34 point is illustrated by the CEICES project, in which teams deliberately ensured that
35 their work on speech could be compared (Batliner et al., 2006). Even there, the teams
36 have continued to favour very substantially different methodologies.
37
38
39

40 In that context, it would be wrong to make strong or specific claims about the state of
41 the art. Hence the section aims to pick out issues that are intellectually interesting, and
42 which impact on the way we think about recognising emotion, rather than to
43 summarise the technology.
44
45

46 **(i) Speech**

47 Two things are striking in contemporary approaches to recognising emotion from
48 speech. The first is the shift from acted corpora towards natural sources. The second is
49 the number of features considered. Classical experimental papers consider small
50 numbers of features - 34 features in Banse & Scherer's 1996 study, 14 in Juslin &
51 Laukka's (2003) review. It is commonplace for contemporary papers to consider
52 thousands of features.
53
54

55 If it is the case that the number of features relevant to perceiving is substantial, then
56 two broad types of interpretations are natural.
57
58

59 One is suggested by evidence that emotional passages of speech are much more likely
60 to be rated as degraded communication than non-emotional passages from the same
interactions (Cowie and Cornelius, 2003). Many of the features associated with

1
2
3
4 emotional speech may essentially reflect impaired control of the complex processes
5 involved in fluent speech production, resulting breakdowns and simplifications that
6 can take a virtually limitless variety of forms. That is consistent with the description
7 of a moderately large feature set provided by Cowie & Douglas-Cowie (in press). The
8 features associated with emotion depend on speaker gender, length of utterance, and
9 person judging, suggesting a rather anarchic process.

10
11
12 The second is that the features amount in effect to a dense picture of some underlying
13 setting – more akin to a picture than to a list of discrete attributes. That is consistent
14 with the observation by Schuller et al. (in press) that information seems to be
15 concentrated in spectral features (in contrast with emphasis on pitch and intensity in
16 early research).

17
18
19 Both may well be partially true if there are differences in the way different levels of
20 emotion are signalled. The Schuller study considered simulated intense emotions. It is
21 not surprising that spectral parameters are important in that context given the link
22 between them and changes of tension and setting in the vocal tract. The Cowie studies
23 used naturalistic material, with moderate levels of emotion predominating.

24
25
26 A wide range of more specific ideas is being explored. There are now systems that
27 track speech through a space with three dimensions (valence, arousal and time),
28 exploiting constraints on expected rate of change (Wollmer et al., 2008). Predictably,
29 arousal is easier to track than activation. Voice quality is still elusive, with evidence
30 both for and against its contribution (Schuller et al., in press). Linguistically
31 motivated descriptions of intonation have been incorporated into analysis, but seem
32 not to enhance recognition (Batliner et al., 2006).

33 34 35 (ii) Face

36 Contemporary research highlights at least three major shifts in thinking about the
37 perception of emotion from facial expression.

38
39
40 The first shift is engaging with the facial patterns produced during dynamic
41 expression of emotion. Scherer and Ellgring (2007) coded the facial actions used by
42 actors simulating strong emotions. They concluded: “We do not find any *complete or*
43 *full* prototypical patterns for basic emotions” (p. 126). Instead, the data showed many
44 activations of one or a few action units. That suggests perception cannot rely on
45 distinctive local patterns: it must integrate information across multiple times and/or
46 multiple modalities.

47
48
49 The second shift is engaging with naturalistic data. Sneddon & McRorie (2007)
50 compared sequences from an acted database with sequences from naturalistic
51 recordings of strong emotions. Raters examined individual frames (in a random
52 order), and rated the emotion conveyed by each. Frame-to-frame change in rated
53 emotion was then derived. It was much greater in the naturalistic material.
54 Impressionistically, that seems to be because people in the naturalistic recordings
55 were shifting focus rapidly, and expressing different reactions to different aspects of
56 their situation. The implication is that the problem of integrating evidence over time
57 and modalities is even greater than studies like Scherer & Ellgring’s imply.
58
59
60

1
2
3 The third shift is engaging with moderate emotional colouring. Classical research
4 found mixed evidence for timing effects, but that changes when the emotions are
5 moderate. Ambadar, Schooler, and Cohn (2005) showed robust effects of movement
6 for the identification of subtle emotions: expressions that were not identifiable in
7 static presentations were clearly apparent in dynamic displays.
8
9

10 A final shift involves the FACS system. It has clear advantages over methods based
11 on direct matching to a few global templates; but it also has problematic features. On
12 one hand, it is designed for extreme expressions, and difficult to apply to moderate
13 emotional colouring: on the other hand, it forces systems to locate points precisely
14 when information seems to be available at a coarse level (Tian et al, 2005). There are
15 few obvious alternatives, but good reasons to look for candidates.
16
17

18 In summary, it seems increasingly likely that classical analyses of facial expression
19 apply neatly to selected ideal cases, but bypass problems that are central to dealing
20 with the variety and indefiniteness of everyday life. Patterns distributed across time
21 and modality need to be found and disentangled, and that is a major challenge.
22
23

24 25 (iii) Gesture

26 Computational research on gesture and emotion is very active, and very diverse. It
27 deals with phenomena from subtle finger movements to dancing; it considers both
28 how they can be detected and the perceptual effect of synthesising them; techniques
29 of analysis range from precise recovery of local structure to global flow; and the
30 techniques used to describe a particular kind of gesture for the purpose of synthesis
31 often bear strikingly little relationship to those used in the context of detection. A
32 brief comment on that kind of field must necessarily be highly selective.
33
34

35 An issue that is particularly interesting arises from synthesis. Refining gesture is
36 regarded as a way to overcome a cluster of issues related to naturalness, 'stiffness',
37 and responsiveness. For instance, humans exhibit 'idle movement' even when they
38 are standing on a single spot. Agents with no idle movement give a disconcertingly
39 'wooden' effect. Genuine smiles tend to be accompanied by head and shoulder
40 movements (Valstar et al., 2007), and agents that smile without those movements
41 have a similar effect. Various gestures – notably smiles and head nods – play an
42 important part in backchannelling during a conversation (Heylen 2007), and it has
43 been argued that the lack of backchannelling contributes to the difficulty of sustaining
44 interaction with synthesised agents.
45
46
47
48

49 These effects expose an aspect of perception that is normally taken for granted.
50 Human emotional engagement depends on perceiving not only what the other party's
51 emotional state is, but also that the other party is engaging emotionally. When agents
52 are unable to give cues that signify engagement, emotion can be identified, but
53 emotional rapport cannot be built. If that is correct, then perceiving those cues is a
54 non-trivial part of emotional life.
55
56

57 One of the keys to exploring these issues is an agent with the ability either to display
58 or to omit the relevant responses, singly or in combination. Schroeder et al, (2008)
59 have reported work towards that.
60

1
2
3
4 (iv) Multi-modality

5 Multimodality has become increasingly central to the domain, but the situation is not
6 straightforward. Computational research concluded quickly that audio and facial
7 expressions presented complementary information (Busso et al 2004) However,
8 Scherer & Ellgring (2007b), analysing a substantial multimodal database, found rather
9 few multimodal patterns (notably high vocal arousal accompanied by stretching of the
10 mouth and arms: and low vocal arousal accompanied by slumped upper body, eyelid
11 droop, and back of hands pointing forward).
12

13
14 All of the studies reported above used acted data. Naturalistic data raises different
15 issues. In a study using the Belfast Naturalistic Database, five raters judged how
16 concordant or discordant audio and visual indicators were. Perfect concordance was
17 very rare, apparently because different channels expressed different aspects of the
18 person's emotional status – positive towards the interlocutor, negative about the
19 events being described. The most marked divergence occurred where raters with
20 access to all the modalities identified the dominant emotion as anger.
21
22

23 Rather surprisingly, all the research points to the same conclusion. Different
24 modalities do tend to complement each other. In acted data, they offer different
25 components of a vector that points to a single conclusion. In naturalistic data, they are
26 likely to reflect different aspects of the way people react to their situation. In either
27 case, perceptual processes are systematically sensitive to information in multiple
28 modalities.
29
30

31
32 (v) Context

33 Computational research has developed appealing models of the way context might
34 contribute to emotion perception. For example, Conati (2002) has described
35 techniques where a probabilistic model assesses affect by integrating evidence from
36 two sources: on one hand, both possible causes of emotion (i.e., the state of the
37 interaction); and on the other, signs that that are expected to be influenced by
38 emotional reactions.
39
40

41 That kind of model raises two kinds of question for experimental work. One is how
42 context affects classification of emotions subtler than the prototypical categories that
43 dominated classical experimental research. The other is how perceivers identify the
44 object of an emotion. The only obvious options involve observing context and
45 understanding speech. Identifying the object was rarely a central topic in experimental
46 research, but as noted earlier, distinguishing between 'angry with me' and 'angry with
47 my assailant' is not a minor issue.
48
49

50 A second type of context effect was highlighted by Caldwell (2000). He showed that
51 the same speech sample evoked reliable impressions of anger in isolation, but was
52 judged neutral in a context which allowed listeners to attune to the speaker's habitual
53 settings. Adaptation to speaker characteristics is a major challenge both for pure
54 science and for technology.
55
56

57 (vi) Hypotheses about mechanisms

58 Initially, technological research explored various types of classification rule.
59 Interesting lines of division have emerged gradually. There is support for rules based
60 on decision trees, but it seems limited. Bayesian rules arouse more interest, partly

1
2
3 because they appear to model other perceptual phenomena well. There is also
4 widespread interest in techniques that take account of time, including Hidden Markov
5 Models and recurrent neural nets.
6
7

8 Among the highest profile issues is whether to use early or late fusion for multimodal
9 inputs. There is evidence that late fusion has advantages technologically (Valstar et
10 al., 2007). That meshes reassuringly with familiar observation. People can and do
11 register that a person's face is telling one emotional story, but his or her voice is
12 telling another. However, everyday observation also warns us that people do not
13 always notice the telltale discrepancy. In addressing that kind of observation,
14 technological research is making its way back to issues at the heart of the
15 impressionistic style.
16
17

18 19 20 **5 HOW DO WE UNDERSTAND THE TASK?**

21 There is a saying that 'fish will be the last to discover water'. Ongoing, fluent
22 responsiveness to emotional colouring in other people's expressions and actions is so
23 fundamental to human life that it is very easy to take for granted. Research has shifted
24 that stance gradually. It began by indicating how the perception of emotion could be
25 improved, and then moved to examine how the ordinary person identified sharply
26 distinct cases. The theme of this paper has been that technological research opens the
27 way for a third reassessment, by trying to match what people do without conscious
28 effort, and sometimes without conscious awareness.
29
30

31
32 No doubt others will disagree with the way that this paper has drawn the shape of the
33 field. But there is a clear need to mark out the shape of the field, and provoking others
34 to do a better job is not a trivial contribution.
35

36 **ACKNOWLEDGEMENT**

37 The research leading to this paper has received funding from the European
38 Community's Seventh Framework Programme (FP7/2007-2013) under grant
39 agreement n° 211486 (SEMAINE)
40
41

42 **REFERENCES**

43
44 Adolphs, R. 2005 Perception and Emotion: How We Recognize Facial Expressions *Current*
45 *Directions in Psychological Science* 15 (5) 222 -226.
46
47

48 Ambadar, Z Schooler, and Cohn, J 2005 Deciphering the enigmatic face: The importance of facial
49 dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5):403-410,
50

51 Ang, J. Dhillon, R. Krupski, A. Shriberg, E.& Stolcke, A. 2002 Prosody-based
52 automatic detection of annoyance and frustration in human-computer dialog. *Proc.*
53 *ICSLP 2002, Denver, Colorado.*
54

55 Bachorowski, J-A. 1999 Vocal Expression and Perception of Emotion *Current*
56 *Directions in Psychological Science* 8 (2), 53-57.
57

58 Banse, R and Scherer, K, R. 1996 Acoustic profiles in vocal emotion expression.
59 *Journal of Personality and Social Psychology* 70 (3), 614-636.
60

1
2
3 Baron-Cohen, S, Golan, O, Wheelwright S, & Hill, J.J. 2004 *Mind Reading: The*
4 *Interactive Guide to Emotions*. London: Jessica Kingsley Publishers.

5
6
7 Bassili J. N 1978 Facial motion in the perception of faces and of emotional expression.
8 *Journal of Experimental Psychology*, 4:373–379.

9
10 Bassili J. N. 1979 Emotion recognition: The role of facial motion and the relative importance
11 of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–
12 2059.

13
14
15 Batliner, A, Fischer, K. Huber, R. Spilker, J. Noeth, E. 2003 How to find trouble in
16 communication. *Speech Communication* 40,117-143.

17
18 Batliner, A. Steidl, S. Schuller B., Seppi D., Laskowski K., Vogt T., Devillers, L.,
19 Vidrascu L, Amir N., Kessous L., Aharonson V. 2006 Combining efforts for
20 improving automatic classification of emotional user states. In Erjavec, T. and Gros,
21 J. (Ed.), *Language Technologies, IS-LTC 2006*. pp. 240-245

22
23
24 Biederman, I 1987 Recognition-by-Components: A Theory of Human Image
25 Understanding *Psychological Review* 94 (2),115-147

26
27
28 Birdwhistell, RL 1970 *Kinesics and context* Philadelphia: University of Pennsylvania
29 Press.

30
31 Burkhardt, F & Sendlmeier W. F. 2000 Verification of acoustical correlates of
32 emotional speech using formant-synthesis. *Proceedings of the ISCA Tutorial and*
33 *Research Workshop (ITRW) on Speech and Emotion*, 151-156.

34
35
36 Busso, C Deng, Z Yildirim, S Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S.,
37 Neumann, U. & Narayanan, S. 2004 Analysis of Emotion Recognition using Facial
38 Expressions, Speech and Multimodal Information *Proc 6th International conference*
39 *on Multimodal Interfaces* pp. 205 – 211.

40
41
42 Cahn, J. E., 1990 The Generation of Affect in Synthesized Speech, *Journal of the*
43 *American Voice I/O Society*, 8, p. 1-19

44
45 Campbell, N 2004 Speech & Expression; the Value of a Longitudinal Corpus
46 *Proceedings of LREC 2004*

47
48
49 Carlson, R., Granström, B. & Lennart, N. 1992 Experiments with emotive speech -
50 acted utterances and synthesized replicas. *ICSLP-1992*, pp. 671-674.

51
52
53 Carroll, J.M. & Russell, J.A 1997 Facial Expressions in Hollywood's Portrayal of
54 Emotion *Journal of Personality and Social Psychology*. 72 (1), 164-176.

55
56 Castanho AP, Otta E. 1999 Decoding spontaneous and posed smiles of children who
57 are visually impaired and sighted. *J Vis Impair Blind* 93, 659–662.

58
59
60 Chakrabarti, B. & Baron-Cohen, S. 2006 Empathizing: neurocognitive developmental
mechanisms and individual differences *Progress in Brain Research*. 156, 403-17

1
2
3 Conati, C. 2002 Probabilistic Assessment of User's Emotions in Educational Games.
4 *Journal of Applied Artificial Intelligence* 16 (7&8), 555-575.

5
6 Cowie, R 2005 What are people doing when they assign everyday emotion terms?
7 *Psychological Inquiry* 16 (1) 11-18

8
9
10 Cowie, R (in press) Describing the forms of emotional colouring that pervade
11 everyday life To appear in P. Goldie (ed) *Oxford Handbook of the Philosophy of*
12 *Emotion*

13
14
15 Cowie, R. & Cornelius, R. 2003 Describing the emotional states that are expressed in
16 speech. *Speech Communication* 40 (1-2): 5-32.

17
18
19 Cowie, R. Douglas-Cowie, E Apolloni, B Taylor, J. Romano, A. & Fellenz, W. 1999
20 What a neural net needs to know about emotion words. In N. Mastorakis (ed.),
21 *Computational Intelligence and Applications*. World Scientific Engineering Society,
22 109-114.

23
24
25 Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W.
26 and Taylor, J. 2001 Emotion recognition in human-computer interaction. *IEEE Signal*
27 *Processing Magazine* 18 (1): 32-80.

28
29
30 Cowie, R. & Douglas-Cowie, E. (in press) Prosodic and related features that signify
31 emotional colouring in conversational speech To appear in S.Hancil (ed) *The Role of*
32 *Prosody in the Expression of Emotions in English and in French* Berne: Peter Lang

33
34
35 Cowie, R. Douglas-Cowie E & Schroeder M. 2000 (eds.), *Speech and Emotion:*
36 *Proceedings of the ISCA workshop*

37
38
39 Crane, E & Gross M 2007 Motion capture and emotion: Affect detection in whole
40 body movement *Affective Computing and Intelligent Interaction 2007* Springer,
41 Berlin: pp. 95-101.

42
43
44 de Gelder, B & Vroomen, J 2000 The perception of emotions by ear and by eye
45 *Cognition and Emotion* 14 (3), 289-311

46
47
48 L. Devillers, R. Cowie, J-C. Martin, E. Douglas-Cowie, S. Abrilian, M. McRorie
49 2006 Real life emotions in French and English TV video clips: an integrated
50 annotation protocol combining continuous and discrete approaches . *Proc. 5th Int.*
51 *Conf. on Language Resources and Evaluation (LREC)*. Genoa, Italy

52
53
54 Dimberg U, Thunberg M, Elmehed K. 2000. Unconscious facial reactions to
55 emotional facial expressions. *Psychol. Sci.* 11, 86-89

56
57
58 Dittrich, WH Troscianko, T., Lea, S.E.G., & Morgan, D 1996 Perception of emotion
59 from dynamic point-light displays represented in dance *Perception* 25 (6), 727-738

60
61
62 Douglas-Cowie, E. Campbell, N. Cowie R. & Roach, P. 2003 Emotional speech:
63 towards a new generation of databases. *Speech Communication* 40 (1-2), 33-60.

1
2
3 Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M.,
4 Martin, J-C., Devillers, L., Abrilian S., Batliner, A., Amir, N. and Karpouzis K.. 2007
5 The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic
6 and Induced Emotional Data. *Affective Computing and Intelligent Interaction 2007*
7 Berlin: Springer Verlag, pp. 488-500
8
9

10 Douglas-Cowie, E, Devillers, L, Martin, J-C, Cowie, R, Savvidou, S., Abrilian, S.,
11 Cox, C. 2005 Multimodal Databases of Everyday Emotion: Facing up to Complexity
12 *Interspeech 2005* pp. 813-816.
13
14

15 Ekman P. 1982. *Emotion in the Human Face*. Cambridge University Press: New
16 York.
17

18 Ekman, P Friesen WV 1976 *Pictures of facial affect* Palo Alto, CA: Consulting
19 Psychologists Press
20

21 Ekman, P., & Friesen WV 1976 Measuring facial movement *Journal of Nonverbal*
22 *Behavior*, 1(1), 56-75
23
24

25 Ekman, P., & Friesen, W.V. 1982 Felt, false and miserable smiles. *Journal of*
26 *Nonverbal Behavior*, 6, 238–252.
27
28

29 Ekman, P & O’Sullivan M 1991 Who can catch a liar? *American Psychologist*, 46
30 (9), 913-920.
31
32

33 Fernández-Dols, JM Wallbott, & H Sanchez F 1991 Emotion category accessibility
34 and the decoding of emotion from facial expression and context *Journal of Nonverbal*
35 *Behavior* 15(2), 107-123.
36
37

38 Fontaine, J., Scherer, K., Roesch, E. & Ellsworth, P. 2007 The World of Emotions Is
39 Not Two-Dimensional *Psychological Science* 18 (12), 1050 – 1057
40

41 Frank MG, Ekman P, Friesen WV. 1993 Behavioral markers and recognizability of
42 the smile of enjoyment. *J Pers Soc Psychol.* 64:83–93
43
44

45 Goldberg, H. 1951 The role of ‘cutting’ in the perception of motion pictures,
46 *Journal of Applied Psychology* 35, 70–71.
47

48 Grammer, K, Kruck, KB, & Magnusson MS 1998 The courtship dance: Patterns of
49 nonverbal synchronization in opposite-sex encounters *Journal of Nonverbal Behavior*
50 22, 3–29
51

52 Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. and Wedin, L. 1980 Perceptual
53 and acoustic correlates of voice qualities. *Acta Otolaryngologica* 90, 441-451.
54
55

56 Hatfield, E. & Rapson, R. L. 2000 Emotional contagion. In W. E. Craighead & C. B.
57 Nemeroff (Eds.). *The Corsini encyclopedia of psychology and behavioral science*. New
58 York: John Wiley & Sons, 493-495.
59

60 Hatfield, E., Cacioppo, J. T., & Rapson, R. L. 1994 *Emotional contagion*. New York:
Cambridge University Press.

1
2
3
4 Hess, U., Banse, R. and Kappas, A. 1995 The intensity of facial expression is
5 determined by underlying affective state and social situation. *J Pers Soc Psychol.* 69,
6 280–288.
7

8
9 Heylen, D., Bevacqua, E., Tellier, M., & Pelachaud, C. 2007 Searching for
10 Prototypical Facial Feedback Signals *Proc IVA 2007* pp.147-153
11

12
13 Juslin, P & Laukka, P 2003 Communication of Emotions in Vocal Expression and
14 Music Performance: Different Channels, Same Code? *Psychological Bulletin* 129(5),
15 770–814
16

17
18 El Kaliouby, R & Robinson P. 2005 Real-Time Inference of Complex Mental States
19 from Facial Expressions and Head Gestures In B. Kisačanin, V. Pavlović & T.S.
20 Huang *Real-Time Vision for Human-Computer Interaction Part II* Berlin: Springer
21 pp. 181- 200
22

23
24 Kim, J & Andre, E. 2008 Emotion Recognition Based on Physiological Changes in
25 Music Listening *IEEE Trans Pattern Analysis & Machine Intelligence* 30 (12), 2067-
26 2083.
27

28
29 Kwon, O. Chan, K. Hao, J. 2003 Emotion recognition by speech signals. *Proc.*
30 *Eurospeech 2003, Geneva* pp. 125-128.
31

32
33 Lacava , P.G., Golan, O., Baron-Cohen, S.,& Myles, B.S. 2007 Using Assistive
34 Technology to Teach Emotion Recognition to Students With Asperger Syndrome A
35 Pilot Study *Remedial and Special Education* 28(3), 174-181
36

37
38 Laukka, P. 2004 *Vocal expression of emotion* PhD thesis, University of Upsaala
39

40
41 Laukkanen, A-M., Vilkman, E., Alku, P. & Oksanen, H. 1996 Physical variations
42 related to stress and emotional state: a preliminary study *Journal of Phonetics* 24,
43 313-335.
44

45
46 Laver, J. 1980 *The Phonetic Description of Voice Quality*. Cambridge: CUP.
47

48
49 Lazarus, R.J. 1999 The Cognition-Emotion Debate: A Bit of History
50 In T Dalglish & MJ Power (eds) *Handbook of cognition and emotion*. Chichester:
51 Wiley. pp 1-19.
52

53
54 Lee, C., Narayanan, S. 2003 Emotion recognition using a data-driven fuzzy inference
55 system. *Proc. Eurospeech 2003, Geneva*, 157-160.
56

57
58 Lieberman, P. & Michaels, S.B. 1962 Some Aspects of Fundamental Frequency and
59 Envelope Amplitude as Related to the Emotional Content of Speech *J. Acoust. Soc.*
60 *Am.* 34 (7) 922-927

Massaro, D.W. 2004 A Framework for Evaluating Multimodal integration by
Humans and A Role for Embodied Conversational Agents *Proc. International
Conference on Multimodal Interfaces 2004*, 24-41

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, W., Stroeve, S. 2000 Automatic recognition of emotion from voice: a rough benchmark. *Proc. ISCA ITRW on Speech and Emotion*, pp. 207-212.
- McRorie, M & Sneddon, I. 2007 Real Emotion Is Dynamic and Interactive *Proc Affective Computing and Intelligent Interaction 2007* Berlin: Springer pp. 759-760
- Milner, A.D. & Goodale, M.A. 1995 *The Visual Brain in Action* Oxford: Oxford University Press.
- Moore, R.K. 2007 Spoken language processing: Piecing together the puzzle *Speech communication* 49(5), 418-435
- Mozziconacci, S. 1998 *Speech variability and emotion: Production and perception* PhD thesis: University of Eindhoven.
- Munn, N.1940 The effect of knowledge of the situation upon judgment of emotion from facial expressions *Journal of Abnormal and Social Psychology* 35, 324-338.
- Murray, I. R., & Arnott, J. L. 1995 Implementation and testing of a system for producing emotion-by-rule in synthetic speech, *Speech Communication* 16, 369-390.
- Nakatsu, R. Tosa, N. Nicholson, J. 1999 Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Proc. IEEE International Workshop on Multimedia Signal Processing* pp. 439-444.
- Niedenthal P.M., Barsalou, L.W., Winkielman, P., Krauth-Gruber, S. & Ric, F. 2005 Embodiment in Attitudes, Social Perception, and Emotion *Personality and Social Psychology Review* 9(3), 184–211.
- Niedenthal, P.M., Halberstadt, J.B., Margolin, J., Innes-Ker, A. 2000 Emotional state and the detection of change in facial expression of emotion *European Journal of Social Psychology* 30(2), 211 – 222
- Oster A, -M. & Risberg, A. 1986 The identification of the mood of a speaker by hearing impaired listeners *STL-QPSR* 4,79-90.
- Pandzic, I. & Forscheimer, R. 2002 The origins of the MPEG-4 Facial animation standard In I. Pandzic & R. Forscheimer (eds) *MPEG-4 Facial animation: The standard, Implementation and Applications* Chichester: Wiley. pp 3-13
- Picard, R. 1997 *Affective computing* Cambridge, Mass: MIT Press
- Pollick, F.E., Paterson, H.M., Bruderlin, A. & Sanford, A.J. 2001 Perceiving affect from arm movement *Cognition* 82 B51–B61.
- Poggi, I 2006 *Le parole del corpo: introduzione alla comunicazione multimodale* Rome: Carocci.
- Prom-on, S., Xu, Y. and Thipakorn, B. 2009 Modeling tone and intonation as target approximation *J. Acoust. Soc. Am.* 125(1), 406 – 424

1
2
3
4 Russell, J. A., Bachorowski, J-A and Fernandez-Dols, J-M. 2003 Facial and vocal
5 expressions of emotion. *Ann. Rev. Psychol.* 54:329–49.

6
7 Scherer, K. R. & Ellgring, H. 2007a Are facial expressions of emotion produced by
8 categorical affect programs or dynamically driven by appraisal? *Emotion* 7, 113–130
9

10 Scherer, K.R. & Ellgring, H. 2007b Multimodal expression of emotion: Affect
11 programs or componential appraisal patterns? *Emotion* 7(1), 158–171
12

13 Schmidt, K.L. & Cohn, J. 2001 Human Facial Expressions as Adaptations:
14 Evolutionary Questions in Facial Expression Research *Am J Phys Anthropol.* 2001;
15 Suppl 33: 3–24.
16
17

18 Schroeder, M. 2001 Emotional Speech Synthesis: A Review *Proc. Eurospeech 2001,*
19 *ISCA, Bonn, Germany.* pp. 561–564
20

21 Schröder, M. (ed) 2008 Elements of an EmotionML 1.0 W3C Incubator Group Report
22 <http://www.w3.org/2005/Incubator/emotion/XGR-emotionml/>
23

24 Schroeder, M Cowie, R., Heylen,D, Pantic, M, Pelachaud, C., & Schuller, B. 2008
25 Towards responsive Sensitive Artificial Listeners Paper presented to the 4th
26 *International Workshop on Human-Computer Conversation, Bellagio* October 2008
27

28 Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G. (in press) *Spectral or Voice Quality?*
29 *Feature Type Relevance for the Discrimination of Emotion Pairs.* To appear in S
30 Hancil (ed.) *The Role of Prosody in the Expression of Emotions in English and in French*
31 Berne: Peter Lang
32

33 Sengers, P., Liesendahl, R., Magar, W., Seibert, C., Müller, B., Joachims, T., Geng,
34 W., Martensson, P., Höök, K. 2002 The Enigmatics of Affect. *Proceedings of DIS*
35 *2002* NY: ACM Press, 87-98
36

37 Stocker, M. & Hegman, E. 1992 *Valuing Emotions* Cambridge: Cambridge University
38 Press
39

40 Terzopoulos, D & Waters, K. 1993 Analysis and synthesis of facial image sequences
41 using physical and anatomical models *IEEE PAMI* 15(6), 569-579
42

43 Tian, Y-I, Kanade, T., & Cohn, J. 2005 Facial Expression Analysis In S.Z Li &
44 A.K.Jain, (eds) *Handbook of face recognition* Berlin: Springer pp. 247-266
45

46 Valstar, M., Gunes H. & Pantic M. 2007 How to Distinguish Posed from Spontaneous
47 Smiles using Geometric Features *Proc ICMI'07* Nagoya Aichi, Japan pp. 38-45
48

49 Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K.R. 2000 Studying the Dynamics of
50 Emotional Expression Using Synthesized Facial Muscle Movements *Journal of*
51 *Personality and Social Psychology* 78 (1) 105-119
52

53 Wilhelm, P & Schoebi, D. 2007 Assessing Mood in Daily Life *European Journal of*
54 *Psychological Assessment* 23(4):258–267
55
56
57
58
59
60

1
2
3
4
5 Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., &
6 Cowie, R. 2008 Abandoning Emotion Classes - Towards Continuous Emotion
7 Recognition with Modelling of Long-Range Dependencies *Proc INTERSPEECH*
8 *2008*

9
10 Yacoub, S. Simske S., Lin, X. & Burns, J. 2003 Recognition of emotions in
11 interactive voice response systems *Proceedings of Eurospeech 2003, Geneva* pp. 729-
12 732

13
14
15 Young, A.W., Rowland, D Calder AJ, Etcoff, NL Seth A, Perrett DI 1997 Facial
16 expression megamix: Tests of dimensional and category accounts of emotion
17 recognition *Cognition* 63, 271-313

18
19
20 Zhou G., Hansen, J. & Kaiser, J. 1999 Methods for stress classification: Nonlinear
21 TEO and linear speech based features. *Proc. IEEE International Conference on*
22 *Acoustics, Speech, and Signal Processing* 1999, vol. IV, pp. 2087-2090
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

FIGURE Captions

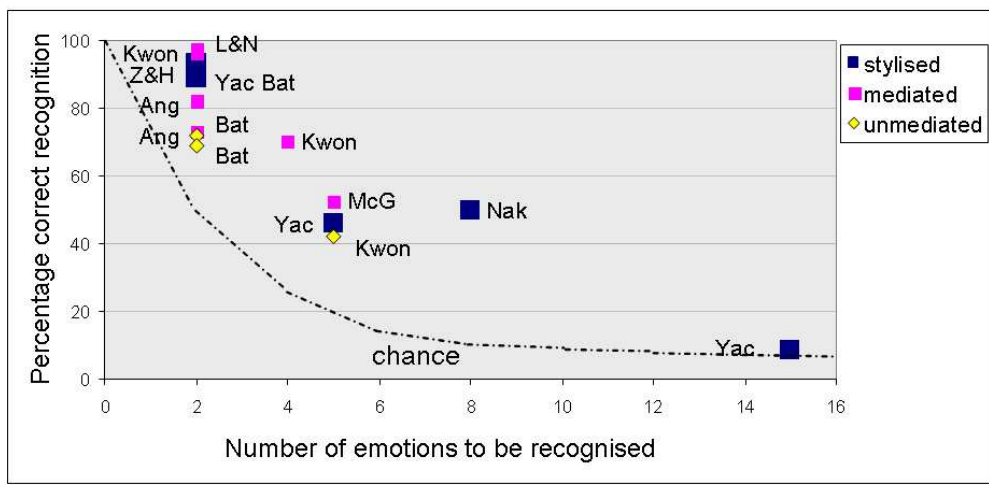
Figure 1: Plot of discrimination results from key early studies of emotion recognition – Lee and Narayan, 2003; Kwon et al., 2003; Zhou & Hansen, 1999; Ang, 2002; Yacoub et al., 2003; Batliner et al. 2003; McGilloway et al. 2000; and Nakatsu et al. 1999; against number of categories to be discriminated (horizontal axis)

Figure 2: Multimodal effects in the Belfast Naturalistic database: Each panel shows average dimensional ratings for (a) the audio-visual versions of the clips (broad light line) and (b) one of the partial versions (thin dark line). Activation ratings are in the upper panels, evaluation in the lower.

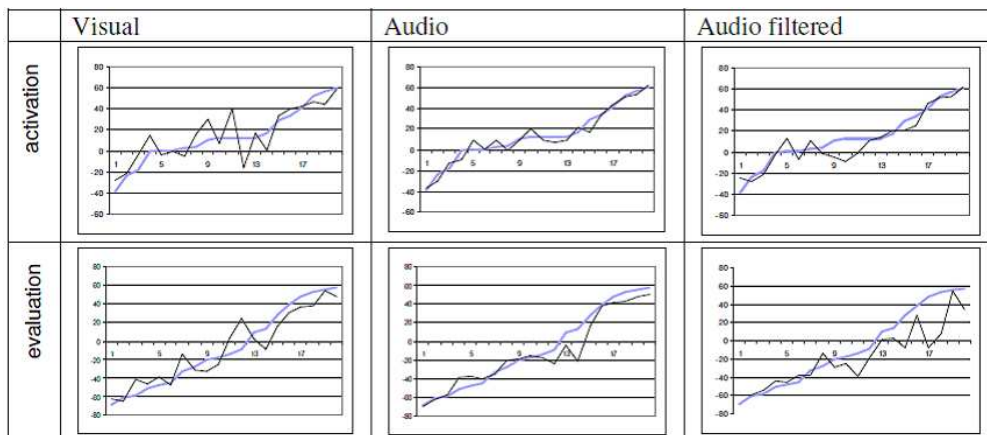
Short title for page heading:

Understanding the task of perceiving emotion

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



246x120mm (96 x 96 DPI)



228x101mm (96 x 96 DPI)

Review Only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60