PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 3515–3525
doi:10.1098/rstb.2009.0139

Perceiving emotion: towards a realistic understanding of the task

Roddy Cowie*

School of Psychology, Queen's University Belfast, Belfast BT7 1NN, UK

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 categorizing 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 and partly cognitive. Another approach is to use dimensions. Both approaches need ways to deal with gradual changes over time and mixed emotions. Attaching target descriptions to a sample poses 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.

Keywords: emotion; perception; expression; face; speech; naturalistic

1. INTRODUCTION

A great deal is known about the perception of emotion, but this 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 many different practical goals, and it is not necessarily easy for people concerned with one goal to see the relevance of research concerned with another goal.

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.

The paper is particularly aimed at highlighting the way technological research is changing the field. Most fundamentally, by engaging with people's ordinary ability to register other people's emotions, it reveals how extraordinary that ability is.

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

2. RESEARCH IN THE IMPRESSIONISTIC STYLE

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

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

*r.cowie@qub.ac.uk

One contribution of 17 to a Discussion Meeting Issue 'Computation of emotions in man and machines'.

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

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

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

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 Facial Action Coding System, known as FACS (Ekman & Friesen 1976*b*), 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-standing 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 standardized.

These analyses provided practical tools, but they also have bearing 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 (Promon *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 synthesize stimuli capable of evoking a specified perceptual response. Ekman & Friesen's (1976*a*) *Pictures of facial affect* is 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 (using Facial Action Parameters, or FAPs) was incorporated into the MPEG 4 standard (Pandzic & Forscheimer 2002). The early 1990s also saw speech synthesizers 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 allow observers to distinguish the relevant categories; but on the other hand, they are clearly not perceptually natural. The issue is taken up later.

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

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

Alongside the work with static images of faces, there was a considerable body of experimental research on the role of movement. Bassili (1978, 1979) demonstrated that category judgements could be made on the basis of facial movements rather than static

configurations. Ekman & Friesen (1982) added the influential idea that timing distinguished different types of smile. But although there was experimental support (Frank *et al.* 1993), the overall pattern of findings tended to be equivocal (Ambadar *et al.* 2005). It was demonstrated that various kinds of body movement—including dance (Dittrich *et al.* 1996), knocking movement (Pollick *et al.* 2001) and gait (Crane & Gross 2007)—can be used to classify an agent's emotional state.

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

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

Controlled pictures such as Ekman and Friesen's lend themselves to morphing, and this was exploited to study category boundaries. Two notable types of findings emerged. There is evidence of categorical perception, meaning that the perceptual effect of objectively equal differences between stimuli are perceptually small if the stimuli lie well within a category and large if they are close to a category boundary (Young *et al.* 1997). There is also evidence that boundaries are labile; Niedenthal *et al.* (2000) showed that they shift with mood.

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

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

What has been outlined above is a body of literature broadly comparable to research in other areas of perception. It has been interwoven with research on two other key themes.

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

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

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

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

4. RESEARCH IN THE TECHNOLOGICAL STYLE

Picard's (1997) book *Affective computing* signalled the arrival of research on automatic techniques for detecting emotion-related states in human beings and responding to them appropriately. It has been influenced by experimental work on the perception of emotion. But over time, technological research has increasingly been drawn to contrasting conceptions of the problem, and different kinds of solution.

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

(a) *Engagement with naturalistic material*

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

The concept of naturalistic material is not straightforward. For example, it is often opposed to 'acted material'. This can lead people to dismiss material that is taken from real conversations on the grounds

that the participants are manipulating the emotions that they show. In the worst case, they then revert to material that shows emotion as they assume it would appear if there were no such manipulation—which is usually produced by actors. Some studies avoid using actors by having random members of the public simulate emotion.

A useful way of putting it is as follows: a material is naturalistic if it is of a kind that might have to be dealt with in an application. The contrast is with idealized material, which is generated to match someone's conception of what an emotion should be like.

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

(i) *Speech*

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

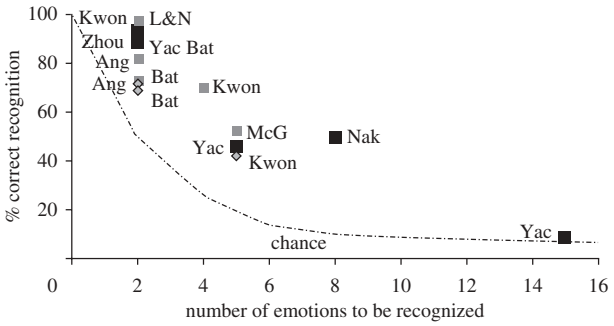

Figure 1. Plot of discrimination results from key early studies of emotion recognition—Lee & Narayanan (2003), Kwon *et al.* (2003), Zhou *et al.* (1999), Ang *et al.* (2002), Yacoub *et al.* (2003), Batliner *et al.* (2003), McGilloway *et al.* (2000) and Nakatsu *et al.* (1999)—against number of categories to be discriminated (horizontal axis). Black squares, stylized; grey squares, mediated; grey diamonds, unmediated.

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 categorizing idealized samples.

(ii) *Naturalism as a criterion in speech synthesis*

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

(iii) *Vision*

In one sense, research on facial expression addressed issues of natural and posed expression long before research on speech, because of long-standing interest in sincerity and deception. However, the spirit of the research generally reflected the impressionistic tradition. It focused on cues that a skilled observer could use to distinguish posed from spontaneous

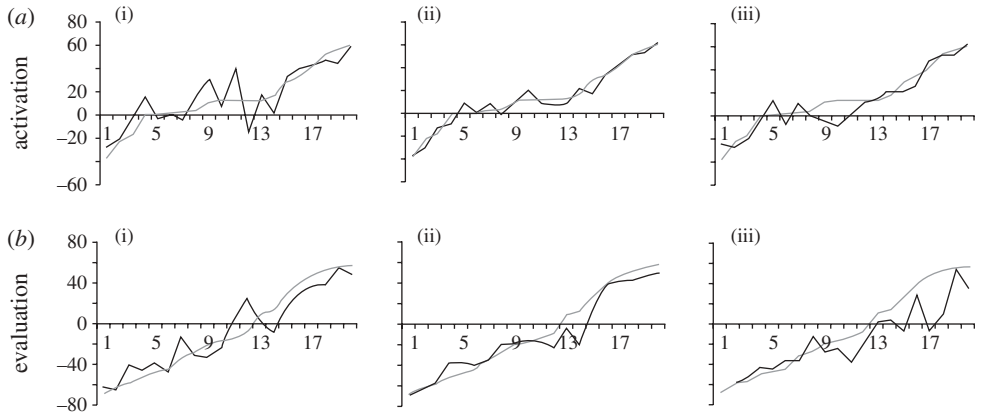

Figure 2. Multi-modal effects in the Belfast Naturalistic Database. Each panel shows average dimensional ratings for the audiovisual versions of the clips (black line) and one of the partial versions (grey dark line). (*a*) Activation ratings and (*b*) evaluation. (*a*,*b*) (i) Visual; (ii) audio; (iii) audio filtered.

smiles. The eye wrinkling associated with the famous Duchenne smile is among the most widely cited, but actually occurs frequently in posed smiles (Schmidt & Cohn 2001). There is better evidence for others, such as asymmetry (Frank *et al.* 1993) and amount of muscle movement (Hess *et al.* 1995). But although the cues exist, people tend not to use them: even children can produce posed smiles that untrained observers fail to discriminate from spontaneous smiles (Castanho & Otta 1999).

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

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

(iv) *Multi-modality*

Most of the work described above is unimodal. However, it includes some multi-modal sources, notably the Belfast Naturalistic Database. Comparing the ratings of the different modalities suggests that there are complex intermodal effects to be understood.

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 audiovisual 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 recognized 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 that difficulty (Cowie 2009) uses the term 'emotional life' to cover all those parts of human life that distinguish it from the life of a being who, like Star Trek's Mr Data, is always unemotional; and 'pervasive emotion' (following Stocker & Hegman 1992) is used to describe what is present when a person is not truly unemotional.

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

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

There is a more generic typology to be considered beyond the distinctions discussed so far. It involves distinguishing classes of phenomena like short-lived, intense emotions; moods; long-lasting 'established' emotions (such as grief or shame); stances; attitudes, and so on. These are practically important, and a competent perceiver should be able to judge whether someone is angry because of a specific event or a long-standing grievance, or is simply in a bad mood. They are also interesting because they relate to frequency of occurrence, and therefore to arguments about what is worth perceiving: moods and stances make up a very large part of emotional life and unbridled emotions rather little (Wilhelm & Schoebi 2007; Cowie 2009).

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

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

Dimensional representations are a means of addressing essentially that problem which have a long history in psychology. The simplest version, which describes emotion in terms of valence and arousal, was imported into technological research early (Cowie *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 & Cornelius 2003). Fuller dimensional schemes have emerged more recently, notably the 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 summarizing 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 judgements 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 that vary over time. The output of emotion perception can therefore be visualized as a family of 'traces' that 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.

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

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

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

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

(c) *What are the cues?*

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

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

(i) *Speech*

Two things are striking in contemporary approaches to recognizing emotion from speech. The first is the shift from acted corpora towards natural sources. The second is the number of features considered. Classical experimental papers consider small numbers of features—34 features in Banse & Scherer's (1996) study and 14 in Juslin & Laukka's (2003) review. It is commonplace for contemporary papers to consider thousands of features.

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

One is suggested by evidence that emotional passages of speech are much more likely to be rated as degraded communication than non-emotional passages from the same interactions (Cowie & Cornelius 2003). Many of the features associated with emotional speech may essentially reflect impaired control of the complex processes involved in fluent speech production, resulting breakdowns and simplifications that can take a virtually limitless variety of forms. This is consistent with the description of a moderately large feature set provided by Cowie & Douglas-Cowie (2009). The features associated with emotion depend on speaker gender, length of utterance and person judging, suggesting a rather anarchic process.

The second is that the features amount in effect to a dense picture of some underlying setting—more akin to a picture than to a list of discrete attributes. That is consistent with the observation by Schuller *et al.* (2009) that information seems to be concentrated in spectral features (in contrast with an emphasis on pitch and intensity in early research).

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

A wide range of more specific ideas is being explored. There are now systems that track speech through a space with three dimensions (valence, arousal and time), exploiting constraints on expected rate of change (Wollmer *et al.* 2008). Predictably, arousal is easier to track than activation. Voice quality is still elusive, with evidence both for and against its contribution (Schuller *et al.* 2009). Linguistically motivated descriptions of intonation have been incorporated into analysis, but seem not to enhance recognition (Batliner *et al.* 2006).

(ii) *Face*

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

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

The second shift is engaging with naturalistic data. McRorie & Sneddon (2007) compared sequences from an acted database with sequences from naturalistic recordings of strong emotions. Raters examined individual frames (in random order), and rated the emotion conveyed by each. Frame-to-frame change in rated emotion was then derived. It was much greater in the naturalistic material. Impressionistically, that

seems to be because people in the naturalistic recordings were shifting focus rapidly and expressing different reactions to different aspects of their situation. The implication is that the problem of integrating evidence over time and modalities is even greater than studies like Scherer and Ellgring's imply.

The third shift is engaging with moderate emotional colouring. Classical research found mixed evidence for timing effects, but that changes when the emotions are moderate. Ambadar *et al.* (2005) showed robust effects of movement for the identification of subtle emotions: expressions that were not identifiable in static presentations were clearly apparent in dynamic displays.

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

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

(iii) *Gesture*

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

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

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

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

(iv) *Multi-modality*

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

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

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

(v) *Context*

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

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

A second type of context effect was highlighted by Cauldwell (2000). He showed that the same speech

sample evoked reliable impressions of anger in isolation, but was judged neutral in a context that allowed listeners to attune to the speaker's habitual settings. Adaptation to speaker characteristics is a major challenge both for pure science and for technology.

(vi) *Hypotheses about mechanisms*

Initially, technological research explored various types of classification rules. Interesting lines of division have emerged gradually. There is support for rules based on decision trees, but it seems limited. Bayesian rules attract more interest, partly because they appear to model other perceptual phenomena well. There is also widespread interest in those techniques that take into account time, including Hidden Markov Models and recurrent neural nets.

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

5. HOW DO WE UNDERSTAND THE TASK?

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

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

The research leading to this paper has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

REFERENCES

Adolphs, R. 2006 Perception and emotion: how we recognize facial expressions. *Curr. Dir. Psychol. Sci.* **15** 222–226. (doi:10.1111/j.1467-8721.2006.00440.x)

Ambadar, Z. S., Schooler, J. W. & Cohn, J. 2005 Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. *Psychol. Sci.* **16**, 403–410. (doi:10.1111/j.0956-7976.2005.01548.x)

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

Bachorowski, J.-A. 1999 Vocal expression and perception of emotion. *Curr. Dir. Psychol. Sci.* **8**, 53–57. (doi:10.1111/1467-8721.00013)

Banse, R. & Scherer, K. R. 1996 Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* **70**, 614–636.

Baron-Cohen, S., Golan, O., Wheelwright, S. & Hill, J. J. 2004 *Mind reading: the interactive guide to emotions*. London, UK: Jessica Kingsley Publishers.

Bassili, J. N. 1978 Facial motion in the perception of faces and of emotional expression. *J. Exp. Psyschol.* **4**, 373–379.

Bassili, J. N. 1979 Emotion recognition: the role of facial motion and the relative importance of upper and lower areas of the face. *J. Pers. Soc. Psychol.* **37**, 2049–2059.

Batliner, A., Fischer, K., Huber, R., Spilker, J. & Noeth, E. 2003 How to find trouble in communication. *Speech Commun.* **40**, 117–143. (doi:10.1016/S0167-6393(02)00079-1)

Batliner, A. *et al.* 2006 Combining efforts for improving automatic classification of emotional user states. In *Proceedings of IS-LTC 2006, Ljubliana*, pp. 240–245.

Bhuta, T., Patrick, L. & Garnett, J. D. 2004 Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J. Voice* **18**, 299–304. (doi:10.1016/j.jvoice.2003.12.004)

Biederman, I. 1987 Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–147. (doi:10.1037/0033-295X.94.2.115)

Birdwhistell, R. L. 1970 *Kinesics and context*. Philadelphia, PA: University of Pennsylvania Press.

Burkhardt, F. & Sendlmeier, W. F. 2000 Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 151–156.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U. & Narayanan, S. 2004 Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. Sixth Int. Conf. on Multimodal Interfaces*, pp. 205–211.

Cahn, J. E. 1990 The generation of affect in synthesized speech. *J. Am. Voice I/O Soc.* **8**, 1–19.

Campbell, N. 2004 Speech and expression; the value of a longitudinal corpus. In *Proc. LREC 2004, Lisbon, Portugal*.

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

Carroll, J. M. & Russell, J. A. 1997 Facial expressions in Hollywood's portrayal of emotion. *J. Pers. Soc. Psychol.* **72**, 164–176.

Castanho, A. P. & Otta, E. 1999 Decoding spontaneous and posed smiles of children who are visually impaired and sighted. *J. Vis. Impair. Blind* **93**, 659–662.

Cauldwell, R. T. 2000 Where did the anger go? The role of context in interpreting emotion in speech. In *Speech and Emotion: Proceedings of the ISCA workshop, Newcastle, Co. Down, N. Ireland, September 2000* (eds R. Cowie, E. Douglas-Cowie & M. Schroeder), pp. 127–131.

Chakrabarti, B. & Baron-Cohen, S. 2006 Empathizing: neurocognitive developmental mechanisms and individual differences. *Prog. Brain Res.* **156**, 403–417. (doi:10.1016/S0079-6123(06)56022-4)

Conati, C. 2002 Probabilistic assessment of user's emotions in educational games. *J. Appl. Artif. Intell.* **16**, 555–575. (doi:10.1080/08839510290030390)

Cowie, R. 2005 What are people doing when they assign everyday emotion terms? *Psychol. Inq.* **16** 11–18.

Cowie, R. 2009 Describing the forms of emotional colouring that pervade everyday life. In *Oxford handbook of the philosophy of emotion* (ed. P. Goldie), pp. 63–94. Oxford, UK: Oxford University Press.

Cowie, R. & Cornelius, R. 2003 Describing the emotional states that are expressed in speech. *Speech Commun.* **40**, 5–32. (doi:10.1016/S0167-6393(02)00071-7)

Cowie, R. & Douglas-Cowie, E. 2009 Prosodic and related features that signify emotional colouring in conversational speech. In *The role of prosody in affective speech* (ed. S. Hancil), pp. 213–240. Berne, Switzerland: Peter Lang.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A. & Fellenz, W. 1999 What a neural net needs to know about emotion words. In *Computational intelligence and applications* (ed. N. Mastorakis), pp. 109–114. World Scientific Engineering Society.

Cowie, R., Douglas-Cowie, E. & Schroeder, M. (eds) 2000 *Speech and Emotion: Proceedings of the ISCA workshop, Newcastle, Co. Down, N. Ireland, 2000.*

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. 2001 Emotion recognition in human–computer interaction. *IEEE Signal Process. Mag.* **18**, 32–80. (doi:10.1109/79.911197)

Crane, E. & Gross, M. 2007 Motion capture and emotion: affect detection in whole body movement. In *Affective computing and intelligent interaction* (eds A. Paiva, R. Prada & R. Picard), *Proceedings of ACII'2007 Lisbon*. Springer Proceedings Series: Lecture Notes in Computer Science, vol. 4738.

de Gelder, B. & Vroomen, J. 2000 The perception of emotions by ear and by eye. *Cogn. Emotion* **14**, 289–311.

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

Dimberg, U., Thunberg, M. & Elmehed, K. 2000 Unconscious facial reactions to emotional facial expressions. *Psychol. Sci.* **11**, 86–89. (doi:10.1111/1467-9280.00221)

Dittrich, W. H., Troscianko, T., Lea, S. E. G. & Morgan, D. 1996 Perception of emotion from dynamic point-light displays represented in dance. *Perception* **25**, 727–738. (doi:10.1068/p250727)

Douglas-Cowie, E., Cowie, R. & Schröder, M. 2000 A new emotion database: considerations, sources and scope. In *Speech and Emotion: Proceedings of the ISCA workshop, Newcastle, Co. Down, N. Ireland, September 2000* (eds R. Cowie, E. Douglas-Cowie & M. Schroeder), pp. 9–44.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. 2003 Emotional speech: towards a new generation of databases. *Speech Commun.* **40**, 33–60. (doi:10.1016/S0167-6393(02)00070-5)

Douglas-Cowie, E., Devillers, L., Martin, J. -C., Cowie, R., Savvidou, S., Abrilian, S. & Cox, C. 2005 Multimodal databases of everyday emotion: facing up to complexity. *Interspeech 2005* 813–816.

Douglas-Cowie, E. K. *et al.* 2007 The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pp. 488–500. Berlin, Germany: Springer-Verlag.

Ekman, P. 1982 *Emotion in the human face*. New York, NY: Cambridge University Press.

Ekman, P. & Friesen, W. V. 1976*a* *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. & Friesen, W. V. 1976*b* Measuring facial movement. *J. Nonverbal Behav.* **1**, 56–75. (doi:10.1007/BF01115465)

Ekman, P. & Friesen, W. V. 1982 Felt, false and miserable smiles. *J. Nonverbal Behav.* **6**, 238–252. (doi:10.1007/BF00987191)

Ekman, P. & O'Sullivan, M. 1991 Who can catch a liar? *Am. Psychol.* **46**, 913–920. (doi:10.1037/0003-066X.46.9.913)

El Kaliouby, R. & Robinson, P. 2005 Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human–computer interaction part II* (eds B. Kisačanin, V. Pavlović & T. S. Huang), pp. 181–200. Berlin, Germany: Springer.

Fernández-Dols, J. M., Wallbott, H. & Sanchez, F. 1991 Emotion category accessibility and the decoding of emotion from facial expression and context. *J. Nonverbal Behav.* **15**, 107–123. (doi:10.1007/BF00998266)

Fontaine, J., Scherer, K., Roesch, E. & Ellsworth, P. 2007 The world of emotions is not two-dimensional. *Psychol. Sci.* **18**, 1050–1057. (doi:10.1111/j.1467-9280.2007.02024.x)

Frank, M. G., Ekman, P. & Friesen, W. V. 1993 Behavioral markers and recognizability of the smile of enjoyment. *J. Pers. Soc. Psychol.* **64**, 83–93.

Goldberg, H. 1951 The role of 'cutting' in the perception of motion pictures. *J. Appl. Psychol.* **35**, 70–71. (doi:10.1037/h0062192)

Grammer, K., Kruck, K. B. & Magnusson, M. S. 1998 The courtship dance: patterns of nonverbal synchronization in opposite-sex encounters. *J. Nonverbal Behav.* **22**, 3–29. (doi:10.1023/A:1022986608835)

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. & Wedin, L. 1980 Perceptual and acoustic correlates of voice qualities. *Acta Otolaryngol.* **90**, 441–451. (doi:10.3109/00016488009131746)

Hatfield, E. & Rapson, R. L. 2000 Emotional contagion. In *The Corsini encyclopedia of psychology and behavioral science* (eds W. E. Craighead & C. B. Nemeroff), pp. 493–495. New York, NY: John Wiley & Sons.

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

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

Heylen, D., Bevacqua, E., Tellier, M. & Pelachaud, C. 2007 Searching for prototypical facial feedback signals. *Proc. IVA*, pp. 147–153.

Juslin, P. & Laukka, P. 2003 Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* **129**, 770–814. (doi:10.1037/0033-2909.129.5.770)

Kim, J. & Andre, E. 2008 Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 2067–2083.

Kwon, O., Chan, K. & Hao, J. 2003 Emotion recognition by speech signals. *Proc. Eurospeech 2003, Geneva*, pp. 125–128.

Lacava, P. G., Golan, O., Baron-Cohen, S. & Myles, B. S. 2007 Using assistive technology to teach emotion recognition to students with asperger syndrome: a pilot study. *Remedial Special Educ.* **28**, 174–181. (doi:10.1177/07419325070280030601)

Laukka, P. 2004 Vocal expression of emotion. PhD thesis, University of Uppsala, Uppsala, Sweden.

Laukkanen, A.-M., Vilkman, E., Alku, P. & Oksanen, H. 1996 Physical variations related to stress and emotional state: a preliminary study. *J. Phonetics* **24**, 313–335. (doi:10.1006/jpho.1996.0017)

Laver, J. 1980 *The phonetic description of voice quality.* Cambridge, UK: Cambridge University Press.

Lazarus, R. J. 1999 The cognition–emotion debate: a bit of history. In *Handbook of cognition and emotion*

(eds T. Dalgleish & M. J. Power), pp. 1–19. Chichester, UK: Wiley.

Lee, C. & Narayanan, S. 2003 Emotion recognition using a data-driven fuzzy inference system. In *Proc. Eurospeech 2003, Geneva*, pp. 157–160.

Lieberman, P. & Michaels, S. B. 1962 Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *J. Acoust. Soc. Am.* **34**, 922–927. (doi:10.1121/1.1918222)

Massaro, D. W. 2004 A framework for evaluating multimodal integration by humans and a role for embodied conversational agents. In *Proc. Int. Conf. on Multimodal Interfaces 2004*, pp. 24–41.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, W. & Stroeve, S. 2000 Automatic recognition of emotion from voice: a rough benchmark. In *Proc. ISCA ITRW on Speech and Emotion*, pp. 207–212.

McRorie, M. & Sneddon, I. 2007 Real emotion is dynamic and interactive. In *Proc. Affective Computing and Intelligent Interaction 2007*, pp. 759–760. Berlin, Germany: Springer.

Milner, A. D. & Goodale, M. A. 1995 *The visual brain in action*. Oxford, UK: Oxford University Press.

Moore, R. K. 2007 Spoken language processing: piecing together the puzzle. *Speech Commun.* **49**, 418–435. (doi:10.1016/j.specom.2007.01.011)

Mozziconacci, S. 1998 Speech variability and emotion: production and perception. PhD thesis, University of Eindhoven, Eindhoven, The Netherlands.

Munn, N. 1940 The effect of knowledge of the situation upon judgment of emotion from facial expressions. *J. Abnormal Soc. Psychol.* **35**, 324–338. (doi:10.1037/h0063680)

Murray, I. R. & Arnott, J. L. 1995 Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Commun.* **16**, 369–390. (doi:10.1016/0167-6393(95)00005-9)

Nakatsu, R., Tosa, N. & Nicholson, J. 1999 Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, pp. 439–444.

Niedenthal, P. M., Halberstadt, J. B., Margolin, J. & Innes-Ker, A. 2000 Emotional state and the detection of change in facial expression of emotion. *Eur. J. Soc. Psychol.* **30**, 211–222. (doi:10.1002/(SICI)1099-0992(200003/04)30:2<211::AID-EJSP988>3.0.CO;2-3)

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S. & Ric, F. 2005 Embodiment in attitudes, social perception, and emotion. *Pers. Soc. Psychol. Rev.* **9**, 184–211. (doi:10.1207/s15327957pspr0903_1)

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 *MPEG-4 facial animation: the standard, implementation and applications* (eds I. Pandzic & R. Forscheimer), pp. 3–13. Chichester, UK: Wiley.

Picard, R. 1997 *Affective computing*. Cambridge, MA: MIT Press.

Poggi, I. 2006 *Le parole del corpo: introduzione alla comunicazione multimodale*. Rome, Italy: Carocci.

Pollick, F. E., Paterson, H. M., Bruderlin, A. & Sanford, A. J. 2001 Perceiving affect from arm movement. *Cognition* **82**, B51–B61. (doi:10.1016/S0010-0277(01) 00147-0)

Prom-on, S., Xu, Y. & Thipakorn, B. 2009 Modeling tone and intonation as target approximation. *J. Acoust. Soc. Am.* **125**, 406–424.

Russell, J. A., Bachorowski, J.-A. & Fernandez-Dols, J.-M. 2003 Facial and vocal expressions of emotion. *Annu. Rev. Psychol.* **54**, 329–49. (doi:10.1146/annurev.psych.54.101601.145102)

Scherer, K. R. 2005 What are emotions? And how can they be measured? *Social Science Information* **44**, 695–729. (doi:10.1177/0539018405058216)

Scherer, K. R. & Ellgring, H. 2007*a* Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion* **7**, 113–130. (doi:10.1037/1528-3542.7.1.113)

Scherer, K. R. & Ellgring, H. 2007*b* Multimodal expression of emotion: affect programs or componential appraisal patterns? *Emotion* **7**, 158–171. (doi:10.1037/1528-3542.7.1.158)

Schmidt, K. L. & Cohn, J. 2001 Human facial expressions as adaptations: evolutionary questions in facial expression research. *Am. J. Phys. Anthropol.* **116**(Suppl. 33), 3–24. (doi:10.1002/ajpa.20001)

Schroeder, M. 2001 Emotional speech synthesis: a review. In *Proc. Eurospeech 2001, ISCA, Bonn, Germany*, pp. 561–564.

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

Schroeder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C. & Schuller, B. 2008 Towards responsive sensitive artificial listeners. In *Fourth Int. Workshop on Human–Computer Conversation, Bellagio, October 2008*.

Schuller, B., Wöllmer, M., Eyben, F. & Rigoll, G. 2009 Prosodic, spectral or voice quality? Feature type relevance for the discrimination of emotion pairs. In *The role of prosody in affective speech* (ed. S. Hancil), pp. 285–307. Berne, Switzerland: Peter Lang.

Sengers, P., Liesendahl, R., Magar, W., Seibert, C., Müller, B., Joachims, T., Geng, W., Martensson, P. & Höök, K. 2002 The enigmatics of affect. In *Proc. DIS 2002*, pp. 87–98. New York, NY: ACM Press

Stocker, M. & Hegman, E. 1992 *Valuing emotions*. Cambridge, UK: Cambridge University Press.

Terzopoulos, D. & Waters, K. 1993 Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE PAMI* **15**, 569–579.

Tian, Y.-I., Kanade, T. & Cohn, J. 2005 Facial expression analysis. In *Handbook of face recognition* (eds S. Z. Li & A. K. Jain), pp. 247–266. Berlin, Germany: Springer.

Valstar, M., Gunes, H. & Pantic, M. 2007 How to distinguish posed from spontaneous smiles using geometric features. In *Proc. ICMI'07, Nagoya Aichi, Japan*, pp. 38–45.

Wehrle, T., Kaiser, S., Schmidt, S. & Scherer, K. R. 2000 Studying the dynamics of emotional expression using synthesized facial muscle. *Mov. J. Pers. Soc. Psychol.* **78**, 105–119. (doi:10.1037/0022-3514.78.1.105)

Wilhelm, P. & Schoebi, D. 2007 Assessing mood in daily life European. *J. Psychol. Assess.* **23**, 258–267. (doi:10.1027/1015-5759.23.4.258)

Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E. & Cowie, R. 2008 Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech 2008, Brisbane, Australia*.

Yacoub, S., Simske, S., Lin, X. & Burns, J. 2003 Recognition of emotions in interactive voice response systems. In *Proc. Eurospeech 2003, Geneva*, pp. 729–732.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A. & Perrett, D. I. 1997 Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition* **63**, 271–313. (doi:10.1016/S0010-0277(97)00003-6)

Zhou, G., Hansen, J. & Kaiser, J. 1999 Methods for stress classification: nonlinear TEO and linear speech based features. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. IV, pp. 2087–2090.