Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 389–405
Neural Networks
www.elsevier.com/locate/neunet

2005 Special Issue

Emotion recognition in human–computer interaction

N. Fragopanagos*, J.G. Taylor

Department of Mathematics, King's College, Strand, London WC2 R2LS, UK

Received 19 March 2005; accepted 23 March 2005

Abstract

In this paper, we outline the approach we have developed to construct an emotion-recognising system. It is based on guidance from psychological studies of emotion, as well as from the nature of emotion in its interaction with attention. A neural network architecture is constructed to be able to handle the fusion of different modalities (facial features, prosody and lexical content in speech). Results from the network are given and their implications discussed, as are implications for future direction for the research.

Keywords: Emotions; Emotion classification; Attention control; Sigma–pi neural networks; Feedback learning; Relaxation; Emotion data sets; Prosody; Lexical content; Face feature analysis

1. Introduction

As computers and computer-based applications become more and more sophisticated and increasingly involved in our everyday life, whether at a professional, a personal or a social level, it becomes ever more important that we are able to interact with them in a natural way, similar to the way we interact with other human agents. The most crucial feature of human interaction that grants naturalism to the process is our ability to infer the emotional states of others based on covert and/or overt signals of those emotional states. This allows us to adjust our responses and behavioural patterns accordingly, thus ensuring convergence and optimisation of the interactive process. This paper is based on the theoretical foundations of, and work carried out within, the collaborative EC project called ERMIS (for emotionally rich man–machine intelligent system), in which we have been involved recently. The aim of ERMIS is the development of a hybrid system capable of recognising people's emotions based on information from their faces and speech, both from the point of view of their prosodic and lexical content. We will develop in particular a neural network architecture and simulation demonstrating its recognition of emotions in speech and face stimuli. It will lead to open questions indicating further lines of enquiry.

The literature on emotions is rich and spans several disciplines, often with no obvious overlap or consolidating outlook. Our view of emotions has thus been shaped by the philosophy of Rene Descartes, the biological concepts of Charles Darwin and the psychological theories of William James, only to mention a few of the gurus of human sciences. Such theoretical concepts should be used as guidelines in putting together an automatic emotion recognition system (such as ERMIS) provided that they are shown to be relevant to more recent knowledge on emotions such as that stemming from the modern neuro-sciences. Indeed, recent technological advances have allowed us to probe the human brain and particularly the emotional circuitry that is involved in recognising emotions, which is yielding a more detailed understanding of the function and structure of emotion recognition in the brain. At the same time technological advances have significantly improved the signal processing techniques applied to the analysis of the physical correlates of emotions (such as the facial and vocal features) thus allowing efficient multi-modal emotion recognition interfaces to be built.

The possible applications of an interface capable of assessing human emotional states are numerous. One of the uses of such an interface is to enhance human judgement of emotion in situations where objectivity and accuracy are required. Lie detection is an obvious example of such situations, although improving on human performance

* Corresponding author.
E-mail address: nickolaos.fragopanagos@kcl.ac.uk (N. Fragopanagos).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.03.006

would require a very effective emotion recognition system. Another example is clinical studies of schizophrenia and particularly the diagnosis of flattened affect that so far relies on the psychiatrists' subjective judgement of subjects' emotionality based on various physiological clues. An automatic emotion-sensitive system could augment these judgements, so minimising the dependence of the diagnostic procedure on individual psychiatrists' perception of emotionality. More generally along those lines, automatic emotion detection and classification can be used in a wide range of psychological and neuro-physiological studies of human emotional expression that so far rely on subjects' self-report of their emotional state, which often proves problematic. In a professional environment, enriching a teleconference session with real-time information on the emotional state of the participants could provide a substitute for the reduced naturalism of the medium, so again assisting humans in their emotional discriminatory capacity.

Another use of an emotion-sensitive system could be to embed it in an automatic tutoring application. An emotion-sensitive automatic tutor can interactively adjust the content of the tutorial and the speed at which it is delivered based on whether the user finds it boring and dreary or exciting and thrilling or even unapproachable and daunting. The system could recommend a brake when signs of weariness are detected. Similarly, emotion-sensitivity can be added to automatic customer services, call centres or personal assistants, for example, to help detect frustration and avoid further irritation, with the options to pass the interaction over to a human, or even terminate it altogether. One could also imagine an emotion-responsive car that can alert the driver when it detects signs of stress or anger that could impair their driving abilities.

The most obvious commercial application of emotion-sensitive systems is the game and entertainment industry with either interactive games that offer the sensation of naturalistic human-like interaction, or pets, dolls and so on that are sensitive to the owner's mood and can respond accordingly. Finally, owing to the shared basis of human emotion recognition and emotional expression, understanding and developing automatic systems for emotion recognition can assist in generating faces and/or voices endowed with convincingly human-like emotional qualities. This can in turn lead to a fully interactive system or agent that can perceive emotion and respond emotionally. This would thereby take human–machine interaction a step closer to human–human interaction.

In the sections that follow we will briefly review some of the prominent theories of emotions and the issues that arise from them. We will then turn to the more modern theoretical advances and experimental evidence and discuss issues that arise separately on the side of the sender and on the side of the receiver. After that we will explore the nature of the emotional features from the various modalities and discuss the available data for training and testing. Finally, we will present an artificial neural network architecture for fusing emotional information from the various modalities under attentional modulation and present the results obtained in the ERMIS framework through this neural network.

2. The psychological tradition

In our effort to construct an automatic emotion recogniser, it is important to examine the ideas proposed on the nature of emotions insofar as they shape the way emotional states are described. These ideas can guide us in determining what an emotional state is and what the relevant features are which distinguish this state from others. It is also crucial to delineate the nature of the mapping of these relevant features to the state's internal representation so that effective models of this mapping can be built. Looking back at the history of emotional theories we should mention Aristotle, who classified emotions into opposites and explained the physiological and hedonic qualities associated with emotions. Later, Rene Descartes introduced the idea that a few emotions (or passions) underlie the whole of human emotional behaviour. After studying the relationship between emotions and facial expressions and bodily movements, Charles Darwin drew the conclusion that emotions are strongly linked to their survival value. He also suggested that emotions have been inherited from animal precursors. During 1880s, the American psychologist William James and the Danish physiologist Carl G. Lange independently reached the conclusion that emotions arise from perception of the physiological state after he had closely examined the peripheral components of emotions such as somatic arousal. A very extensive review of these classic theories as well as more contemporary ones can be found in Solomon (2003).

More recently, Arnold (1960) and Lazarus (1968) introduced the cognitive appraisal theory of emotions by proposing that emotions arise when a stimulus, event or situation is cognitively assessed to be carrying a personal meaning. This personal meaning is determined by personal goals and concerns and shaped by past experiences. Moreover, depending on the outcome of this cognitive appraisal an appropriate emotional response is generated. In this way appraisal theorists bring together the high-level cognitive components of emotional processing and the more low-level limbic and somatic response components that together form a complex circuitry that allows us to experience emotions even in the absence of explicit awareness of the emotion-arousing episode.

2.1. Issues arising from psychological studies

The theories of emotions mentioned so far have inspired many researchers to continue the theoretical investigation of emotions in various directions, thus making available a wide spectrum of ideas and concepts that taken together can capture the main aspects of the nature of emotions. However through such investigation, several issues of contention

have arisen that are crucial to be addressed and begun to be resolved in order to design suitable automatic emotion recognition architectures.

2.1.1. Basic emotions and their universality

Following a long tradition going back to Descartes and Darwin that supports the existence of a small, fixed number of discrete (basic) emotions, Silvan Tomkins proposed in 1962 (Tomkins, 1962) that there exist nine basic affective states (two are positive, one is neutral and six are negative), each indicated by a specific configuration of facial features. This assumption has been perpetuated by many researchers who followed (Ekman et al., 1972; Izard, 1971; Oatley & Johnson-Laird, 1987) with each researcher producing their own list of basic emotions that are different in the number and the type of basic emotions with those on the others' lists. This disparity is to say the least confusing in trying to understand the characteristics of the internal representations of the various emotional states considered to be most crucial for the development of an automatic emotion recognition system. Furthermore, while one would expect a set of basic emotions to be consistently recognised across cultures—in other words, being universal—evidence suggests that there is minimal universality at least in the recognition of emotions from facial expressions (Russell, 1994) although this view has been challenged by Ekman (1994a). Rather than engaging in this irresolvable dispute on the number and type of basic emotions, we have opted for a more flexible solution to the problem of the representations of emotional states. We do this by adopting a representation of the emotional state by a point in the continuous 2D space whose co-ordinates are the activation and evaluation involved in the emotional state. This thereby provides a continuous emotional representational map. We will discuss the nature of this map in detail later on.

2.1.2. Evolutionarily hard-wired vs. socially learned emotions

Another long-standing debate in emotion theory, which persists to date, is whether emotions are innate or learned. At one extreme, evolutionary theorists believe in the Darwinian tradition that evolution has crafted emotions in the brain as a result of a long environment-driven adaptation to better serve the behavioural imperatives of our ancestors (Ekman, 1994b; Izard, 1992; Neese, 1990; Tooby & Cosmides, 1990). Strong differentiation of emotional states within the limbic system lends some support to this approach, although such differentiation need not necessarily be genetically hard-wired or be based on discrete emotion-specific brain systems. Indeed at the other extreme, many theorists take the social constructivist approach (Averill, 1980; Ortony & Turner, 1990) which emphasises the role of higher cortical processes (such as those involved in complex social behaviour) in differentiating emotions. This camp does not accept that the strong differentiation of emotions in the limbic system is innate; but rather that it is conceivable that the limbic system contains areas that are differentially sensitive to the arousal level (activation) and to the valence (evaluation) of stimuli or events to which the subject is exposed in a non-emotion-specific way. This would allow for social influence to shape emotional responsiveness and would justify the emotional variance reported to exist across different cultural populations. At the same time, it would suggest that people from within the same social population should perceive emotion coherently. We will see later that there are indeed subtle differences between people even in a given culture as to the signals they detect from others as to the emotional states of those they are observing.

2.1.3. 'Primary' vs. 'secondary' emotions

Finally, we turn to discuss the division of emotions into two categories: 'primary' and 'secondary' emotions. What Damasio (1994) calls 'primary' emotions are the more primitive emotions such as startle-based fear, as well as innate aversions and attractions. These are said to arise automatically in the low-level limbic circuit. On the other hand, 'secondary' emotions are more subtle and sophisticated in that they require the involvement of cognitive processing to arise. These are likely to involve high-level cortical processing and even require conscious awareness. This division of emotions directly relates to the issues discussed above. Thus the primary emotions are equivalent to the basic emotions, a thesis strongly supported by the theorists who support the basic emotions, and who usually also argue that these emotions are evolutionary crafted in the limbic system. The secondary emotions would be argued, by the supporters of innate basic emotions, to be a blend of basic emotions much in the way that different colours can be created by the mixture of red, green and blue. On the other hand, the social constructivists would argue that secondary emotions are social constructs built on a set of rudimentary emotions such as startle and affinity/disgust. It is crucial to fully appreciate this division of emotions into primary and secondary since it is the secondary emotions that we are more concerned within the design of human–computer interfaces. However primary emotions, such as anger, certainly can surface in the sorts of interactions we are considering here, even in the interaction of a human with their computer!

2.2. Input- and output-specific issues

Before proceeding to review, the technical literature on the signs that indicate emotions we should raise some more general issues regarding the input to the automatic emotion recogniser, i.e. the information transmitted by the emotion sender, as well as issues relating to the output of the automatic emotion recogniser, i.e. the information captured by the emotion receiver.

2.2.1. Input-specific issues

2.2.1.1. Display rules. Evidence suggests that people learn to voluntarily inhibit spontaneous emotional expression in order to comply with culture, gender and group membership. This phenomenon was termed 'display rules' (Ekman, 1975). For instance, in most cultures expressions of anger and grief are considered unsociable and are discouraged, often being replaced by an attempted or 'fake' smile. An automatic emotion recogniser has to be able to go beyond the obvious signs of archetypical emotions to cope with these situations.

2.2.1.2. Deception. Another issue that automatic emotion recognisers have to deal with is deception. Often people deliberately misrepresent their emotional states usually in an attempt to conceal information about themselves that can be embarrassing or incriminating. This can be overcome by the use of physiological measurements such as those used in lie detectors, for instance, but this is not often an available option especially in friendly human–machine interaction. Besides, deception is part of human social interaction and does not always serve a malign purpose.

2.2.1.3. Systematic ambiguity. Often similar configurations of the features typically used to extract emotional information, such as the facial and vocal features, correspond to different behavioural states and not necessarily emotional ones. For instance, lowered eyebrows may indicate anger, but instead may indicate concentration, depending on the behavioural context. Environmental context also influences the way we alter those features to convey information that could be misleading if context is not taken into account. For instance, shouting could signify anger but it could also be a requirement of communication in a noisy environment. This systematic ambiguity of individual signs of emotions presents a challenge to any automatic emotion recogniser, and makes intelligent fusion of all available modalities and context an imperative.

2.2.2. Output-specific issues

2.2.2.1. Category labels. The output of an automatic emotion recogniser is the emotional state recognised after processing the features from the available modalities. However, an emotional state is an abstract notion which needs to be explicitly represented in order to be of use in any application. The most natural way to achieve this is by assigning a label to each emotional state from the armoury of emotional labels provided by everyday language. Restricting the choice of labels to what we earlier called basic or primary emotions, such as fear, disgust, anger, sadness, happiness and so on, would not be an effective solution to the problem as these emotions are rarely experienced in their pure form in realistic situations. Nor should we forget the lack of convergence with respect to number and type that burdens the concept of basic or primary emotions. Thus inclusion of secondary emotions is deemed essential to begin to efficiently describe the wealth of possible emotions experienced. There are two ways of achieving this inclusion. One is a straightforward extension of the list of labels to include those that describe more subtle emotions. For instance, Whissell (1989) lists 107 words describing emotional states, while Plutchik (1980) lists 142. Another way to describe secondary emotions is by mixing and matching the basic emotional labels as if in a palette of primary colours. Although the former way to deal with secondary emotions is more explicit, it places a heavy load on the automatic classification system, thereby rendering this an intractable method. In addition, the problem of deciding on a definitive list of secondary emotional labels still persists. The latter method, using a mix and match of primary emotional labels, is simply too awkward to explicitly represent emotional states, certainly so for any human inspection of the results.

2.2.2.2. Activation–evaluation space. An alternative solution to the problem of representing emotional states is using a continuous 2D space to which the emotional states are mapped. One dimension of this space corresponds to the valence of the emotional state and the other to the arousal or activation level associated with it. Cowie and colleagues (Cowie et al., 2001) have called this representation the 'activation–evaluation space'. This bipolar affective representation approach is supported in the literature (Carver, 2001; Russell & Barrett, 1999) as well being well founded through cognitive appraisal theory. An emotional state is 'valenced', i.e. is perceived to be positive or negative depending on whether the stimulus, event or situation that caused this emotional state to ensue was evaluated (appraised) by the agent of the emotional state as beneficial or detrimental. This appraisal process that assigns the positive or the negative sign to the emotional state is a key idea in cognitive appraisal theory (Ellsworth, 1994). The arousal effect of emotion on the other hand goes back as far as Darwin, who suggested that emotion predisposes us to act in certain ways. More recently from an appraisal-theoretic point of view, Frijda (1986) proposed that emotions are to be equated with action tendencies. Thus rating an emotional state on an activation scale, i.e. the strength of the drive to act as a result of that emotional state, is an appropriate complement to the valence rating. These two values together will yield a robust but flexible solution to the issue of the most appropriate emotional state representation to be used. It is also possible to relate the explicit emotional categorical labelling of emotional states to the activation–evaluation space values by representing the emotional labels themselves as points on this space. In such a translation, basic emotional labels would not map on to the activation–evaluation space uniformly. Rather they tend to form a roughly circular pattern. This is a feature which has inspired Plutchik to suggest that this may be an intrinsic structural

property of emotion. So he described emotion using an angular measure ranging from acceptance (0) to disgust (180) and from apathetic (90) to curious (270), as well as the distance from the centre, which thereby defines the strength of the emotion. More generally speaking, although the activation–evaluation space is a powerful tool to describe emotional states, there will always arise some loss of information from the collapse of the structured, high-dimensional space of the possible emotional states to a rudimentary 2D space. Moreover, different results can be obtained through the different ways of performing this collapse. However, we will not consider this issue further here, but take the activation–evaluation space as basic to our approach.

2.2.2.3. Time related categories. Aside from emotions in their narrow sense, emotional states can be related to other structures that have similar affective qualities but quite different time courses. Moods, for instance, have a longer life than emotions and can therefore affect behaviour on a larger time scale. Moreover, moods are not generated instantaneously in response to a particular object, as emotions are. Thus moods are usually experienced in a more global and diffused fashion. Nevertheless, in language the same emotional word might describe a short-lived emotion or a more protracted mood. For instance, the word 'sad' can be used to describe an emotion in response to some disappointing news but can also be used to describe the mood of a griever. Emotional traits have an even longer life as they reflect enduring inclinations to enter certain emotional states. Again the word-label 'happy' can be assigned to an emotion, a mood or a trait equally well. Thus it is clear that an automatic emotion recogniser would benefit from the use of more than one temporal scale of analysis of the signs of emotional states. In this way the emotional states recognised at each instant can be attributed to the appropriate cause (emotion, mood, trait, etc.) and mixed effects can be disentangled.

2.3. Input features—output features—training/testing data for prosody

2.3.1. Features

Speech carries a significant amount of information about the emotional state of the speaker in the form of its prosody or paralinguistic content. We take prosody to mean the way words (or non-verbal utterances) are spoken as opposed to the actual words which is the linguistic content of speech. Prosody can be quantified by the values (and the changes of those values in time) of acoustic properties such as the pitch, the amplitude or intensity and the spectral content of speech. Indeed, there has been ample research on the way such acoustic properties correlate with different emotional states. For instance, studies have shown that anger and happiness/joy are generally characterized by high mean pitch, wider pitch range, high speech rate, increases in high frequency energy, and usually increases in rate of articulation (Murray & Arnott, 1993). Sadness, as well as boredom, is characterized by decrease in mean pitch, slightly narrow pitch range, and slower speaking rate (Murray & Arnott, 1993). These are but a few of numerous empirical observations that relate the acoustic properties of speech with the emotional state of the speaker. In fact due to their empirical nature, these observations tend to have a certain degree of variance or even disparity depending on the researchers and the material used for the studies. A limitation however that seems to consistently manifest itself in most studies of speech prosody and emotion is that prosody can only really provide information about the arousal or activation level of the speaker and not so easily about the valance of their emotional state. This is evident from the examples given above where anger and joy share the same vocal characteristics while being quite opposite in valence and similarly for sadness and boredom. A very extensive examination of the relationship between prosody and emotion can be found in Cowie et al. (2001).

2.3.2. Computational studies

Several computational studies have been carried out that aim to extract automatically prosodic variables and, based on them, thence classify the emotional state of the speaker. We will only outline two of the most promising ones. The ASSESS system (Cowie et al., 2001) extracts the pitch contour, the intensity and the spectral content of speech as well as detecting the number and duration of pauses. ASSESS then performs extensive statistical calculations that result in a large number of statistical moments of various orders of those variables. The pause detection is an important operation as it allows for meaningful segments of speech—into so-called 'tunes'—to be defined as the portion of speech lying between two pauses. These tunes can then provide a unit of analysis for the study of the fluctuations of the extracted vocal variables. Thus, the output of ASSESS is a large set of statistical moments of the basic features extracted by means of signal processing and can be delivered on a tune-by-tune basis or on a larger chunk basis. The ASSESS output can then be used as input to any statistical analysis tool or neural network.

ASSESS was the pre-processor/feature analyser actually used for ERMIS, with the resulting output fed to a brain-inspired neural network (ANNA) which we will introduce in a later section, as well as later present results and discuss their interesting features. Another important computational study was carried out by Scherer's group (Banse & Scherer, 1996) who also used the basic speech features (fundamental frequency, energy, speech rate and spectral measures) as well as some statistical moments of those features, and then performed discriminant analysis to match speech samples to specific basic emotional states. Classification was of the order of 50% correct, which is about the same order of classification as achieved by human judges.

2.3.3. Neural sites

Turning to the neural sites that appear to be involved in the recognition of emotions from prosody we are confronted with a wealth of findings, mostly from lesion studies, that indicate that these neural sites are distributed between the left and right hemispheres, with an emphasis on structures in the right hemisphere, and in particular the right inferior frontal regions. The latter sites, together with more posterior regions in the right hemisphere and the left frontal regions, as well as subcortical structures, support emotion recognition based on the various auditory features. A review of recent studies that support the above neural allocation of prosodic processing can be found in Adolphs et al. (2002). In the same study the authors tested 66 patients with focal brain damage against 14 control subjects, in their ability to recognise emotion from prosody. The main conclusions drawn, based on their findings, were that firstly, primary and high level auditory cortices are involved in the extraction and perceptual processing of the various prosodic cues. Secondly, the amygdala and the orbital and polar frontal cortices appear to be responsible for translating these prosodic cues into emotional information regarding the speech source. Thirdly, following emotion categorisation, a simulation of the associated bodily response ensues, mediated by motor and premotor structures. Finally, somatosensory structures are responsible for transforming the simulated bodily responses back to somatic and sensory representations thus giving the sensation of the emotion conveyed by the speech source.

2.4. Faces

2.4.1. Features

Facial expression is a fundamental carrier of emotional information and is used widely in all cultures and civilisations to express as well as perceive emotion. In an automatic emotion recognition system, the task is to extract those features from a face that are most indicative of a person's emotional state. These include, but are not restricted to the eyes, the eye-brows, the mouth and the nose. The extracted features can be analysed statically (usually at the extreme of a sequence of facial movements) or dynamically by monitoring and measuring the variation of these features in time. Much of the recent research in the field focuses on the latter. Most of the techniques developed to analyse facial expressions are based on the seminal work by Ekman and Friesen (1978) who developed the so-called the Facial Action Coding System (FACS). FACS is predicated on human facial anatomy and introduces the concept of 'action units' (AUs) as causes of facial movement. An AU is an assembly of several facial muscles that generate a certain facial action by their movement. The mapping between action units and muscle units is not one-to-one, as the same muscles can elicit a series of action units (or effectively of facial expressions). The work by Ekman and Friesen was original based on the assumption that there exists a set of basic emotions, so most research in extracting emotional information from the faces has also been based on this assumption, often with very high classification performance. However, we have already discussed in previous sections how the notion of basic emotions can be limiting, as it does not allow for the full spectrum of emotions to be represented and consequently used for realistic emotion classification beyond the archetypical emotions. In particular the social emotions do not so easily enter this approach.

2.4.2. Computational studies

The first task to be performed in any facial expression analysis approach is face detection or tracking. This can be based on colour, used as a clue to disentangle the face from its background, or on a set of given templates in the form of gray values, active contours, graphs or wavelets. Pose estimation is also an important aspect of face detection or tracking, as different viewing angles cause substantial change in the appearance of the face. Following face detection or tracking, two major approaches exist for the mapping of facial features to emotions. One is the target-oriented approach, whereby facial expression analysis is conducted statically at the apex or extremity of the expression (mug-shot) with the aim to detect the presence of static cues such as wrinkles as well as the positions and shapes of facial features. This approach has proved cumbersome, so that few techniques based on this approach have been successful. The other approach to facial expression recognition is gesture-oriented. This approach requires a set of successive frames of a facial expression, as it involves measurements of image properties on a frame-by-frame basis so that gradients and variances can be extracted. One of the gesture-oriented approaches is based on optical flow, whereby dense motion fields are computed in selected areas of the face and then mapped to facial emotions by means of motion templates extracted as sums over a set of test motion fields. Another gesture-oriented approach is based on feature tracking whereby prominent features such as edges, corner-like and high-level patterns (e.g. eyes, brows, nose and mouth) are firstly extracted for each frame of a sequence, followed by motion analysis by means of relating the features from one frame to another. A third gesture-oriented approach aligns a 3D model of the face and head to the image data so as to extract object motion as well as orientation. Our ERMIS partner, the National Technical University of Athens (NTUA), used a feature tracking approach which is based on the extraction of the MPEG-4 compliant Facial Definition Parameters (FDPs) which are in turn used to calculate the Facial Animation Parameters (FAPs). We will show in a later section results obtained when the latter are used as inputs to ANNA for emotion classification in ERMIS.

2.4.3. Neural sites

The first stage of processing emotional faces is a feed-forward sweep through primary and higher level visual

cortices ending up in associative cortices (temporal) and more specifically in the fusiform gyrus known to be especially active during the processing of faces. Even at the early phases of this sweep, a crude categorisation of the faces into emotional or not can ensue, based on the structural properties of the faces. In the event of an emotional face being detected, projections at various levels of the visual and the associative cortices to the amygdala can alert the latter of the onset of an emotional face. Some scientists believe that even a subcortical route, presumably via the superior colliculus, can carry similar crude emotional information very fast up to the amygdala. The issue of how early and from how low-level sources the amygdala can be activated is still open to debate. Nevertheless, once the amygdala is alerted to the onset of an emotional face it can in turn draw attention back to that face so as to delineate the categorisation of the face in detail. It can do so, as mentioned in the either by employing its direct projections back to the visual and associative cortices to enhance processing of this face or via activation of the orbitofrontal cortex that can, in turn, initiate a re-prioritisation of the salience of this face within the prefrontal goal areas and drive attention back to the face. Finally, the amygdala can also generate or simulate a bodily response through its connections to motor cortices, hypothalamus and brainstem nuclei which can then be transformed back to somatosensory representations providing effectively a simulation of the other person's emotional state. This is a possible mechanism by which we experience empathy. For a more extensive review of the neural correlates of emotion recognition from faces see Adolphs (2002).

It may alternatively be that the amygdala is the basic repository of the emotional characteristic of face (and other) inputs elicited at a pre-attentive stage. Cortical representations may be more neutral, whereas the amygdala representation of a stimulus provides a low level but crucial first component of the emotional character of a stimulus, so as to cause attention to be drawn to the stimulus (as indicated above) for further analysis. The early cortical analysis, on this view, would consist of structural features extracted from the stimulus input, which would alert the coarse emotional code carried by the amygdale for that input. More detailed analysis could be done later by the orbitofrontal cortex, so as to allow more deliberated decisions as to an appropriate response to be made.

2.5. Words

Extracting emotional information from the lexical content or meaning of the words used by a speaker, using an automatic system, is not a trivial task. Often some of the words we use carry a strong and clear emotional charge. More often, we convey emotionality on a higher level through socially learned semantic schemata. We will first review the existing approaches to automatic emotional content extraction from words and then proceed to present the relevant system we developed for ERMIS. So the existing approaches can be classified into four classes:

(1) Keyword spotting;
(2) Lexical affinity;
(3) Statistical natural language processing;
(4) Hand-crafted models.

We will not consider the fourth class here since it is not very general, and only concentrate on the first three.

2.5.1. Keyword spotting

This is the most direct approach, involving a simple look-up table from a given lexical entry to an emotional value, be it of one or a number of categories of emotions or of emotional dimensions. Thus, Ortony's Affective Lexicon provides an often-used source of affective words, grouped into affective categories. Similarly, Elliot's Affective Reasoner watches for 198 affect key words like 'distressed' or 'enraged', plus affect intensity modifiers, like 'extremely' or 'somewhat'. The Whissell dictionary transforms any word into a pair of values, one for activation, the other for evaluation. This can thereby give a finer descriptor than can the Affective lexicon, which can only produce emotional categories. The weakness of all of these systems is that they have poor recognition of affect when there is negation, and only surface features are involved. Thus a lot of sentences convey underlying meaning rather than a surface one. Thus the sentence 'My husband filed for divorce and he wants to take custody of the children' has very strong emotional content, which is not detected at the surface level by the affective transform systems.

2.5.2. Lexical affinity

This approach is an extension of the keyword spotting technique in the sense that apart from picking up the obvious emotional keywords, it assigns a probabilistic 'affinity' for a particular emotion to arbitrary words. These probabilities are often part of linguistic corpora, which gives this approach one of its disadvantages, namely, that the assigned probabilities are biased toward corpus-specific genre of texts. Another disadvantage of this approach is that it misses out on emotional content that resides deeper than the word-level on which this technique operates. For example, the word 'accident', having been assigned a high probability of indicating a negative emotion, would not contribute correctly to the emotional assessment of phrases like 'I avoided an accident' or 'I met my girlfriend by accident'.

2.5.3. Statistical natural language processing

Statistical natural language processing approaches operate by combining features of keyword spotting and lexical affinity techniques not only analysed on a word-by-word basis but on the basis of the statistics of these features taken over some significant portion of text. Various methods, such

as latent semantic analysis (LSA), have been used for the emotional assessment of texts but they all suffer from an important drawback; namely, the unit of text over which statistical analysis is run needs to be rather large to be effective (in text terms, paragraph-level and above), rendering this methodology inappropriate for use on a sentence-level and below. Nevertheless, given the restrictions for use, statistical natural language processing yields good results for affect classification as it can importantly detect some underlying semantic context in the text.

2.5.4. Conclusions on existing approaches

Overall the various methods all suffer from incorrect description of deeper emotional content since they are only able to capture the surface emotional content, be that for a long sequence of text or not. But we consider that the word-spotting method, possibly enhanced by further context-driven assistance, can provide a useful component in a multi-modal approach to emotion recognition. Such multi-modality is not considered in the analyses above. We will present results of such a multi-modal analysis, using ANNA, shortly.

2.5.5. Text post-processing module for ERMIS

Various parameterisations and coordinates systems have been proposed in the literature for the quantification of the emotional content of words. As noted above, we have adopted the 2D emotional space of activation and evaluation. Activation values indicate how dynamic an emotional state is, i.e. how active or passive it makes the person in that state. Evaluation (or pleasantness) values designate the 'sign' of an emotional state; in other words, how positive or negative the person feels. This set of parameters is a good global descriptor of an emotional state and has been used by Prof. Whissell of the Laurentian University to compile the 'Dictionary of Affect in Language'. This dictionary comprises around 9000 words and their corresponding activation and evaluation values as rated by students at her university. The dictionary also includes the values of a parameter called 'imagery' which corresponds to how strong an image an emotional state ensues, but we will not be concerned with this. We note that statistical analysis of the dictionary terms has confirmed statistical independence of all three parameters (insignificant cross-correlations).

The speech recognition engine that precedes the text post-processing module converts the speech audio signal to text. Next, the words in the extracted text are passed on to the text post-processing module where they are mapped to the 2D activation–evaluation space, thus, forming a trajectory that corresponds to the movement of emotion in the speech stream. An example of such trajectories is demonstrated in Fig. 1, where the emotionally contradictory phrases 'I am very happy' and 'I am very angry' are shown next to each other.

The most crucial feature of the text post-processing module is the window over which the two values for each

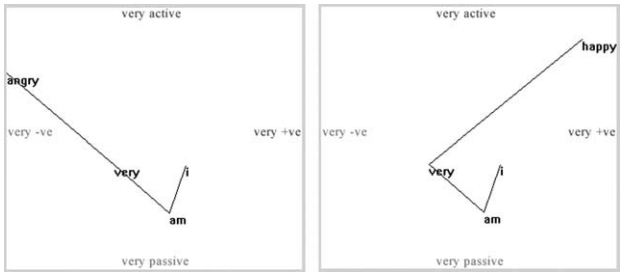

Fig. 1. Emotional trajectories of emotionally contradictory phrases.

word is calculated. These results will then be used as features to be fused with the other features produced by the prosodic and the facial analysis and, finally, give as output the best estimate of the emotional state. Further tests need to be done to define the length and type of this window, i.e. how many words will be contained and by which criteria these sets of words will be chosen. This is expected to affect crucially the performance of the text post-processing module. In what follows we will refer to the activation–evaluation values produced by the text post-processing module as DAL values after Whissell's emotional lexicon which is the core of the text post-processing module.

3. Training and testing material

An automatic emotion recognition system that employs learning architectures (e.g. neural networks), such as the one developed for ERMIS, requires sufficient training and testing material. This material should contain two streams: an input stream and an output stream. The input stream would comprise the extracted relevant features from the various modalities (prosody, faces, words, etc.) and the output would comprise the emotional class or category or more generally the emotional representation of the episode for which the input features were extracted. By episode we mean any piece of text, audio, video or multimodal combinations of these that was used for analysis.

There exist many databases of such episodes from different scenarios and contexts developed in academia or industry for the purposes of emotional classification studies. An overview of these databases can be found in Cowie et al. (2001), while a more up-to-date and detailed report on these databases can be found in deliverable D5c of the European Network of Excellence HUMAINE at: http://emotion-research.net/deliverables/D5c.pdf. We will only mention a couple of points with respect to those databases and then proceed to discuss the database that was developed for ERMIS.

Most of the existing databases of uni or multimodal emotional material have been generated by actors acting out one emotion or another. It is empirically apparent that acted emotions are quite different from natural emotions as they measure significantly differently on the use of various

analytical tools (as well as by human judgement). Furthermore, actors tend to over-act the emotion they are supposed to portray, thus restricting the available material to 'full-blown' emotions. Therefore, although the use of actors to generate databases of emotional material appears to be an easy solution, and thus obviously attractive, one should be very cautious about using such material and not assume that the results obtained with acted emotion would also apply with accuracy to natural emotional expression.

Another limitation of the existing databases is that the output stream mentioned above is often missing or expressed in the space of basic emotions. We have already stressed how limiting is the space of basic emotions for describing the full range of emotional expressions. Therefore, it is crucial that the 'ground truth' for an emotional episode, that is to be used for training or testing any emotion recognition system, should be described in a flexible and comprehensive way, such as the activation–evaluation space described above.

For ERMIS there was developed a scenario, called SALAS, where emotions would be naturally elicited by users while being recorded on audio and video tape, and then rated by humans on the activation and evaluation space. The SALAS scenario is a development of the ELIZA concept introduced by Weizenbaum (1966). The user communicates with a system, whose responses give the impression of sympathetic understanding, thereby allowing a sustained interaction to build up various emotional states in the user. The system takes no account of the user's meaning, but simply picks from a repertoire of stock responses on the basis of surface cues extracted from the user's contributions. In the case of SALAS, the surface cues involve emotional tone. In a 'Wizard of Oz' type arrangement, a human operator identifies the emotional tone, and uses it to select the sub-repertoire from which the system's response is to be drawn. A second constraint on the selection is that the user selects one of four 'artificial listeners' for the human subject to interact with at any given time. Each listener will try to direct the user towards a particular emotional state—'Spike' will try to make the user angry, 'Pippa' will try to make him or her happy, etc.

SALAS took its present form as a result of a good deal of pilot work. In its present form, it provides a framework within which users express a considerable range of emotions in ways that are virtually unconstrained. The process depends on users' co-operation—they are told that it is like an 'emotional gym', and they have to use the machinery it provides to exercise their emotions. But if they do enter into the spirit, they can move through a very considerable emotional range in a recording session or a series of recording sessions: the 'artificial listeners' are designed to let them do exactly that.

After obtaining the input stream of the training/testing material in the form of audio–visual recordings of several SALAS sessions, an output stream is generated by having other human subjects assess the emotional content of the input material. This is achieved through the use of a program called FEELTRACE which tracks in real-time the movement of a pointer (driven by a computer mouse) around a 2D activation–evaluation space projected on a computer monitor while subjects view the material (of the user reacting emotionally with the SALAS program) to be assessed. We thus obtain a stream of (x, y)-coordinate values in the activation–evaluation space that is synchronised with the input stream and can serve as the 'ground truth' or 'supervisor', in neural network terms, during the training of our learning system. Taken together the two streams—the input (speech/prosody/lexical content), and the output (as the FEELTRACE trajectories made by a set of human assessors)—comprise a fully usable training/testing data collection.

4. Extracting artificial neural network architectures: ANNA

4.1. The architecture

One of the most important effects of emotion is their ability to capture attention whether it is 'bottom-up' attention directed to stimuli or events that have been automatically registered as emotional, or it is 'top-down' attention re-engaged to a stimulus or event that has been evaluated as important to the current needs and goals after a cognitive appraisal mediated by a complex emotional–cognitive network. This emotion–attention interaction has been extensively discussed in the previous paper, and forms the theoretical basis for the artificial neural network architecture we have developed for ERMIS. The input of this neural network comprises features, from various modalities, which correlate with the user's emotional state. These are fed to a hidden layer, denoted EMOT, as shown in Fig. 1, and representing the emotional state of the user as being extracted from the input message. The output is a label of this state (FEEL). Attention acts (as well supported by experiment) as a feedback modulation on the feature inputs so as to amplify or inhibit the various feature inputs according to whether they are or are not useful for the emotional state detection. The architecture is thus based on a purely a standard feed-forward neural network, but with the addition of a feedback layer (denoted IMC in Fig. 1) modulating the activity in the inputs to the hidden layer (EMOT). IMC denotes 'inverse model controller', a concept taken from engineering control theory and transported into attention (Taylor, 2003). As such the IMC is the generator of a control signal to re-orient attention to the emotionally impelling input stimulus. The overall information flow of the resulting network, ANNA (=artificial neural network for attention and emotion) is shown in Fig. 2.

More general architectures for the interaction of attention and emotion can be defined in place of that of Fig. 2, being based on a more complete decomposition of the modules in

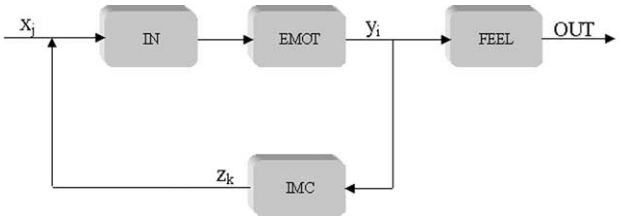

Fig. 2. Information flow in ANNA. IMC, inverse model controller; FEEL, output emotional state classification; EMOT, emotional (hidden) state; IN, stimulus input in one or several modalities (speech, facial contours, body orientation). The feedback from the IMC acts as a modulation on the input to the EMOT hidden layer.

the brain into working memory and monitor sites, as well as prefrontal goal sites (Fragopanagos, Kockelkoern, & Taylor, accepted; Taylor, 2003) (as also described in the previous paper in relation to the simulation of the emotional–attentional blink). This uses the more general CODAM model (Taylor, 2003), which contains such essential brain modules as to allow report to take place (and hence consciousness), as well as helping to specify computational mechanisms for the scarce resource character of attention. This expansion of the attention control circuitry, beyond that used in Fig. 2, is necessary, it is claimed, to enable the beginning of simulations of conscious emotional experience. For only by such an extension would it be possible to incorporate activity related to the experience of 'what it is like to be in an emotional state'. However, the data available for ANNA only allowed us to take somewhat simple neural architectures, with relatively few connection weights. Thus we had to leave out of Fig. 2 these further CODAM-like components.

4.2. The learning rules

The equations that govern the responses of the neurons in the various layers of ANNA are as follows.

Assuming linear output:

$$\mathrm{OUT} = \sum_i a_i y_i \tag{1}$$

Hidden layer (EMOT) response:

$$y_i = f\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} z_k\right]\right) \tag{2}$$

Feedback layer (IMC) response:

$$z_k = f\left(\sum_i B_{ki} y_i\right) \tag{3}$$

Fig. 3 illustrates the connectivity, neuron activity, and weight assignments in ANNA as described in Eqs. (1)–(3).

Before starting training ANNA, we need to have obtained for each input x_j the set of asymptotically converged y_i's and z_k's by running the network until these values have converged to a suitable value. So if we combine

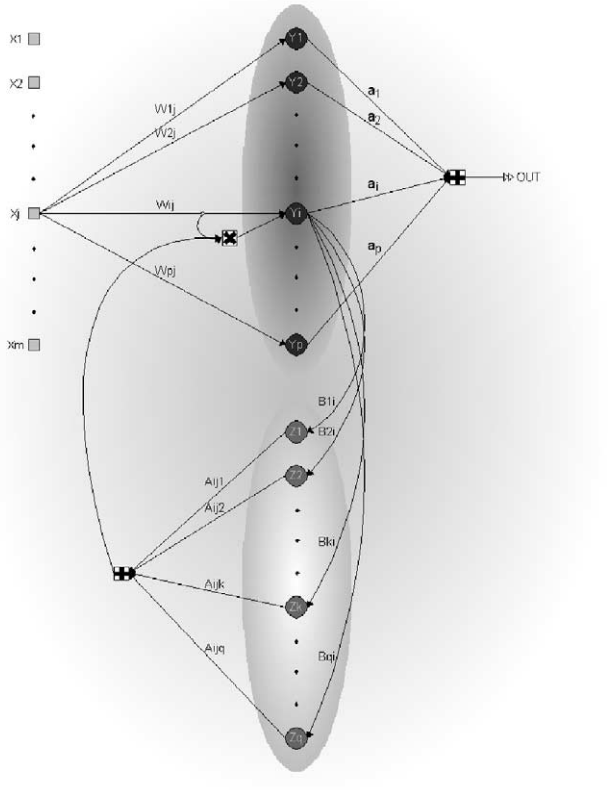

Fig. 3. The connectivity of ANNA (note gain modulation of inputs to EMOT layer by feedback from the IMC layer).

the ANNA Equations (2) and (3) in one iterative equation for the variables y_i, we get the equations:

$$y_i(t+1) = f\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} f\left(\sum_l B_{kl} y_l(t)\right)\right]\right) \tag{4}$$

We note that the feedback in Eq. (4) does not involve an additive threshold modulation, as is popular in many simulations of attention. Instead, it uses a product modulation, leading to a quadratic sigma–pi form of activation. This modulation has strong experimental support, as summarised and analysed in terms of its possible cholinergic basis, in Hartley et al. (2005).

We iterate (4) $\forall i$ until some global convergence criterion (e.g. $\sum_i |y_i(t+1) - y_i(t)|^2 < \delta$) is satisfied. This way we obtain the $y_i(\infty)$'s and $z_k(\infty)$'s.

Then we can train by gradient descent on the error surface $E = \frac{1}{2}\sum_m [\mathrm{OUT}^{(m)} - t^{(m)}]^2$, where t denotes the target output, based on the training set $\{x_j^{(m)}, \mathrm{OUT}^{(m)}\}_{j,m}$ with m denoting the pattern.

The training rules are:

$$\Delta\vartheta = -\varepsilon\frac{\partial E}{\partial\vartheta}, \quad \text{for } \vartheta = \{a_i, \omega_{ij}, A_{ijk}, B_{kl}\} \tag{5}$$

where if we define $\delta^{(m)} = \mathrm{OUT}^{(m)} - t^{(m)}$ and

$$f'_{y_i(\infty)} = f'\left(\sum_j \omega_{ij}x_j\left[1 + \sum_k A_{ijk}z_k(\infty)\right]\right) \quad \text{and}$$

$$f'_{z_k(\infty)} = f'\left(\sum_i B_{ki}y_i(\infty)\right)$$

$$\frac{\partial E}{\partial a_i} = \sum_m \delta^{(m)}y_i^{(m)} \tag{6}$$

$$\frac{\partial E}{\partial \omega_{kl}} = \sum_m \delta^{(m)}\sum_i \frac{\partial y_i^{(m)}}{\partial \omega_{kl}}$$

$$= \sum_m \delta^{(m)}\sum_i (L^{-1})_{ik} f'_{y_k^{(m)}(\infty)}\left[1 + \sum_n A_{k\,\ln} z_n^{(m)}(\infty)\right]x_l^{(m)} \tag{7}$$

$$\frac{\partial E}{\partial A_{k\,\ln}} = \sum_m \delta^{(m)}\sum_i \frac{\partial y_i^{(m)}}{\partial A_{k\,\ln}}$$

$$= \sum_m \delta^{(m)}\sum_i (L^{-1})_{ik} f'_{y_k^{(m)}(\infty)}\omega_{kl}x_l^{(m)}z_n^{(m)}(\infty) \tag{8}$$

$$\frac{\partial E}{\partial B_{kl}} = \sum_m \delta^{(m)}\sum_i \frac{\partial y_i^{(m)}}{\partial B_{kl}}$$

$$= \sum_m \delta^{(m)}\sum_s (L^{-1})_{is} f'_{y_s^{(m)}(\infty)}\sum_j \omega_{sj}x_j f'_{z_k^{(m)}(\infty)}A_{sjk}y^{(m)}(\infty) \tag{9}$$

where

$$L_{ir} = \delta_{ir} - f'_{y_i^{(m)}(\infty)}\sum_j \omega_{ij}x_j\sum_k f'_{z^{(m)}(\infty)}A_{ijk}B_{kr} \tag{10}$$

The derivation of the partial derivatives of Eqs. (7)–(9) with respect to the weights can be found in Appendix A.

In the ERMIS framework, ANNA is being used to fuse features that can predict the user's emotional state while belonging to different modalities such as linguistic, paralinguistic (prosodic) and facial. It is precisely this sort of application where one has an input vector comprising highly uncorrelated and diverse features that ANNA was designed for. These features need to be appropriately amplified or attenuated to reveal an optimal strongly predictive subset of the original set. Therefore, it is not only the ERMIS framework for which ANNA is suitable, but rather any application requiring similar cross-modal integration would benefit immensely from the unique attentional feedback feature that characterises ANNA. This leads to an adaptive recurrent 'sigma–pi' ANN, whose training laws are given in detail in Appendix A. ANNA is relevant to any situation in which attention feedback could improve processing, such as in complex scenes or with complex inputs (as in auditory channels, for example).

5. Results

5.1. General features of the analysis

A selection of SALAS sessions was analysed by the respective ERMIS partners who extracted the relevant features from the voice, faces and word stream. These sessions were also evaluated as to their emotional content by four subjects using the FEELTRACE program. The resulting streams of input and output data were in turn analysed by use of ANNA. The results are shown in Table 1, which gives the full set of ASSESS–FAPs–DAL training results, as well as the testing results.

To explain what has been done to assess the value of the different modalities towards emotion recognition (text emotion content through DAL, prosody through ASSESS feature analysis, and facial characteristics through the 17 FAP values), we note that there are seven combinations of the three sets of input features: each one of the three input modalities separately, the three pairs of modalities used together, or all three input feature sets, across all modalities, used simultaneously. These seven possible combinations of modality inputs are shown successively going down the rows of Table 1. There are four separate data entries, one for each of the four viewers used to give a running estimate by FEELTRACE of the emotional state of the particular subject being observed in the database; the viewers are denoted by their initials (JD, DR, EM, CC). There has also been a choice made of output to be used, as either denoted by 'A' for activation value, 'E' for evaluation value, or 'A+E' to denote the full 2D activation–evaluation co-ordinates arising from the FEELTRACE assessment by each of the viewers. We also note that, assuming random responses, ANNA would achieve a 50% success rate when considering the positive or negative values of the A or E 1D outputs; the corresponding random level for quadrant matching in the A+E case will be expected to be 25%. As we can see, the results imply considerably better success rates than being purely random. We discuss the results in more detail in the next sub-section.

5.2. Specific features of the training process

Before that, we give some details of the training process itself. More specifically, the results were obtained from ANNA on 500 training epochs, three runs for each dataset, the final results being averaged (with associated standard deviation calculated). In all training sessions five hidden-layer (EMOT) neurons and five feedback-layer

Table 1
The table is divided into three parts side-by-side

INPUT SPACE	OUT	FT	AVG	SD	INPUT SPACE	OUT	FT	AVG	SD	INPUT SPACE	OUT	FT	AVG	SD
ASSESS	A+E	JD	0.48	0.04	ASSESS	A	JD	0.82	0.02	ASSESS	E	JD	0.56	0.04
ASSESS	A+E	DR	0.37	0.05	ASSESS	A	DR	0.62	0.03	ASSESS	E	DR	0.64	0.03
ASSESS	A+E	EM	0.61	0.06	ASSESS	A	EM	0.86	0.01	ASSESS	E	EM	0.59	0.03
ASSESS	A+E	CC	0.48	0.04	ASSESS	A	CC	0.64	0.05	ASSESS	E	CC	0.77	0.02
FAPs	A+E	JD	0.52	0.02	FAPs	A	JD	0.82	0.03	FAPs	E	JD	0.64	0.04
FAPs	A+E	DR	0.31	0.02	FAPs	A	DR	0.63	0.04	FAPs	E	DR	0.57	0.05
FAPs	A+E	EM	0.60	0.01	FAPs	A	EM	0.88	0.02	FAPs	E	EM	0.59	0.03
FAPs	A+E	CC	0.37	0.04	FAPs	A	CC	0.54	0.02	FAPs	E	CC	0.59	0.03
DAL	A+E	JD	0.51	0.04	DAL	A	JD	0.85	0.03	DAL	E	JD	0.64	0.03
DAL	A+E	DR	0.39	0.02	DAL	A	DR	0.66	0.01	DAL	E	DR	0.62	0.02
DAL	A+E	EM	0.58	0.03	DAL	A	EM	0.98	0.01	DAL	E	EM	0.72	0.01
DAL	A+E	CC	0.32	0.01	DAL	A	CC	0.75	0.08	DAL	E	CC	0.60	0.08
ASSESS+FAPs	A+E	JD	0.51	0.01	ASSESS+FAPs	A	JD	0.78	0.00	ASSESS+FAPs	E	JD	0.69	0.03
ASSESS+FAPs	A+E	DR	0.44	0.04	ASSESS+FAPs	A	DR	0.60	0.02	ASSESS+FAPs	E	DR	0.65	0.03
ASSESS+FAPs	A+E	EM	0.65	0.05	ASSESS+FAPs	A	EM	0.89	0.05	ASSESS+FAPs	E	EM	0.69	0.01
ASSESS+FAPs	A+E	CC	0.35	0.04	ASSESS+FAPs	A	CC	0.63	0.05	ASSESS+FAPs	E	CC	0.60	0.04
DAL+FAPs	A+E	JD	0.51	0.04	DAL+FAPs	A	JD	0.86	0.03	DAL+FAPs	E	JD	0.56	0.09
DAL+FAPs	A+E	DR	0.35	0.04	DAL+FAPs	A	DR	0.53	0.02	DAL+FAPs	E	DR	0.67	0.03
DAL+FAPs	A+E	EM	0.71	0.04	DAL+FAPs	A	EM	0.95	0.02	DAL+FAPs	E	EM	0.66	0.05
DAL+FAPs	A+E	CC	0.33	0.07	DAL+FAPs	A	CC	0.73	0.07	DAL+FAPs	E	CC	0.60	0.04
ASSESS+DAL	A+E	JD	0.57	0.01	ASSESS+DAL	A	JD	0.90	0.03	ASSESS+DAL	E	JD	0.67	0.03
ASSESS+DAL	A+E	DR	0.37	0.02	ASSESS+DAL	A	DR	0.61	0.02	ASSESS+DAL	E	DR	0.65	0.05
ASSESS+DAL	A+E	EM	0.61	0.07	ASSESS+DAL	A	EM	0.98	0.01	ASSESS+DAL	E	EM	0.77	0.02
ASSESS+DAL	A+E	CC	0.49	0.07	ASSESS+DAL	A	CC	0.73	0.02	ASSESS+DAL	E	CC	0.66	0.01
ASSESS+FAPs+DAL	A+E	JD	0.47	0.07	ASSESS+FAPs+DAL	A	JD	0.87	0.03	ASSESS+FAPs+DAL	E	JD	0.61	0.01
ASSESS+FAPs+DAL	A+E	DR	0.41	0.01	ASSESS+FAPs+DAL	A	DR	0.63	0.05	ASSESS+FAPs+DAL	E	DR	0.67	0.04
ASSESS+FAPs+DAL	A+E	EM	0.67	0.01	ASSESS+FAPs+DAL	A	EM	0.98	0.02	ASSESS+FAPs+DAL	E	EM	0.71	0.01
ASSESS+FAPs+DAL	A+E	CC	0.43	0.01	ASSESS+FAPs+DAL	A	CC	0.73	0.08	ASSESS+FAPs+DAL	E	CC	0.67	0.04

In each part the possible set of combinations of input feature modalities (7) for each of the FEELTRACE subject (4) are given in terms of the training set used as either activations (A) or evaluation (E). The final column in each part gives the variance of the results over nine testing runs.

(IMC) neurons were employed, with the learning rate fixed at 0.001. Also of each dataset, four parts were used for the training and one part was used for testing the net on 'unseen' inputs. In Table 1, that follows. Input space stands for the type of features used; Out stands for the output dimension used for training (A: activation, E: evaluation); FT stands for the particular FeelTracer used as supervisor; AVG denotes the average quadrant match (for 2D-space) or average half-plane match for (1D-space) over three runs and SD denotes the standard deviation for the above average. We also add that PCA was used on the ASSESS features so as to reduce from about 500 input features to around 7–10 as describing most of the volatility of the series over time.

5.3. Analysis of results in Table 1

The results are shown in Table 1.

We will now turn to analyse the results from Table 1. We note first the smallness of the standard deviations for all

the results, indicating considerable stability across the various trained networks used to calculate this parameter. This gives more confidence in the results. Moreover, various alternate parameter choices for numbers of hidden nodes, etc., were assessed on a range of benchmark problems in order to arrive at the above parameter choices made in the simulation. At the same time some runs were performed with alternate numbers of hidden nodes, but did not produce any improvements. Furthermore, the results have enough stability to indicate that these results from the training of ANNA can be accepted as a reasonable approximation to what can be best achieved in trained recognition system.

Next we note that the results for classification using the A output only are relatively high, in three cases up to 98%, and with two more at 90% or above. In particular we note the effectiveness of the data coming from the FeelTracer EM, with the average success rates of 86, 88, 98, 89, 95, 98 and 98% for the A input. A high level of success is obtained with Feeltracer JD, although not at quite such a high level. There are consistently lower values for the FeelTracer DR, all in the 60–66% band, and CC, who varies more across the modalities used, with values of 64, 54, 75, 63, 73, 73 and 73%.

Let us now consider the effect of different input combinations on the success rate, when we use only the output A. From Table 1, we see that there is not a significant difference between the input combination choices of DAL + FAPs, ASSESS + DAL, ASSESS + FAPs + DAL, and DAL on its own.

The E output leads to considerably lower success rates, with not so much disparity across the FeelTracers. On the other hand, the A + E output also has EM as the best FeelTracer to model, with success rates of 61, 60, 58, 65, 71, 61 and 67%, all well above the chance level of 25%. The best successes arise from the input combinations ASSESS + FAPs + DAL, ASSESS + DAL, ASSESS + FAPs and DAL.

Finally, there is more difficulty to reach a similar assessment with the full A + E full FeelTrace output, due to the considerable variation across the FeelTracers. For EM the optimal choices of input combinations are DAL + FAPs (71%), ASSESS + FAPs + DAL (67%), and ASSESS + FAPs (65%)—so especial added value from the FAPs. However for the FeelTracer CC, the optimal inputs are ASSESS (48%), ASSESS + DAL (49%) or ASSESS + FAPs + DAL (43%). This implies that added value for CC arises especially from the ASSESS prosodic features. Different combinations adding value do not arise so clearly for the other two FeelTracers: for JD the results are very similar across all input combinations, with success levels 48, 52, 51, 51, 51, 57 and 47%, while for DR these values are 37, 31, 39, 44, 35, 37 and 41%.

5.4. Conclusions on the results

The results presented above lead to the general conclusion that it is possible to obtain reasonably good results on classification of the emotional states of human subjects. However, a very important further conclusion is that different FeelTracers may be using different modalities on which to judge the emotional state in which a particular subject presently finds themselves. We used in this discussion only the A + E output results as relevant, since that is what each FeelTracer is trying to produce from their viewing of the emotional states of the subject. For EM, then, the most important cued modality would be that from facial structure, for CC it would be from prosody, while for the other two, JD and DR, there was no clear preponderance of modality from which cues were being extracted.

This difference between the different results of different FeelTracers, and in particular their use of different modalities in judging the emotional states of those they are observing, had already been referred to in the earlier part of the paper. Thus it was noted that there may be cultural differences of emotion state recognition. Here, as we mentioned in that context, but did not support then, we now present data that show that even in a given culture it would seem that there are differences as to what are important clues as to the emotional states of others, what are not.

6. Conclusions

In this paper we have introduced the framework of the EC project ERMIS. The aim of this project was to build an automatic emotion recognition system able to exploit multimodal emotional markers such as those embedded in the voice, face and words spoken. We discussed the numerous potential applications of such a system for industry as well as in academia. We then turned to the psychological literature to help lay the theoretical foundation of our system and make use of insights from the various emotion theories proposed in shaping the various components of our automatic emotion recognition system such as the input and output representations as well as the internal structure. We thus looked at the different proposals with respect to the size (basic emotions hypothesis) and nature (discrete or continuous) of the emotional space as well as its origin (evolution or social learning) and hierarchical structure (primary, secondary).

We then proceeded to discuss several issues that pertain to emotion recognition as suggested by psychological research. These were explored separately for the input and the output of the automatic recognition system. The input-related issues involved inconsistencies in the expression of emotion due to social norm, deceptive purposes as well as natural ambiguity of emotional expression. The output-related issues pertained to the nature of the representation of emotional states (discrete and categorical or continuous and non-categorical) as well as the time course of emotional states and the implications of different time scales for automatic recognition. We then proceeded to examine the features that can be used as emotional markers in the various modalities as well as the computational studies carried out based on those features. We also looked at the neural

correlates of the perception of those features or markers. With respect to the word stream we reviewed the state-of-the-art in emotion extraction from text and presented our DAL system developed for ERMIS.

We then presented an artificial neural network called ANNA developed for the automatic classification of emotional states driven by a multimodal feature input. The novel feature of ANNA is the feedback attentional loop designed to exploit the attention-grabbing effect of emotional stimuli to further enhance and clarify the salient components of the input stream. Finally we presented the results obtained through the use of ANNA on training and testing material based on the SALAS scenario developed within the ERMIS framework.

The results obtained by using ANNA indicate that there can be crucial differences between subjects as to the clues they pick up from others about the emotional states of the latter. This was shown in the Table 1 of results, and discussed in Sections 5.3 and 5.4. It is that feature, as well as to possible differences across what different human objects also 'release' to others about their inner emotional states, whose implications for constructing an artificial emotion state detector we need to consider carefully. Environmental conditions, such as poor lighting or a poor audio recorder, will influence the input features available to the system. Independently of that, we must take note of the further variability of the training data available for the system.

Acknowledgements

We would like to acknowledge help from all of the partners in ERMIS, especially Roddie Cowie and Ellie Douglas-Cowie from QUB, as well as our colleagues from NTUA led by Stefanos Kollias. We would also like to acknowledge financial help from the EC under project ERMIS, under which this work was carried out.

Appendix A. ANNA equations (reference)

$$y_i = f\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} z_k\right]\right)$$

$$z_k = f\left(\sum_i B_{ki} y_i\right)$$

Define:

$$f'_{y_i(\infty)} = f'\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} z_k(\infty)\right]\right)$$

$$f'_{z_k(\infty)} = f'\left(\sum_i B_{ki} y_i(\infty)\right)$$

$$\frac{\partial y_i(\infty)}{\partial \omega_{lm}} = f'_{y_i(\infty)}\left\{\sum_j \frac{\partial \omega_{ij}}{\partial \omega_{lm}} x_j \left[1 + \sum_K A_{ijk} z_k(\infty)\right] + \sum_j \omega_{ij} x_j \sum_k A_{ijk} \frac{\partial z_k(\infty)}{\partial \omega_{lm}}\right\} \tag{A1}$$

$$\frac{\partial z_k(\infty)}{\partial \omega_{lm}} = f'_{z_k(\infty)} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \tag{A2}$$

(A1) and (A2)⇒

$$\begin{aligned}\frac{\partial y_i(\infty)}{\partial \omega_{lm}} &= f'_{y_i(\infty)} \sum_j \delta_{il}\delta_{jm} x_j \left[1 + \sum_k A_{ijk} z_k(\infty)\right] \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \\ &\Rightarrow \frac{\partial y_i(\infty)}{\partial \omega_{lm}} \\ &= f'_{y_i(\infty)} \delta_{il} x_m \left[1 + \sum_k A_{imk} z_k(\infty)\right] \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \Rightarrow\end{aligned}$$

$$\begin{aligned}&\sum_r \delta_{ir} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \\ &= f'_{y_i(\infty)} \delta_{il} x_m \left[1 + \sum_k A_{imk} z_k(\infty)\right] \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \\ &\Rightarrow \sum_r \left\{\delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}\right\} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} \\ &= f'_{y_i(\infty)} \delta_{il} x_m \left[1 + \sum_k A_{imk} z_k(\infty)\right]\end{aligned} \tag{A3}$$

Define (A3)⇒

$$L_{ir} = \delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}$$

$$\sum_r L_{ir} \frac{\partial y_r(\infty)}{\partial \omega_{lm}} = \delta_{il} f'_{y_i(\infty)} \left[1 + \sum_k A_{imk} z_k(\infty)\right] x_m \tag{A4}$$

Rewrite (A4) in matrix form:

$$\mathbf{L}\frac{\partial \mathbf{y}(\infty)}{\partial \omega_{lm}} = \begin{bmatrix} \delta_{1l} f'_{y_1(\infty)}\left[1 + \sum_k A_{1mk} z_k(\infty)\right] \\ \delta_{2l} f'_{y_2(\infty)}\left[1 + \sum_k A_{2mk} z_k(\infty)\right] \\ \vdots \\ \delta_{Pl} f'_{y_P(\infty)}\left[1 + \sum_k A_{Pmk} z_k(\infty)\right] \end{bmatrix} x_m$$

Multiply both sides by the inverse of $\mathbf{L}$:

$$\frac{\partial \mathbf{y}(\infty)}{\partial \omega_{lm}} = \mathbf{L}^{-1}\begin{bmatrix} \delta_{1l} f'_{y_1(\infty)}\left[1 + \sum_k A_{1mk} z_k(\infty)\right] \\ \delta_{2l} f'_{y_2(\infty)}\left[1 + \sum_k A_{2mk} z_k(\infty)\right] \\ \vdots \\ \delta_{Pl} f'_{y_P(\infty)}\left[1 + \sum_k A_{Pmk} z_k(\infty)\right] \end{bmatrix} x_m$$

Retain ξth component:

$$\begin{aligned} \frac{\partial y_\xi(\infty)}{\partial \omega_{lm}} &= \sum_s (L^{-1})_{\xi s}\delta_{sl} f'_{y_s(\infty)}\left[1 + \sum_k A_{smk} z_k(\infty)\right] x_m \\ &\Rightarrow \frac{\partial y_\xi(\infty)}{\partial \omega_{lm}} \\ &= (L^{-1})_{\xi l} f'_{y_l(\infty)}\left[1 + \sum_k A_{lmk} z_k(\infty)\right] x_m \end{aligned} \tag{A5}$$

Appendix B

ANNA equations (reference)

$$y_i = f\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} z_k\right]\right)$$

$$z_k = f\left(\sum_i B_{ki} y_i\right)$$

$$\frac{\partial y_i(\infty)}{\partial A_{lmn}} = f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k \left\{\frac{\partial A_{ijk}}{\partial A_{lmn}} z_k(\infty) + A_{ijk}\frac{\partial z_k(\infty)}{\partial A_{lmn}}\right\} \tag{B1}$$

$$\frac{\partial z_k(\infty)}{\partial A_{lmn}} = f'_{z_k(\infty)} \sum_r B_{kr}\frac{\partial y_r(\infty)}{\partial A_{lmn}} \tag{B2}$$

(B1) and (B2)⇒

$$\begin{aligned} \frac{\partial y_i(\infty)}{\partial A_{lmn}} &= f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k \left\{\delta_{il}\delta_{jm}\delta_{kn} z_k(\infty) \right. \\ &\quad \left. + f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr}\frac{\partial y_r(\infty)}{\partial A_{lmn}}\right\} \Rightarrow \frac{\partial y_i(\infty)}{\partial A_{lmn}} \\ &= f'_{y_i(\infty)}\delta_{il}\omega_{im} x_m z_n(\infty) \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr}\frac{\partial y_r(\infty)}{\partial A_{lmn}} \Rightarrow \end{aligned}$$

$$\begin{aligned} &\sum_r \delta_{ir}\frac{\partial y_r(\infty)}{\partial A_{lmn}} \\ &= f'_{y_i(\infty)}\delta_{il}\omega_{im} x_m z_n(\infty) \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} \sum_r B_{kr}\frac{\partial y_r(\infty)}{\partial A_{lmn}} \\ &\Rightarrow \sum_r \left\{\delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}\right\}\frac{\partial y_r(\infty)}{\partial A_{lmn}} \\ &= f'_{y_i(\infty)}\delta_{il}\omega_{im} x_m z_n(\infty) \end{aligned} \tag{B3}$$

Define (B3)⇒

$$M_{ir} = \delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}$$

$$\sum_r M_{ir}\frac{\partial y_r(\infty)}{\partial A_{lmn}} = \delta_{il} f'_{y_i(\infty)}\omega_{im} x_m z_n(\infty) \tag{B4}$$

Rewrite (B4) in matrix form:

$$\mathbf{M}\frac{\partial \mathbf{y}(\infty)}{\partial A_{lmn}} = \begin{bmatrix} \delta_{1l} f'_{y_1(\infty)}\omega_{1m} \\ \delta_{2l} f'_{y_2(\infty)}\omega_{2m} \\ \vdots \\ \delta_{Pl} f'_{y_P(\infty)}\omega_{Pm} \end{bmatrix} x_m z_n(\infty)$$

Multiply both sides by the inverse of $\mathbf{M}$:

$$\frac{\partial \mathbf{y}(\infty)}{\partial A_{lmn}} = \mathbf{M}^{-1}\begin{bmatrix} \delta_{1l} f'_{y_1(\infty)}\omega_{1m} \\ \delta_{2l} f'_{y_2(\infty)}\omega_{2m} \\ \vdots \\ \delta_{Pl} f'_{y_P(\infty)}\omega_{Pm} \end{bmatrix} x_m z_n(\infty)$$

Retain ξth component:

$$\begin{aligned} \frac{\partial y_\xi(\infty)}{\partial A_{lmn}} &= \sum_s (M^{-1})_{\xi s}\delta_{sl} f'_{y_s(\infty)}\omega_{sm} x_m z_n(\infty) \Rightarrow \frac{\partial y_\xi(\infty)}{\partial A_{lmn}} \\ &= (M^{-1})_{\xi l} f'_{y_l(\infty)}\omega_{lm} x_m z_n(\infty) \end{aligned} \tag{B5}$$

Appendix C

ANNA equations (reference)

$$y_i = f\left(\sum_j \omega_{ij} x_j \left[1 + \sum_k A_{ijk} z_k\right]\right)$$

$$z_k = f\left(\sum_i B_{ki} y_i\right)$$

$$\frac{\partial y_i(\infty)}{\partial B_{lm}} = f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k A_{ijk} \frac{\partial z_k(\infty)}{\partial B_{lm}} \tag{C1}$$

$$\begin{aligned}\frac{\partial z_k(\infty)}{\partial B_{lm}} &= f'_{z_k(\infty)} \left\{\sum_r \frac{\partial B_{kr}}{\partial B_{lm}} y_r(\infty) + \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial B_{lm}}\right\} \\ &\Rightarrow \frac{\partial z_k(\infty)}{\partial B_{lm}} \\ &= f'_{z_k(\infty)} \left\{\sum_r \delta_{kl}\delta_{rm} y_r(\infty) + \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial B_{lm}}\right\} \\ &\Rightarrow \frac{\partial z_k(\infty)}{\partial B_{lm}} \\ &= f'_{z_k(\infty)} \left\{\delta_{kl} y_m(\infty) + \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial B_{lm}}\right\}\end{aligned} \tag{C2}$$

(C1) and (C2) $\Rightarrow$

$$\begin{aligned}\frac{\partial y_i(\infty)}{\partial B_{lm}} &= f'_{y_i(\infty)} \sum_j \omega_{ij} x_j A_{ijl} f'_{z_l(\infty)} y_m(\infty) \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k A_{ijk} f'_{z_k(\infty)} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial B_{lm}} \\ &\Rightarrow \sum_r \delta_{ir} \frac{\partial y_r(\infty)}{\partial B_{lm}} \\ &= f'_{y_i(\infty)} \sum_j \omega_{ij} x_j A_{ijl} f'_{z_l(\infty)} y_m(\infty) \\ &\quad + f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k A_{ijk} f'_{z_k(\infty)} \sum_r B_{kr} \frac{\partial y_r(\infty)}{\partial B_{lm}} \\ &\Rightarrow \sum_r \left\{\delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}\right\} \\ \frac{\partial y_r(\infty)}{\partial B_{lm}} &= f'_{y_i(\infty)} \sum_j \omega_{ij} x_j f'_{z_l(\infty)} A_{ijl} y_m(\infty)\end{aligned} \tag{C3}$$

Define (C3) $\Rightarrow$

$$N_{ir} = \delta_{ir} - f'_{y_i(\infty)} \sum_j \omega_{ij} x_j \sum_k f'_{z_k(\infty)} A_{ijk} B_{kr}$$

$$\sum_r N_{ir} \frac{\partial y_r(\infty)}{\partial B_{lm}} = f'_{y_i(\infty)} \sum_j \omega_{ij} x_j f'_{z_l(\infty)} A_{ijl} y_m(\infty) \tag{C4}$$

Rewrite (C4) in matrix form:

$$\mathbf{N} \frac{\partial \mathbf{y}(\infty)}{\partial B_{lm}} = \begin{bmatrix} f'_{y_1(\infty)} \sum_j \omega_{1j} x_j A_{1jl} \\ f'_{y_2(\infty)} \sum_j \omega_{2j} x_j A_{2jl} \\ \vdots \\ f'_{y_P(\infty)} \sum_j \omega_{Pj} x_j A_{Pjl} \end{bmatrix} f'_{z_l(\infty)} y_m(\infty)$$

Multiply both sides by the inverse of $\mathbf{N}$:

$$\frac{\partial \mathbf{y}(\infty)}{\partial B_{lm}} = \mathbf{N}^{-1} \begin{bmatrix} f'_{y_1(\infty)} \sum_j \omega_{1j} x_j A_{1jl} \\ f'_{y_2(\infty)} \sum_j \omega_{2j} x_j A_{2jl} \\ \vdots \\ f'_{y_P(\infty)} \sum_j \omega_{Pj} x_j A_{Pjl} \end{bmatrix} f'_{z_l(\infty)} y_m(\infty)$$

Retain ξth component:

$$\frac{\partial y_\xi(\infty)}{\partial B_{lm}} = \sum_s (N^{-1})_{\xi s} f'_{y_s(\infty)} \sum_j \omega_{sj} x_j f'_{z_l(\infty)} A_{sjl} y_m(\infty) \tag{C5}$$

References

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, *12*, 169–177.

Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: A 3-D lesion study. *Emotion*, *2*, 23–51.

Arnold, M. B. (1960). *Emotion and personality, Neurological and physiological aspects*, vol. 2. New York: Columbia University Press.

Averill, J. R. (1980). A constructionist view of emotion. In R. Plutchik, & H. Kellerman, *Emotion: Theory, research and experience* (vol. 1) (pp. 305–339). New York: Academic Press.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Persolity and Social Psychology*, *70*, 614–636.

Carver, C. S. (2001). Affect and the functional bases of behavior: On the dimensional structure of affective experience. *Personality and Social Psychology Review*, *5*, 345–356.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, *1*, 32–80.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam Publishing Group.

Ekman, P. (1975). Face muscles talk every language. *Psychology Today*, *9*, 39.

Ekman, P. (1994a). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, *115*, 268–287.

Ekman, P. (1994b). All emotions are basic. In P. Ekman, & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 7–19). New York: Oxford University Press.

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. San Fransisco, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press.

Ellsworth, P. (1994). Some reasons to expect universal antecedents of emotion. In P. Ekman, & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions*. New York: Oxford University Press.

Fragopanagos, N., Kockelkoern, S., & Taylor, J. G. (2005). A neurodynamic model of the attentional blink. *Cognitive Brain Research* (in press).

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Hartley, M., Taylor, N., & Tayor, J. G. (2005). Attention as Sigma-Pi Controlled Ach-Based Feedback. Proceeding of IJCNN05 to be published).

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion–cognition relations. *Psychology Review*, *99*, 561–565.

Lazarus, R.S. (1968). Emotions and adaptation: Conceptual and empirical relations. In W.J. Arnold (Ed.), *Nebraska symposium on motivation* (vol. 16) (pp. 175–270). Lincoln: University of Nebraska Press.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*, 1097–1108.

Neese, R. M. (1990). Evolutionary explanations of emotions. *Human Nature*, *1*, 261–289.

Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion*, *1*, 29–50.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychology Review*, *97*, 315–331.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York: Harper and Row.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*, 102–141.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*, 805–819.

Solomon, R. C. (2003). *What is an emotion?: Classic and contemporary readings*. New York: Oxford University Press.

Tayor, J. G., (2005). Paying attention to consciousness. Progress in Neurobiology *71*,305–335.

Tomkins, S. S. (1962). *Affect, imagery, consciousness*. New York: Springer.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, *11*, 407–424.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, *9*, 36–45.

Whissell, C. M. (1989). The dictionary of affect in language. In R. Plutchik, & H. Kellerman (Eds.) , *Emotion: Theory, research and experience*. *The measurement of emotions* (vol. 4) (pp. 113–131). New York: Academic Press.