J Multimodal User Interfaces
DOI 10.1007/s12193-013-0122-3

ORIGINAL PAPER

# Induction, recording and recognition of natural emotions from facial expressions and speech prosody

**Kostas Karpouzis · George Caridakis · Roddy Cowie · Ellen Douglas-Cowie**

Received: 13 April 2012 / Accepted: 28 March 2013


**Abstract** Recording and annotating a multimodal database of natural expressivity is a task that requires careful planning and implementation, before even starting to apply feature extraction and recognition algorithms. Requirements and characteristics of such databases are inherently different than those of acted behaviour, both in terms of unconstrained expressivity of the human participants, and in terms of the expressed emotions. In this paper, we describe a method to induce, record and annotate natural emotions, which was used to provide multimodal data for dynamic emotion recognition from facial expressions and speech prosody; results from a dynamic recognition algorithm, based on recurrent neural networks, indicate that multimodal processing surpasses both speech and visual analysis by a wide margin. The SAL database was used in the framework of the Humaine Network of Excellence as a common ground for research in everyday, natural emotions.

**Keywords** Affective computing · Multimodal emotion recognition · Emotion induction methods

K. Karpouzis · G. Caridakis (✉)
Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Athens, Greece
e-mail: gcari@image.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ntua.gr

R. Cowie · E. Douglas-Cowie
School of Psychology, Queen's University, Belfast, Belfast, UK
e-mail: r.cowie@qub.ac.uk

E. Douglas-Cowie
e-mail: e.douglas-cowie@qub.ac.uk

## 1 Introduction to databases of natural interaction

Recognizing everyday emotions is a challenging problem, either in the unimodal case (where faces, gestures or speech prosody and utterances are considered alone); or in the multimodal case (in any combination of the above modalities). It depends on dealing in a co-ordinated way with issues ranging from psychology to signal processing. Nevertheless, there are now approaches that yield promising results.

Not the least of the problems is finding suitable data. The commonest pattern is that algorithms are tested on data produced for that specific study, making comparison across studies extremely difficult. Furthermore, in most cases, 'produced' has been equivalent to 'acted', in the sense that one or more people are asked to display certain emotions. These are often extreme versions of the 'basic emotions' proposed by Ekman and Friesen [28], facing straight into a camera or speaking directly into a high-quality microphone in front of them.

Even though the seminal application of an emotion recognition algorithm would be to classify the user's affective state with a particular label or another representation, the outcome of this process can be used in a variety of more complex applications. The first step in such an algorithm is always feature extraction, either from a holistic point of view or by localizing facial features or extracting statistical measurements from speech and physiological signals. These data can also be used to estimate cognitive states, such as engagement, attention or boredom, which are extremely beneficial in intelligent human-machine interaction, along with some emotional states provided by the emotion recognition algorithm. Consider, for instance, an emotion recognition system operating within a car and estimating fatigue or lack of attention to the road from the part of the driver; this can be used to trigger a response from the car computer (e.g. increasing

Published online: 26 April 2013

the volume of the car stereo or prompting the driver to stop for rest at the next car station) or alert nearby cars or the road infrastructure that the driver is not in optimal state. Similar pervasive applications can be envisioned in domestic or work environments, extending beyond health care to non-verbal feedback about a TV program or search engine search (satisfaction/dissatisfaction). In addition, unimodal features can be used to estimate the user's focus of attention (e.g. using head pose and eye gaze or speech localization), which would be alleviate the user interface burden in a smart home setting.

It is reasonably obvious that everyday emotions are rarely as extreme and easily distinguishable as those in the early image databases (e.g., the AR Face Database). There is also growing evidence that the expression of emotion is not qualitatively the same when it is acted as it is when emotions are induced or natural [4,6,17,24,56,55,64]. As a result, although acted data may be appropriate for testing feature extraction and recognition algorithms, emotion recognition systems trained on it cannot be expected to generalize to unconstrained, everyday conditions.

On the other hand, fully naturalistic recordings present their own problems, both of interpretation and of recording quality [13,27]. As a result, there is great interest in data generated by techniques designed to elicit emotion deliberately. Such approaches can produce data that can be both tractable and reasonably naturalistic. Many induction techniques are in use in the machine learning context—often involving computing tasks [2,4] or computer games [5,57,60] and sometimes involving human-human interaction [1,3,42]. More recently, [7] use a variant of the Velten mood induction technique [58] where participants read aloud a number of predefined sentences that put them in particular emotional state. The sentences were shown in three coherent blocks with first positive, then neutral and finally negative sentences in order to put the users gradually into the desired mood. Although the natural and naturalistic, spontaneous emotion elicitation potential of read speech is questionable it is sometimes useful for focused research questions on more than natural emotion analysis (e.g. cross-cultural). Additionally, data corpora such as Hollywood Human Action (HoHA) [39] and Acted Facial Expressions In The Wild (AFEW) [23] are based on excerpts from existing movies and TV shows rather than designing and recording a new corpus. This technique has been effective in emotion recognition research in naturalistic environments since the 'good' (method) actors have the ability to convey natural emotions. For an extended review of induction techniques, see Cowie et al. [20].

One of the challenges for that approach is to collect records of human-machine conversation, because machines are not actually able to carry out conversations. However, there are obvious reasons to try, since it seems very likely that human-machine interactions will differ from human-human interactions in significant ways. Not the least of these is that for the foreseeable future, human-machine interactions will break down in ways that human-human interactions do not and it is important to have ways of recognising the signs of breakdown. In this paper, we introduce a specific type of induction technique that focuses on conversation between a human and an agent simulating a machine. It is designed to capture a broad spectrum of emotional states, expressed in 'emotionally coloured discourse' of the type likely to be displayed in everyday conversation. The system with which the user interacts is called a Sensitive Artificial Listener, or SAL for short.

It is becoming clear that there is no one ideal type of data for emotion recognition research. Different applications require different types of data. Hence, the data obtained using SAL techniques occupies a particular niche. The rest of this section indicates what that is.

### 1.1 Extracted features

It has recently been recognised [55] that different applications impose different requirements for capturing, storing and dealing with the data and the extracted features. For example, an emotion recognition application and an automated call-centre application may both depend on speech in order to estimate the general emotional state of the user; however, in the latter case, transforming and encoding the speech signal in order to be transmitted via a phone line effectively cuts out information stored in the higher frequencies, this limiting the choice of features that the recognizer depends upon. In this case, a general purpose emotional speech recognizer may fail, since the information contained in the speech signal will generally not be enough.

In that context, it is instructive to compare two of the multimodal databases which were used within the Humaine Network of Excellence [36], the Sensitive Artificial Listener material and EmoTV [22]. They are broadly comparable in terms of the expressivity of their content, but they differ vastly in terms of the quality of the low level video signal. Specifically, the SAL database contains faces that cover most of the video frame (interocular distance is typically in the order of 170 pixels, see Fig. 1a), while in the EmoTV (Fig. 1b) database, faces are much smaller (less than 75 pixels in some cases).

### 1.2 Multimodality and interaction

EmoTV recordings show a rich variety of gestures, where there is very little in SAL data: initial results show that these contain useful information. Relationships to physical context (such as what the person is looking at or pointing to) are much more complex in EmoTV, and potentially very informative. The interaction between the user and the system is


**Fig. 1** Indicative instances form natural expressivity corpora [SAL (**a**) and EmoTV (**b**) databases] illustrating the challenges in extracting prominent facial features: arbitrary head pose and low resolution

restricted: for instance, the system cannot out-argue or correct the user, and so the emotions associated with events like that do not appear. SAL material is not acted in a simple sense, but it is to some extent a game. These limitations are probably acceptable in a considerable range of applications, but it needs to be clear that SAL material is only useful if they are. It is a useful summary to say that SAL data models sociable, sedentary conversation.

## 1.3 Emotion descriptions

This section deals with the choice of emotion descriptions for SAL material. The term 'emotion descriptions' is used to refer to different ways of representing emotions: the term emotional model is used to talk about a computational system that implements a particular theory of emotion in specific context (and which will typically use a particular kind of emotion description).

As with recording, there is no ideal way to describe emotional content: context determines which choices make sense. The issue is sometimes to know a person's true emotion even if it is being well concealed, sometimes to react as another person probably would [19]. Sometimes one specific state, such as distress, is the target; sometimes a broad sense of emotional tone is enough. In sociable conversation, the obvious goal is to react to broad emotional tone as another person probably would, and the description chosen reflects that. Key options are sketched here for reference:

### *1.3.1 Categorical representations*

Humans' default approach to describing emotion is to use everyday words that describe qualitatively different types of emotional state. That will be called categorical representation. Various sets of categories have been proposed on different grounds, including evolutionarily basic emotion categories, most frequent everyday emotions, and application-specific emotion sets. Outcomes include Ekman's list of six basic emotions in Ekman and Friesen [28], the Ortony et al. list of emotion words in [45], and the Feeltrace core vocabulary in Cowie et al. [15].

There are two main problems with using categorical terms in the context of sociable conversation. First, an early study of databases containing sociable conversation (though the term was not used at that stage) observed that standard categorical terms are rarely a very accurate fit to the emotions observed there [14]. People typically use multiple terms, and inter-rater agreement is quite low. Second, the grounds for attributing many categorical terms, particularly terms that describe less extreme emotions, are logically rather complex, depending on quite deep understanding of the topics being discussed [13]. Automatic recognition of these states would depend on processing of speech and semantics that cannot be expected in the short term.

### *1.3.2 Appraisal theories and representations*

Many psychological theorists regard category terms as a rough and ready way of partitioning a space defined by theoretical constructs. Appraisal theory is a highly regarded form of that approach. It is grounded in cognitive theories of emotion [41], which argue that emotions are closely related to the situation that is being experienced (or, indeed, imagined) by the agent.

Appraisal representations characterise emotional states in terms of the way the emotional person is evaluating people, things, or events that are currently important to him or her, in terms of properties such as their familiarity, intrinsic pleasantness, relevance to goals. That kind of representation invites a two-way flow of information between external signs of emotion and representations of the context, highlighting what is emotionally significant, in humans or AI systems [40,52,34]. Appraisal theories are very common in emotion modelling since their structure makes it feasible to simulate their postulates in computational models, and some have been formulated explicitly in order to be implemented in a computer. Such an example is the OCC theory [45].

There have been attempts to apply appraisal-based descriptions to material from sociable interactions, but with

limited success [4,22]. There are two likely reasons. First, it is not clear how often appraisal categories have external signs as direct correlates. Second, the evaluations involved in sociable interaction tend to be very complex, with emotional significance attached to many entities at once, some physically present, some subjects of overt discussion, some privately imagined.

#### *1.3.3 Dimensional descriptions*

Dimensional descriptions have a long history in psychology. They capture very broad properties of emotional states, which emerge from statistical analysis. The most commonly considered are valence (negative/positive) and level of arousal (active/passive). A third which it commonly named is variously described as power, control, or approach / avoidance. Its most obvious advantage is the ability to distinguish between fear and anger, both of which involve negative valence and high activation. In anger, the subject of the emotion feels that he or she is powerful/in control; in fear, power/control is felt to lie elsewhere.

There is a case for believing that the main dimensions describe a real entity, 'core affect', that underpins emotional life [50]; but even if that is not so, it is clear that the dimensions are provide a serviceable summary. In the context of automatic recognition, dimensional representation is attractive mainly because it allows emotional states to be described in a way that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where there is likely to be a wide range of emotional states, which do not fall into discrete categories, and vary relatively continuously over time. A further attraction is the fact that dimensional descriptions can be translated into and out of verbal descriptions. This is possible because emotion words can, to an extent, be understood as referring to positions in activation-evaluation space [62]. That is exploited in the Feeltrace interface [18], where words are placed at the appropriate positions in the space so that users will be encouraged to adopt a consistent interpretation of the axes.

#### *1.3.4 Discrete and continuous descriptions*

It is natural to express categorical representations by identifying time periods during which a particular category is regarded as present, and dimensional representations by drawing a 'trace' which shows the person's position in the space from instant to instant. However, the issues are actually separate. Dimensional information can be expressed in a discrete representation (e.g. by identifying which quadrant of the space a person is in), and information about categories can be expressed in traces (e.g. by identifying how well a particular category term applies at any given instant).

### 1.4 Practicalities of labeling emotion-related states

The principles that have been sketched have been translated into working tools over a period of years. The aim is to establish how a particular display would probably be perceived, and so the relevant ground truth is the impression of emotion that observers form from what they see and hear. In the data considered here, they record that impression by moving a cursor as they listen and watch. Moving the cursor in the Feeltrace circle has the major advantage that it obtains a substantial proportion of the relevant information in a single viewing, and in a good deal of work, only that information has been considered. Later studies have used a larger number of one-dimensional scales. That provides more data, but it remains to be established how much more information it yields. That is to a large extent a statistical problem. Work on it is under way [20].

With the two basic dimensions, agreement among raters is reasonably high if it is measured on material where the range of emotional states is reasonably wide. Correlations are of the order of 0.8 for valence and 0.65 for activation [14, 22]. Agreement is slightly higher when raters simply record the intensity of emotion irrespective of its character, lower for category terms like anger and sadness, and considerably lower for power [22].

SAL clips are a key part of the HUMAINE database, which represents an attempt to integrate the various resources that would be needed to capture the essentials of the impressions that people form when they observe another person engaged in an emotionally coloured interaction. Some of its subtler features are used below to illustrate some specific issues. However, most of the analysis uses the basic, and most reliable dimensions, valence and activation.

## 2 The SAL technique

The aim of our effort is to develop systems capable of analyzing the non-extreme expressions of emotion that occur in sociable conversation. The work has been underpinned by data from the SAL induction technique [30,36]. SAL was developed within the ERMIS and HUMAINE projects. It is broadly akin to Weizenbaum's ELIZA [61]. The ELIZA framework simulates a Rogerian therapy, during which clients talk about their problems to a listener that provides responses that induces further interaction without passing any comment or judgment. But where ELIZA's responses are keyed to words in the client's input, SAL's are keyed to the client's emotional state. The responses that SAL gives are stock phrases, which have been chosen to elicit emotionally coloured reactions in the user.

In the versions that generated the data, the SAL operator chooses which statement to use at any given time from a menu

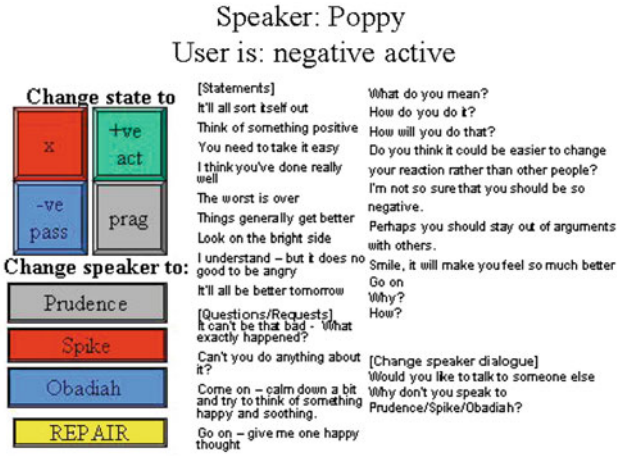


**Fig. 2** Example phrases used in the SAL framework

that is organised to simulate four personalities—Poppy (who aims to make people happy), Obadiah (who aims to make people gloomy), Spike (who aims to make people angry) and Prudence (who aims to make people pragmatic). Users choose at any time which 'personality' they want to talk to. The response that is chosen will depend on the 'personality' that is active and the user's state, which is classified in terms of the four quadrants of the Feeltrace circle. Figure 2 illustrates the options that the Poppy character may use when the user is (in Poppy's judgment) negative and active. It is one of 16 main menus defined by the four characters and four user states. The combination creates an environment rich enough to provoke exchanges that are extended, and quite highly coloured emotionally.

There have been different versions of SAL. In the framework of the HUMAINE project, a large audiovisual database was generated using a Wizard of Oz technique [21]. It was code named SAL 0. An experimenter had scripts for each combination of character and user state, and read from them, using different tones of voice for the four characters. A second version, SAL 1, used extended and re-organised scripts. The original version of SAL 1 was in English, and was successful enough for versions to be developed in Hebrew and Greek with adjustments to suit cultural norms and expectations. Progressively more fully automated versions have been developed under the SEMAINE project [54], Schroeder et al. 2009; Schroeder 2010), but they are not discussed here.

### 2.1 Recordings

Recording was a key concern in the development of SAL. The requirement of capturing both audio and visual inputs means that there must be compromise between demands of naturalness and signal processing. If the environment is too heavily oriented towards optimising recording quality, users are unlikely to show the everyday, relaxed emotionality that would cover most of the emotion representation space. On the other hand, visual and audio analysis algorithms cannot be expected to cope with totally unconstrained head and hand movement, subdued lighting, and mood music. Major issues may also arise from the different requirements of the individual modalities: while head mounted microphones might suit analysis of speech, they can raise severe difficulties for visual analysis. Eventually arrangements were developed to ensure that on the visual side, the face was usually almost frontal and well and evenly lit to the human eye; that it was always easy for a human listener to make out what was being said; and that the setting allowed most human participants to relax and express emotion within a reasonable time.

### 2.2 Collected data

The SAL scenarios engaged users in quite protracted conversations where there is a range of emotions and emotional intensities. The data generated is rich in facial and non verbal signals (e.g. aspects of pitch, spectral characteristics, and timing). Experience shows that listeners learn to 'use' the system, and this suggests that longitudinal use by small numbers may be preferable to occasional use by many. That is reflected in two bodies of SAL data.

The SAL 0 data obtained relatively short recordings from 20 users, 10 male and 10 female, totalling 105 min of footage. This has been segmented into files and is in avi and mpeg format. The technical specifications of the recorded data are 25 fps 720 × 576 resolution for the video and 320 kb/s for the audio. There are accompanying files of what was said.

The SAL 1 data consisted of longer recordings from four users. Each recorded for two sessions, each of approximately 30 min. The data is segmented into files in avi and mpeg format and four raters have labelled the data using the dimensional Feeltrace tool described above. That material is releasable under an agreement governing the use of the data. Greek and Hebrew versions were also developed. The Hebrew version has undergone a number of translations and data has now been collected from 5 users, totalling 2.5 h.

Substantial parts of the data have also been labelled in a more detailed way as part of the HUMAINE Database [25, 35]. Figure 3 illustrates the approach. It presents a SAL 1 sequence labelled as part of the HUMAINE Database. The ability of the scenario to evoke emotion is apparent from the user's facial expression is evident and this is borne out by the accompanying traces from a rater for emotion intensity and activation—first and fourth trace lines respectively. The point at which the shot of the face is taken (marked by the vertical red line) corresponds to a rise in the emotional intensity and degree of activation perceived by the rater. The rater's traces are also included for the degree to which the rater perceives the user to be acting or masking her emotion—second and third trace lines respectively). The pattern of the ActTrace

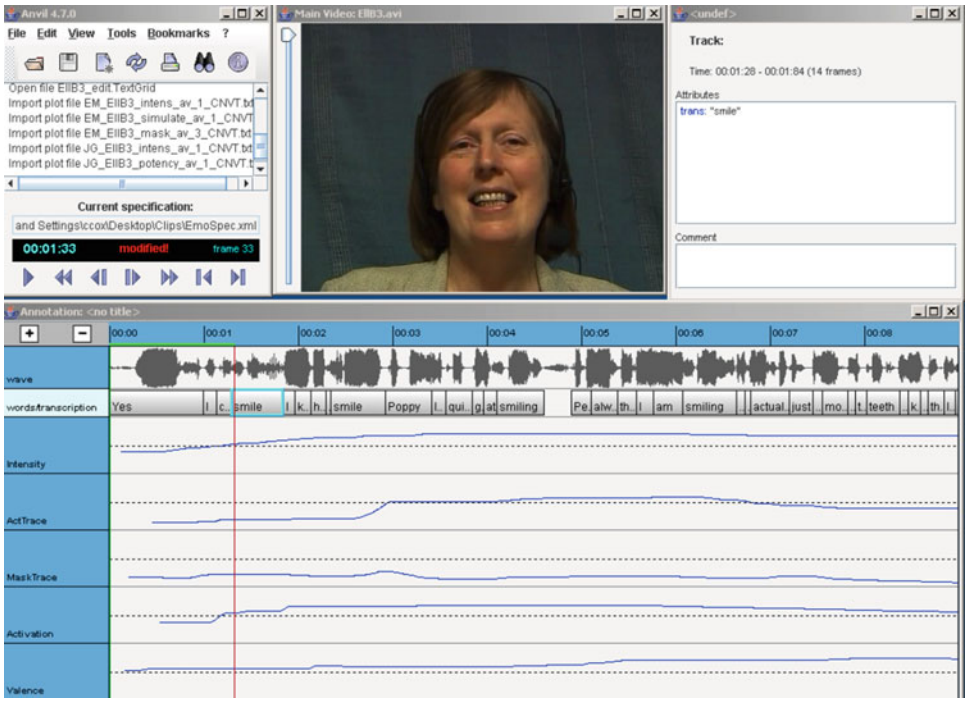


**Fig. 3** Labelled SAL data from the HUMAINE database

line indicates a low level of perceived acting at the start rising to absence of acting as the intensity of the emotion rises, thus indicating the naturalness of the emotion generated.

The SAL data that is already available is of sufficient quantity and quality to train machine recognition systems. Published reports of research using the material include [33] and [38]. More recent research reports very high recognition rates when the multimodal character of the data is exploited [16].

### 2.3 Recognition of natural facial expressions

The methodology used to detect facial features on the SAL database was first shown in Ioannou et al. [37,38] and Zeng et al. [65] offers an excellent review on methods used to recognise emotions in dynamic visual and multimodal databases. The first step is to locate the face, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, i.e. eyes, eyebrows, mouth and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria [63] to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation.

Measurement of Facial Animation Parameters (FAPs) requires the availability of a frame where the subject's expression is found to be neutral; in our case, this neutral frame is manually selected from the video sequences. The final feature masks are used to extract 19 Feature Points (FPs) [49]; Feature Points obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the Facial Animation Parameters (FAPs). Confidence levels on FAP estimation are derived from the equivalent feature point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

### 2.4 Features from prosody

An important difference between the visual and audio modalities is related to the duration of the sequence that we need to observe in order to be able to gain an understanding of the sequence's content. In case of video, a single frame is often enough in order for us to understand what the video displays and always enough for us to be able to process it and extract information. On the other hand, an audio signal needs to have a minimum duration for any kind of processing to be able to be made.

Therefore, instead of processing different moments of the audio signal, as we did with the visual modality, we need to process sound recordings in groups. Obviously, the meaningful definition of these groups will have a major role in

the overall performance of the resulting system. In this work we consider sound samples grouped as tunes [43], i.e. as sequences demarcated by pauses. The basis behind this is that, although expressions may change within a single tune, the underlying human emotion does not change dramatically enough to shift from one quadrant to another. For this reason, the tune is not only the audio unit upon which we apply our audio feature detection techniques but also the unit considered during the operation of the overall emotion classification system.

Among the multitude of available features, selecting the right set of audio features to consider for classification is far from a trivial procedure. In order to overcome this, we start by extracting an extensive set of 377 audio features. This comprises features based on intensity, pitch, Mel Frequency Cepstral Coefficient (MFCC), Bark spectral bands, voiced segment characteristics and pause length. In Castellano et al. [11], Prosogram [44], a method employing prosodic representation based on perception was used to analyze an acted dataset. Prosogram is based on a stylization of the fundamental frequency data (contour) for vocalic (or syllabic) nuclei and provides pitch and length information for each voiced nucleus. According to a 'glissando threshold' in some cases we don't get a fixed pitch but one or more lines to define the evolution of pitch for this nucleus. This representation is in a way similar to the 'piano roll' representation used in music sequencers. This method, based on the Praat environment, offers the possibility of automatic segmentation based both on voiced part and energy maxima. From this model—representation stylization we extracted several types of features: pitch interval based features, nucleus length features and distances between nuclei.

Given that the classification model used in this work is based on a neural network, using such a wide range of features as input to the classifies means that the size of the annotated data set as well as the time required for training will be huge. In order to overcome this we need to statistically process the acoustic feature, so as to discriminate the more prominent ones, thus performing feature reduction. In Caridakis et al. [9], this was achieved by combining two well known techniques: analysis of variance (ANOVA) and Pearson product-moment correlation coefficient (PMCC). ANOVA is used first to test the discriminative ability of each feature. This resulting in a reduced feature set, containing about half of the features tested. To further reduce the feature space we continued by calculating the PMCC for all of the remaining feature pairs; PMCC is a measure of the tendency of two variables measured on the same object to increase or decrease together. Groups of highly correlated (>90 %) features were formed, and a single feature from each group was selected. The overall process results in reducing the number of audio features considered during classification from 377 to only 32, those in Caridakis et al. [9]. Eyben et al. [32] uses a similar selection of 39 features in tune-based processing; those features were extracted using the openSMILE feature extractor [31].

## 3 Multimodal recognition

One of the aims of the recognition effort is to provide at least a general emotion estimate when working in environments where information is limited or defectively encoded. In these cases, the issue of having to work with imperfect data brings up a number of interesting questions:

- are my features wrong?, e.g. the visual features extraction process has failed to locate the eyebrows of the user
- are the features missing?, e.g. the user's hand has occluded the face
- are the features not needed or irrelevant?, e.g. maybe I don't need to know about the user's eyebrow to estimate the emotional state.

Taking into account more than one modality is a key way of dealing with the above questions. What is missing in one modality may be better represented in another; besides this, information in related modalities may be complementary, e.g. image and speech in the case of visemes [49] and phonemes. Work detailed in this section supports this theoretical statement by showing that early recognition results, i.e. when using a single modality, are improved in the case of multimodal recognition. The effectiveness of the multimodal approach and the incorporation of dynamics into the emotion recognition architecture are established when comparing Tables 1, 2 (Sect. 3.2) for unimodal, static speech based and multimodal, dynamic approach respectively.

### 3.1 Fusion of visual and acoustic features

As primary material we consider the audiovisual content collected using the SAL approach. This material was labelled using Feeltrace [18] by four labellers. The activation-valence coordinates from the four labellers were initially clustered into quadrants and were then statistically processed so that a majority decision could be obtained about the unique emotion describing the given moment. The corpus under investigation was segmented into 1,000 tunes of varying length. For every tune the input vector, as far as facial features is concerned, consisted of the FAPs produced by the processing of the frames of the tune. The acoustic input vector consisted of only one value per SBPF (Segment Based Prosodic Feature) per tune. The fusion was performed on a frame basis, meaning that the values of the SBPFs were repeated for every frame of the tune. This approach was preferred because it preserved the maximum of the available information since SBPFs are

only meaningful for a certain time period and cannot be calculated per frame. In terms of technical implementation fusing static and dynamic features is not a trivial task. To tackle the problem of features dynamics heterogeneity the architecture of the Recurrent Neural Network (see Sect. 3.2) was modified in order to facilitate the incorporation of multiple modalities. For the visual modality features we maintain the conventional input neurons used in most RNNs, while for the auditory modality features we use static value neurons. During the temporal evolution of the RNN operation while visual features corresponding to the next frames are fed to the first input neurons of the network, the static input neurons retain their original values for the auditory modality features.

### 3.2 Recurrent neural networks

A wide variety of machine learning techniques have been used in emotion recognition approaches [19,38,46]. Especially in the multimodal case [47,65], they all employ a large number of audio, visual or physiological features, a fact which usually impedes the training process; therefore, we need to find a way to reduce the number of utilized features by picking out only those related to emotion. An obvious choice for this is neural networks, since they enable us to pinpoint the most relevant features with respect to the output, usually by observing their weights. Although such architectures have been successfully used to solve problems that require the computation of a static function, where output depends only upon the current input, and not on any previous inputs, this is not the case in the domain of emotion recognition. One of the reasons for this is that expressivity is a dynamic, time-varying concept, where it is not always possible to deduce an emotional state merely by looking at a still image. As a result, Bayesian approaches which lend themselves nicely to similar problems [53] need to be extended to include support for time-varying features. Picard [48] proposes the use of Hidden Markov Models (HMMs) to model discrete emotional states (interest, joy or distress) and use them to predict the probability of each one, given a video of a user. However, this process needs to build a single HMM for each of the examined cases (e.g. each of the universal emotions), making it more suitable in cases where discrete emotions need to be estimated. In our case, building dedicated HMMs for each of the quadrants of the emotion representation would not suffice, since each of them contains emotions expressed with highly varying features (e.g. anger and fear in the negative/active quadrant).

Compared to earlier attempts at the same data, the dynamic approach outperforms frame-based ones described in Wallace et al. [59]

A more suitable choice would be RNNs (Recurrent Neural Networks) where past inputs influence the processing of future inputs [29]. RNNs possess the nice feature of modelling explicitly time and memory, catering for the fact that emotional states are not fluctuating strongly, given a short period of time. Additionally, they can model emotional transitions and not only static emotional representations, providing a solution for diverse feature variation and not merely for neutral to expressive and back to neutral, as would be the case for HMMs.

The implementation of a RNN we used was based on an Elman network [10,29]. The output classes were 4 (3 for the possible emotion quadrants, since the data for the positive/passive quadrant was negligible, and one for neutral affective state) resulting in a dataset consisting of around 10,000 records. The training/testing dataset was on a 3 to 1 ratio. The classification efficiency, for facial only and audio only, was measured at 67 and 73 % respectively but combining the two modalities we achieved a recognition rate of 79 %. This fact illustrates the ability of the proposed method

**Table 1** Confusion matrix for tunes lasting more than 10 frames

| | Neutral (%) | Q1 (%) | Q2 (%) | Q3 (%) | Q4 (%) | Totals (%) |
|---|---|---|---|---|---|---|
| Neutral | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 |
| Q1 | 0.0 | 98.2 | 1.7 | 0.0 | 0.0 | 100.0 |
| Q2 | 1.7 | 1.7 | 96.4 | 0.0 | 0.0 | 100.0 |
| Q3 | 0.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 |
| Q4 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 100.0 |
| Totals | 8.6 | 50.0 | 16.4 | 8.6 | 16.1 | 100.0 |

**Table 2** Unimodal speech based recognition results

| | Neutral | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|
| % | 74.3 | 73.9 | 80.8 | 79.1 | 82.6 |

**Table 3** Current approach compared to previous recognition attempts

| Methodology | Rule based | Possibilistic rule based | Dynamic and multimodal |
|---|---|---|---|
| Classification rate | 78.4 % | 65.1 % | 98.5 % |

to take advantage of multimodal information and the related analysis. Overall, the operation of this approach in normal operating conditions (as such we consider the case in which tunes have a length of at least 10 frames) is accompanied by a classification rate of 98.55 %, which is very high, as illustrate in Table 3, even for controlled data, let alone for naturalistic data. The threshold of 10 frames was imposed to alleviate the fact that the short-memory function of the Elman network can be used to adapt to ten instances, while the effect of input data 'older' than that is minimal; on the other hand, since the optimal performance of this function is achieved when processing 10 samples, the network could not be adapted to learn and generalize in shorter sequences [51]. From a human-computer interaction point of view, this is acceptable, since the duration of such a tune (10 frames) is less than half a second.

Common applications of recurrent neural networks include complex tasks such as modeling, approximating, generating and predicting dynamic sequences of known or unknown statistical characteristics. While conventional, static in some sense, neural networks provide one response, in the form of a value or vector of values at their output, after considering a given input, RNNs provide such inputs after each different time step. So, while the incorporation of dynamics and the choice of an appropriate classifier prove to be an advantageous approach it also introduces additional issues that have to be tackled. One question to answer is at which time step the network's output should be read for the best classification decision to be reached due to the fact that the very first outputs of a RNN are not very reliable. Dynamics are rarely modeled and recognized coherently until an adequate length of the data is introduced to the network. On the other hand, it is not always safe to utilize the output of the very last time step(s) as the classification result of the network due to possible fluctuations during these steps. This problem is quite similar to decision fusion for multiple modalities and different classifiers in the case of late fusion multimodal approaches but in the temporal domain. This has been tackled by adding a weighting integrating module to the output of the neural network in order to enhance our classification model with robustness and increase its stability [7].

## 4 Limitations: future directions

A great deal of research has focused on the detection of (mainly) visual features in pre-recorded, acted datasets and the utilization of machine learning algorithms to estimate the illustrated emotions. Even in cases of multimodality, features are fed into the machine learning algorithms without any real attempt to find structure and correlations between the features themselves and the estimated result. Neural networks are a nice solution to finding such relations, thus coming up with comprehensible connections between the input (features) and the output (emotion).

The fact that we use naturalistic and not acted data introduces a number of interesting issues, for example segmentation of the discourse in tunes. During the experiment, tunes containing a small number of frames (less than 5 frames, i.e. 0.2 s) were found to be error prone and classified close to chance level (not better than 37 %). This is attributed to the fact that emotion in the speech channel needs at least half a second to be expressed via wording, as well as to the internal structure of the Elman network which works better with a short-term memory of ten frames. From a labelling point of view, ratings from four labellers are available; in some cases, experts would disagree in more than 40 % of the frames in a single tune. In order to integrate this fact, the decision system has to take into account the inter-rater disagreement, by comparing this to the level of disagreement with the automatic estimation.

A future direction regarding the features themselves is to model the correlation between phonemes and FAPs. In general, FPs from the mouth area do not contribute much when the subject is speaking; however, consistent phoneme detection could help differentiate expression-related deformation (e.g. a smile) to speech-related. Regarding the speech channel, the multitude of the currently detected features is hampering the training algorithms. To overcome this, we need to evaluate the importance/prominence of features so as to conclude on the influence they have on emotional transition. This can be achieved through statistical analysis (PCA analysis, K-Means Cluster Analysis, Two Step Cluster Analysis, Hierarchical Cluster Analysis) or Sensitivity Analysis.

Another important concept this is left untouched in related efforts is that of context [12,26]; Duric introduces the concept of W5+ questions, ('who?', 'when?', 'why?', 'where?' and 'how?') to identify additional information shown in the video or audio clip, besides the expressed emotion. Indeed, knowing the environment in which humans express themselves and interact with other humans or machines is important to adapt the training and testing procedure [8] and also provide richer, instead of just better, recognition. Context and personality information is also extremely useful in the case

of conflicting cues across the different supported modalities, e.g. when a person smiles or laughs, and thereby the image analysis part detects 'joy', but says something along the lines of 'I'll get you for that!'. In this case, backing up the feature extraction and recognition processes with an interoperable knowledge representation ensures than prior knowledge about a specific person and/or environment is taken into account, with all the nice features that knowledge technologies offer (e.g. alignment of different sources of information, handling uncertainty or noise in some of the features). A decisive step towards that approach would be the establishment of a common, standard format to represent affect- or emotion-related annotation (Schroeder 2007, 2010), in which case, data interoperability issues would be resolved.

## 5 Conclusions

The general problem of recognising naturalistic human emotion is difficult, and it will remain difficult for the foreseeable future. However, there are interesting subproblems that can be 'picked off' and dealt with.

Broad analysis of the kind of data that SAL provides appears to be one of those subproblems. that is interesting partly because the work has natural extensions in several directions. it has been pointed out that richer descriptions of emotional state can be generated, and new modalities can be introduced, most obviously analysis based on the linguistic content of utterances. The task of generating data itself invites true automation, and progress has been made towards replacing the script reader in SAL 0 and SAL 1 with a system that selects form a similar script automatically (Schroeder et al. 2009). That raises intriguing questions about the behaviour of an agent trying to sustain a conversation with a human. In sum, building on the core of success that has been achieved with the SAL paradigm appears to be well worthwhile.

Regarding the recognition part, the striking outcome is confirmation that with this kind of relatively natural data, multimodal recognition outperforms unimodal. Visual analysis performs more poorly that speech, because of two reasons: firstly, unit segmentation is based on the speech channel (tunes) and units contain, by definition, only one emotion, so related samples are, again by definition, correlated; and secondly, detection of visual expressivity from the mouth area is hampered by voice-related mouth deformation. However, the benefit from using both channels is close to 10 %, which makes the effort worthwhile.

## References


1. Abassi AR, Uno T, Dailey M, Afzulpurkar NV (2007) Towards knowledge-based affective interaction: situational interpretation of affect. In: Paiva A, Prada R, Picard R (eds) Affective computing and intelligent interaction. Springer LNCS, Lisbon, Berlin, pp 452–463
2. Auberge V, Audibert N, Rilliard A (2004) E-Wiz: a trapper protocol for hunting the expressive speech corpora in lab. In Proceedings of 4th international conference on language resources and evaluation (LREC), pp 179–182
3. Bachorowski JA (1999) Vocal expression and perception of emotion. Curr Dir Psychol Sci 8(2):53–57
4. Batliner A, Fischer K, Huber R, Spilker J, Noeth E (2003) How to find trouble in communication. Speech Commun 40:117–143
5. Bechara A, Damasio A, Damasio H, Anderson S (1994) Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50:7–15
6. Busso C, Narayanan S (2008) Recording audio-visual emotional databases from actors: a closer look. In: Proceedings of international conference on language resources and evaluation, workshop on emotion: corpora for research on emotion and affect, Marrakech, Morocco
7. Caridakis G, Karpouzis K, Wallace M, Kessous L, Amir N (2010) Multimodal user's affective state analysis in naturalistic interaction. J Multimodal User Interfaces 3(1–2):49–66 (Springer)
8. Caridakis G, Karpouzis K, Kollias S (2008) User and context adaptive neural networks for emotion recognition. Neurocomputing 71(13–15):2553–2562 (Elsevier)
9. Caridakis G, Malatesta L, Kessous L, Amir N, Raouzaiou A, Karpouzis K (2006) Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proceedings of international conference on multimodal interfaces, Banff, Alberta
10. Caridakis G, Raouzaiou A, Karpouzis K, Kollias S (2006) Synthesizing gesture expressivity based on real sequences. In: Proceedings of workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC, 2006 Conference. Genoa
11. Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction. LNCS, vol 4868. Springer, Heidelberg
12. Chen J, Wechsler H (2007) Human computer intelligent interaction using augmented cognition and emotional intelligence. LNCS Volume 4563/2007. Springer, Berlin, pp 205–214
13. Cowie R (2005) What are people doing when they assign everyday emotion terms? Psychol Inq 16(1):11–18
14. Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Commun 40:5–32 (Elsevier)
15. 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: Mastorakis N (ed) Computational intelligence and applications. World Scientific Engineering Society, pp 109–114
16. Cowie R, Douglas-Cowie E, Karpouzis K, Caridakis G, Wallace M, Kollias S (2008) Recognition of emotional States in natural human-computer interaction. In: Tzovaras D (ed) Multimodal user interfaces. Springer Berlin, pp 119–153
17. Cowie R, Douglas-Cowie E, Mckeown G, Gibney C The challenges of dealing with distributed signs of emotion: theory and empirical evidence. In: Proceedings ACII 2009 (published IEEE), vol 1 pp 351–356
18. Cowie R, Douglas-Cowie E, Savvidou S, Mcmahon E, Sawey M, Schroeder M (2000) FEELTRACE: an instrument for recording perceived emotion in real time. In: Proceedings of ISCA workshop on speech and emotion, Northern Ireland, pp 19–24
19. 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 33–80
20. Cowie R, Mckeown G (2010) Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme SEMAINE deliverable D6b Downloaded from http://www.semaine-project.eu/. 10/11/2010

21. Dahlbaeck N, Jonsson A, Ahrenberg L (1993) Wizard of Oz studies: why and how. In: Proceedings of 1st international conference on intelligent user interfaces, Orlando, Florida, pp 193–200
22. 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: Proceedings of 5th international conference on language resources and evaluation, Genoa
23. Dhall, A, Goecke R, Lucey, S, Gedeon T (2011) Acted facial expressions in the wild database. Technical Report TR-CS-11-02, Australian National University
24. Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40:33–60 (Elsevier)
25. Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, Mcrorie M, Martin J-C, Devillers L, Abrilian S, Batliner A, Amir N, Karpouzis K (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Proceedings of 2nd international conference on affective computing and intelligent interaction, Lisbon
26. Duric Z (2002) Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proc IEEE 90(7):1272–1289
27. Ekman P (1993) Facial expression and emotion. Am Psychol 48(4):384–392
28. Ekman P, Friesen W (1978) The facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, San Francisco
29. Elman JL (1990) Finding structure in time. Cognit Sci 14:179–211
30. Emotionally Rich Man-machine Intelligent System (Ermis) (2008) IST-2000-29319. http://www.image.ntua.gr/ermis. Last retrieved 1 Sept 2008
31. Eyben F, Wollmer M, Schuller B (2009) openEAR–introducing the Munich open-source emotion and affect recognition toolkit. In: Proceedings of of ACII. Amsterdam, The Netherlands, pp 576–581
32. Eyben F, Wollmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R (2010) On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J Multimodal User Interfaces 3:7–19 (Special Issue on Real-time Affect Analysis and Interpretation: Closing the Loop in Virtual Agents, Springer)
33. Fragopanagos N, Taylor J (2005) Emotion recognition in human-computer interaction. Neural Netw 18:389–405 (Elsevier)
34. Frijda NH (1986) The emotions, studies in emotion and social interaction. Cambridge University Press, New York
35. Humaine Database. http://www.emotion-research.net/download/pilot-db. Last retrieved 1 September 2008
36. Humaine IST, Human-Machine Interaction Network on Emotion, 2004–2007. http://www.emotion-research.net. Last retrieved 1 Sept 2008
37. Ioannou S, Caridakis G, Karpouzis K, Kollias S (2007) Robust feature detection for facial expression recognition. EURASIP J Image Video Process (2)
38. Ioannou S, Raouzaiou A, Tzouvaras V, Mailis T, Karpouzis K, Kollias S (2005) Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Netw 18(4):423–435. (Special Issue on Emotion: Understanding & Recognition, Elsevier)
39. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE Conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, London, pp 1–8
40. Lazarus RS (1991) Emotion and adaptation. Oxford University Press, New York
41. Lazarus RS, Folkman S (1987) Transactional theory and research on emotions and coping. Eur J Pers 1(3):141–169
42. Martin J-C, Devillers L, Zara A, Maffiolo V, Lechenadec G (2006) The EmoTABOU corpus. Humaine Summer School, Genova, Italy, September 22–28
43. Mcgilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S (2000) Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proceedings of the ISCA workshop on speech and emotion
44. Mertens P (2004) The prosogram: semi-automatic transcription of prosody based on a tonal perception model. In: Bel B, Marlien I (eds) Proceedings of of Speech Prosody, Japan
45. Ortony A, Collins A, Clore GL (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
46. Pantic M, Rothkrantz L (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal Mach Intell 22(12):1424–1445
47. Pantic M, Sebe N, Cohn J, Huang T (2005) Affective multimodal human-computer interaction. In: Proceedings of the 13th annual ACM international conference on Multimedia, pp 669–676
48. Picard R (1997) Affective computing. MIT Press, Cambridge
49. Raouzaiou A, Tsapatsoulis N, Karpouzis K, Kollias S (2002) Parameterized facial expression synthesis based on MPEG-4. EURASIP J Appl Signal Process 2002(10):1021–1038
50. Russell JA, Feldman-Barrett L (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 76:805–819
51. Schaefer A, Zimmermann HG (2007) Recurrent neural networks are universal approximators. Int J Neural Syst 17(4):253–263
52. Scherer KR (1987) Toward a dynamic theory of emotion: the component process model of affective states. Geneva Stud Emot Commun 1:1–98
53. Sebe N, Cohen I, Huang TS (2005) Handbook of pattern recognition and computer vision. World Scientific, River Edge
54. Semaine IST, The sensitive agent project. http://www.semaine-project.eu. Last retrieved 1 Sept 2008
55. Valstar M, Gunes H, Pantic M (2007) How to distinguish posed from spontaneous smiles using geometric features. In:. Massaro D, Takeda K, Roy D, Potamianos A (eds) Proceedings of the 9th international conference on multimodal interfaces, ICMI 2007, Nagoya, Aichi, Japan, November 12–15, pp 38–45
56. Valstar M, Pantic M, Ambadar Z, Cohn J (2006) Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In: Proceedings of the 8th international conference on multimodal interfaces, ACM, New York, pp 162–170
57. van Reekum C, Johnstone T, Banse R, Etter A, Wehrle T, Scherer K (2004) Psychophysiological responses to appraisal dimensions in a computer game. Cognit Emot 18(663–688)
58. Velten E (1998) A laboratory task for induction of mood states. Behav Res Therapy 35:72–82
59. Wallace M, Ioannou S, Raouzaiou A, Karpouzis K, Kollias S (2006) Dealing with feature uncertainty in facial expression recognition using possibilistic fuzzy rule evaluation. Int J Intell Syst Technol Appl 1(3–4)
60. Wang N, Marsella S (2006) Evg: an emotion evoking game. In: Proceedings of 6th international conference on intelligent virtual agents. Springer LNCS, pp 282–291
61. Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):35–36
62. Whissel CM (1989) The dictionary of affect in language. In: Plutchnik R, Kellerman H (eds) Emotion: theory, research and experience: vol 4, the measurement of emotions. Academic Press, New York
63. Young JW (1993) Head and face anthropometry of adult U.S. civilians. FAA Civil Aeromedical Institute, 1963–1993

64. Zeng Z, Pantic M, Roisman G, Huang TS (2007) A survey of affect recognition methods: audio, visual and spontaneous expressions. In: Proceedings of the 9th international conference on multimodal interfaces. ACM, New York, pp 126–133

65. Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58