Chapter 10
Continuous Analysis of Affect from Voice and Face

Hatice Gunes, Mihalis A. Nicolaou, and Maja Pantic

10.1 Introduction

Human affective behavior is multimodal, continuous and complex. Despite major advances within the affective computing research field, modeling, analyzing, interpreting and responding to human affective behavior still remains a challenge for automated systems as affect and emotions are complex constructs, with fuzzy boundaries and with substantial individual differences in expression and experience [7]. Therefore, affective and behavioral computing researchers have recently invested increased effort in exploring how to best model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum e.g., from -1 to $+1$) of affective behavior in terms of latent dimensions (e.g., arousal, power and valence) and appraisals, rather than in terms of a small number of discrete emotion categories (e.g., happiness and sadness). This chapter aims to (i) give a brief overview of the existing efforts and the major accomplishments in modeling and analysis of emotional expressions in dimensional and continuous space while focusing on open issues and new challenges in the field, and (ii) introduce a representative approach for

H. Gunes (✉) · M.A. Nicolaou · M. Pantic
Imperial College, London, UK
e-mail: h.gunes@imperial.ac.uk

M.A. Nicolaou
e-mail: mihalis@imperial.ac.uk

M. Pantic
e-mail: m.pantic@imperial.ac.uk

H. Gunes
University of Technology Sydney (UTS), Sydney, Australia

M. Pantic
University of Twente, Twente, The Netherlands

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_10, © Springer-Verlag London Limited 2011

multimodal continuous analysis of affect from voice and face, and provide experimental results using the audiovisual Sensitive Artificial Listener (SAL) Database of natural interactions. The chapter concludes by posing a number of questions that highlight the significant issues in the field, and by extracting potential answers to these questions from the relevant literature.

The chapter is organized as follows. Section 10.2 describes theories of emotion, Sect. 10.3 provides details on the affect dimensions employed in the literature as well as how emotions are perceived from visual, audio and physiological modalities. Section 10.4 summarizes how current technology has been developed, in terms of data acquisition and annotation, and automatic analysis of affect in continuous space by bringing forth a number of issues that need to be taken into account when applying a dimensional approach to emotion recognition, namely, determining the duration of emotions for automatic analysis, modeling the intensity of emotions, determining the baseline, dealing with high inter-subject expression variation, defining optimal strategies for fusion of multiple cues and modalities, and identifying appropriate machine learning techniques and evaluation measures. Section 10.5 presents our representative system that fuses vocal and facial expression cues for dimensional and continuous prediction of emotions in valence and arousal space by employing the bidirectional Long Short-Term Memory neural networks (BLSTM-NN), and introduces an output-associative fusion framework that incorporates correlations between the emotion dimensions to further improve continuous affect prediction. Section 10.6 concludes the chapter.

10.2 Affect in Dimensional Space

Emotions and affect are researched in various scientific disciplines such as neuroscience, psychology, and cognitive sciences. Development of automatic affect analyzers depends significantly on the progress in the aforementioned sciences. Hence, we start our analysis by exploring the background in emotion theory, perception and recognition.

According to research in psychology, three major approaches to affect modeling can be distinguished [31]: categorical, dimensional, and appraisal-based approach. The categorical approach claims that there exist a small number of emotions that are basic, hard-wired in our brain, and recognized universally (e.g. [18]). This theory on universality and interpretation of affective nonverbal expressions in terms of basic emotion categories has been the most commonly adopted approach in research on automatic measurement of human affect. However, a number of researchers have shown that in everyday interactions people exhibit non-basic, subtle and rather complex affective states like thinking, embarrassment or depression. Such subtle and complex affective states can be expressed via dozens of anatomically possible facial and bodily expressions, audio or physiological signals. Therefore, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information [82]. Hence, a number of researchers advocate the use of dimensional description of human affect, where

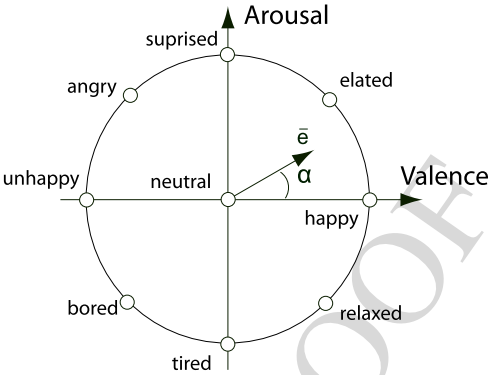


Fig. 10.1 Russel's valence-arousal space. The angle is represented by α while the vector $\bar{e}$ represents the emotion (*point*) as a parameter of arousal and valence. The figure is by courtesy of [77]

affective states are not independent from one another; rather, they are related to one another in a systematic manner (see, e.g., [31, 82, 86]). It is not surprising, therefore, that automatic affect sensing and recognition researchers have recently started exploring how to model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum from -1 to $+1$, without discretization) of affective behavior in terms of latent dimensions, rather than in terms of a small number of discrete emotion categories.

The most widely used dimensional model is a circular configuration called *Circumplex of Affect* (see Fig. 10.1) introduced by Russell [82]. This model is based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum. The proposed poles are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant), as illustrated in Fig. 10.1. Another well-accepted and commonly used dimensional description is the 3D emotional space of pleasure—displeasure, arousal—nonarousal and dominance—submissiveness [63], at times referred to as the *PAD emotion space* [48] or as *emotional primitives* [19].

Scherer and colleagues introduced another set of psychological models, referred to as componential models of emotion, which are based on the appraisal theory [25, 31, 86]. In the appraisal-based approach emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world (relevant concerns/needs) [25, 27, 31, 86]. Despite pioneering efforts of Scherer and colleagues (e.g., [84]), how to use the appraisal-based approach for automatic measurement of affect is an open research question as this approach requires complex, multicomponential and sophisticated measurements of change. One possibility is to reduce the appraisal models to dimensional models (e.g., 2D space of arousal-valence).

Ortony and colleagues proposed a computationally tractable model of the cognitive basis of emotion elicitation, known as OCC [71]. OCC is now established as a standard (cognitive appraisal) model for emotions, and has mostly been used in affect synthesis (in embodied conversational agent design, e.g. [4]).

Each approach, categorical or dimensional, has its advantages and disadvantages. In the categorical approach, where each affective display is classified into a single category, complex mental states, affective state or blended emotions may be too dif-

ficult to handle [108]. Instead, in dimensional approach, observers can indicate their impression of each stimulus on several continuous scales. Despite exhibiting such advantages, dimensional approach has received a number of criticisms. Firstly, the usefulness of these approaches has been challenged by discrete emotions theorists, such as Silvan Tomkins, Paul Ekman, and Carroll Izard, who argued that the reduction of emotion space to two or three dimensions is extreme and resulting in loss of information. Secondly, while some basic emotions proposed by Ekman, such as happiness or sadness, seem to fit well in the dimensional space, some basic emotions become indistinguishable (e.g., fear and anger), and some emotions may lie outside the space (e.g., surprise). It also remains unclear how to determine the position of other affect-related states such as confusion. Note, however, that arousal and valence are not claimed to be the only dimensions or to be sufficient to differentiate equally between all emotions. Nonetheless, they have already proven to be useful in several domains (e.g., affective content analysis [107]).

10.3 Affect Dimensions and Signals

An individual's inner emotional state may become apparent by subjective experiences (how the person feels), internal/inward expressions (bio signals), and external/outward expressions (audio/visual signals). However, these may be incongruent, depending on the context (e.g., feeling angry and not expressing it outwardly).

The contemporary theories of emotion and affect consider appraisal as the most significant component when defining and studying emotional experiences [81], and at the same time acknowledge that emotion is not just appraisal but a complex multifaceted experience that consists of the following stages (in order of occurrence):

1. *Cognitive Appraisal*. Only events that have significance for our goals, concerns, values, needs, or well-being elicit emotion.
2. *Subjective feelings*. The appraisal is accompanied by feelings that are good or bad, pleasant or unpleasant, calm or aroused.
3. *Physiological arousal*. Emotions are accompanied by autonomic nervous system activity.
4. *Expressive behaviors*. Emotions are communicated through facial and bodily expressions, postural and voice changes.
5. *Action tendencies*. Emotions carry behavioral intentions, and the readiness to act in certain ways.

This multifaceted aspect of affect poses a true challenge to automatic sensing and analysis. Therefore, to be able to deal with these challenges, affect research scientists have ended up making a number of assumptions and simplifications while studying emotions [7, 72]. These assumptions can be listed.

1. *Emotions are on or off at any particular point in time*. This assumption has implications on most data annotation procedures where raters label a user's expressed emotion as one of the basic emotion categories or a specific point in a

dimensional space. The main issue with this assumption is that the boundaries for defining the expressed emotion as on or off are usually not clear.

2. *Emotion is a state that the subject does not try to actively change or alleviate*. This is a common assumption during the data acquisition process where the subjects are assumed to have a simple response to the provided stimulus (e.g., while watching a clip or interacting with an interface). However, such simple passive responses do not usually hold during daily human–computer interactions. People generally regulate their affective states caused by various interactions (e.g., an office user logging into Facebook to alleviate his boredom).
3. *Emotion is not affected by situation or context*. This assumption pertains to most of the past research work on automatic affect recognition where emotions have been mostly investigated in laboratory settings, outside of a social context. However, some emotional expressions are displayed only during certain context (e.g., pain).

Affect research scientists have made the following simplifications while studying emotions [7, 72]:

1. *Emotions do occur in asynchronous communication* (e.g., via a prerecorded video/sound from a sender to a receiver). This simplification does not hold in reality as human nonverbal expressive communication occurs mostly face-to-face.
2. *Interpersonal emotions do arise from communications with strangers* (e.g., laboratory studies where people end up communicating with people they do not know). This simplification is unrealistic as people tend to be less expressive with people they do not know on an interpersonal level. Therefore, an automatic system designed using such communicative settings is expected to be much less sensitive to its user's realistic expressions.

Overall, these assumptions and simplifications are far from reality. However, they have paved the initial but crucial way for automatic affect recognizers that attempt to analyze both the felt (e.g., [9, 10, 59]) and the internally or the externally expressed (e.g., [50, 54]) emotions.

10.3.1 Affect Dimensions

Despite the existence of various emotion models described in Sect. 10.2, in automatic measurement of dimensional and continuous affect, valence (how positive or negative the affect is), activation (how excited or apathetic the affect is), power (the sense of control over the affect), and expectation (the degree of anticipating or being taken unaware) appear to make up the four most important affect dimensions [25]. Although ideally the intensity dimension could be derived from the other dimensions, to guarantee a complete description of affective coloring, some researchers include intensity (how far a person is away from a state of pure, cool rationality) as the fifth dimension (e.g., [62]). Solidarity, antagonism and agreement have also been in the list of dimensions investigated [13]. Overall, search for optimal low-dimensional representation of affect remains open [25].

10.3.2 Visual Signals

Facial actions (e.g., pulling eyebrows up) and facial expressions (e.g., producing a smile), and to a much lesser extent bodily postures (e.g., head bent backwards and arms raised forwards and upwards) and expressions (e.g., head nod), form the widely known and used visual signals for automatic affect measurement. Dimensional models are considered important in this task as a single label may not reflect the complexity of the affective state conveyed by a facial expression, body posture or gesture. Ekman and Friesen [17] considered expressing discrete emotion categories via face, and communicating dimensions of affect via body as more plausible.

A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, Russell [82] mapped the facial expressions to various positions on the two-dimensional plane of arousal-valence, while Cowie et al. [13] investigated the emotional and communicative significance of head nods and shakes in terms of arousal and valence dimensions, together with dimensional representation of solidarity, antagonism and agreement.

Although in a stricter sense not seen as part of the visual modality, motion capture systems have also been utilized for recording the relationship between body posture and affect dimensions (e.g., [57, 58]). For instance, Kleinsmith et al. [58] identified that scaling, arousal, valence, and action tendency were the affective dimensions used by human observers when discriminating between postures. They also reported that low-level posture features such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and right shoulder) appear to help in effectively discriminating between the affective dimensions [57, 58].

10.3.3 Audio Signals

Audio signals convey affective information through explicit (linguistic) messages, and implicit (acoustic and prosodic) messages that reflect the way the words are spoken. There exist a number of works focusing on how to map audio expression to dimensional models. Cowie et al. used valence-activation space (similar to valence-arousal) to model and assess affect from speech [11, 12]. Scherer and colleagues have also proposed how to judge emotional effects on vocal expression, using the appraisal-based theory [31].

In terms of affect recognition from audio signals the most reliable finding is that pitch appears to be an index into arousal [7]. Another well-accepted finding is that mean of the fundamental frequency (F0), mean intensity, speech rate, as well as pitch range [46], "blaring" timbre [14] and high-frequency energy [85] are positively correlated with the arousal dimension. Shorter pauses and inter-breath stretches are indicative of higher activation [99].

There is relatively less evidence on the relationship between certain acoustic parameters and other affect dimensions such as valence and power. Vowel duration and power dimension in general, and lower F0 and high power in particular, appear

to have correlations. Positive valence seems to correspond to a faster speaking rate, less high-frequency energy, low pitch and large pitch range [85] and longer vowel durations. A detailed literature summary on these can be found in [87] and [88].

10.3.4 Bio Signals

The bio signals used for automatic measurement of affect are galvanic skin response that increases linearly with a person's level of arousal [9], electromyography (frequency of muscle tension) that is correlated with negatively valenced emotions [41], heart rate that increases with negatively valenced emotions such as fear, heart rate variability that indicates a state of relaxation or mental stress, and respiration rate (how deep and fast the breath is) that becomes irregular with more aroused emotions like anger or fear [9, 41].

Measurements recorded over various parts of the brain including the amygdala also enable observation of the emotions felt [79]. For instance, approach or withdrawal response to a stimulus is known to be linked to the activation of the left or right frontal cortex, respectively.

A number of studies also suggest that there exists a correlation between increased blood perfusion in the orbital muscles and stress levels for human beings. This periorbital perfusion can be quantified through the processing of thermal video (e.g., [102]).

10.4 Overview of the Current Technology

This section provides a brief summary of the current technology by describing how affective data are acquired and annotated, and how affect analysis in continuous space is achieved.

10.4.1 Data Acquisition and Annotation

Cameras are used for acquisition of face and bodily expressions, microphones are used for recording audio signals, and thermal (infrared) cameras are used for recording blood flow and changes in skin temperature. 3D affective body postures or gestures can alternatively be recorded by utilizing motion capture systems (e.g., [57, 58]). In such scenarios, the actor is dressed in a suit with a number of markers on the joints and body segments, while each gesture is captured by a number of cameras and represented by consecutive frames describing the position of the markers in the 3D space. This is illustrated in Fig. 10.2 (second and third rows).

In the bio signal research context, the subject being recorded usually wears a headband or a cap on which electrodes are mounted, a clip sensor, or touch type

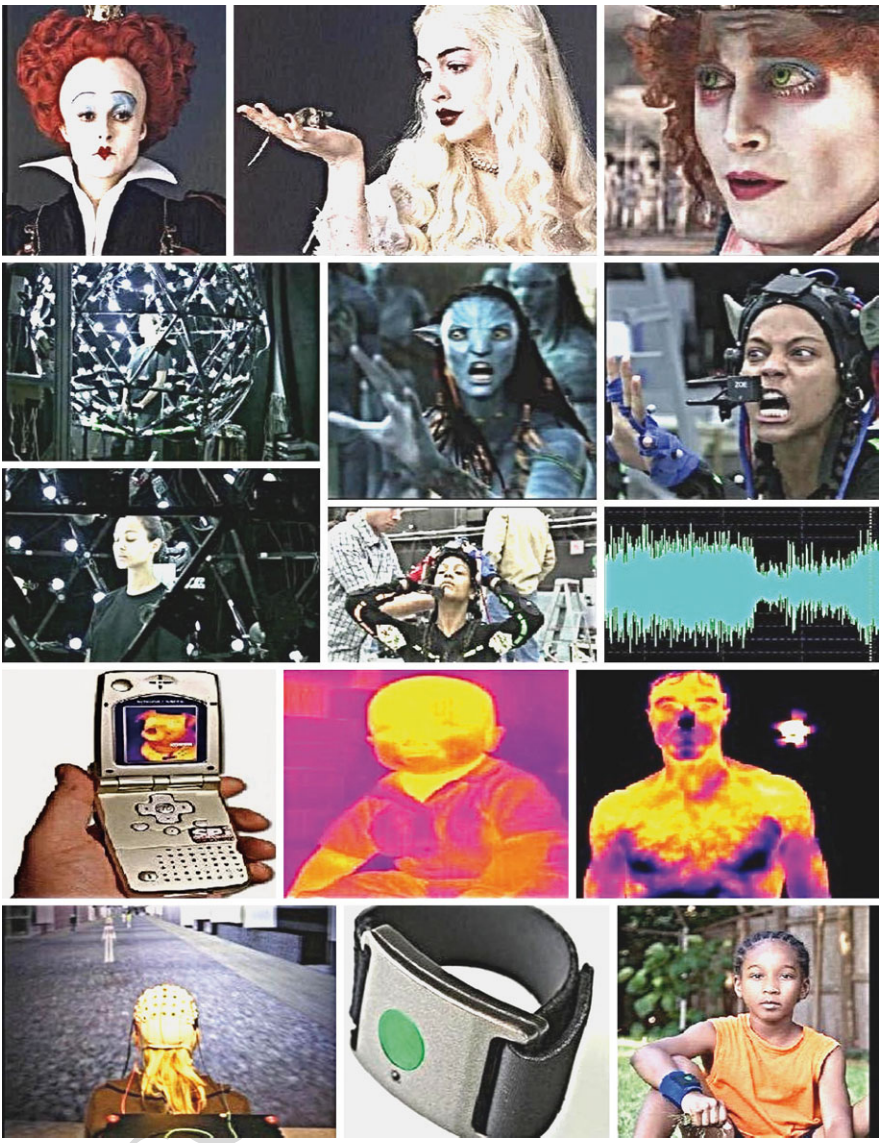


Fig. 10.2 Examples of sensors used in multimodal affective data acquisition: (*1st row*) camera for visible imagery (face and body), (*2nd & 3rd rows*) facial and body motion capture, and audio signals (used for animation and rendering), (*4th row*) infrared camera for thermal imagery, and (*5th row*) various means for recording bio signals (brain signals, heart and respiration rate, etc.)

electrodes (see Fig. 10.2, last row). The subject is then stimulated with emotionally-evocative images or sounds. Acquiring affect data without subjects' knowledge is strongly discouraged and the current trend is to record spontaneous data in more

constrained conditions such as an interview (e.g., [10]) or interaction (e.g., [62]) setting, where subjects are still aware of placement of the sensors and their locations.

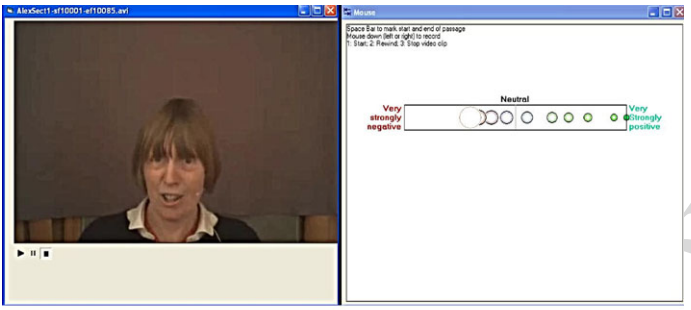
Annotation of the affect data is usually done separately for each modality, assuming independency between the modalities. A major challenge is the fact that there is no coding scheme that is agreed upon and used by all researchers in the field that can accommodate all possible communicative cues and modalities. In general, the Feeltrace annotation tool is used for annotating the external expressions (audio and visual signals) with continuous traces (impressions) in the dimensional space. Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, +1]$, to rate their impression about the emotional state of the subject [11] (see the illustration in Fig. 10.3(a)). For annotating the internal expressions (bio signals), the level of valence and arousal is usually extracted from subjective experiences (subjects' own responses) (e.g., [59, 79]) due to the fact that feelings, induced by an image or sound, can be very different from subject to subject. The Self Assessment Mannequin (SAM) [60], illustrated in Fig. 10.3(b), is the most widely used means for self assessment.

When discretized dimensional annotation is adopted (as opposed to continuous one), researchers seem to use different intensity levels: either a ten-point Likert scale (e.g., 0-low arousal, 9-high arousal) or a range between -1.0 and 1.0 (divided into a number of levels) [37]. The final annotation is usually calculated as the mean of the observers' ratings. However, whether this is the best way of obtaining ground-truth labels of emotional data is still being discussed. Overall, individual coders may vary in their appraisal of what is happening in the scene, in their judgment of the emotional behavior of the target individual, in their understanding of the terms 'positive emotion' and 'negative emotion' and in their movement of the computer mouse to translate their rating into a point on the onscreen scale. Furthermore, recent findings in dynamic emotional behavior coding indicate that the temporal pattern of ratings appears similar across cultures but that there exist significant differences in the intensity levels at which participants from different cultural backgrounds rate the emotional behaviors [96]. Therefore, how to obtain and use rich emotional data annotations, from multiple and multi-cultural raters, needs serious consideration.

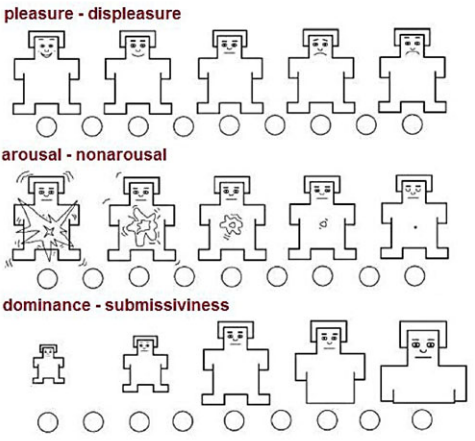
10.4.2 Automatic Dimensional Affect Prediction and Recognition

After affect data have been acquired and annotated, representative and relevant features need to be extracted prior to the automatic measurement of affect in dimensional and continuous space. The feature extraction techniques used for each communicative source are similar to the previous works (reviewed in [40]) adopting a categorical approach to affect recognition.

In dimensional affect analysis emotions are represented along a continuum. Considering this, systems that target automatic dimensional affect measurement should be able to predict the emotions continuously. However, most of the automatic recognition systems tend to simplify the problem by quantizing the continuous labels into



(a)



(b)

Fig. 10.3 Illustration of (**a**) the Feeltrace annotation tool [11], and (**b**) the Self Assessment Mannequin (SAM) [60]

a finite number of discrete levels. Hence, the most commonly employed strategy in automatic dimensional affect prediction is to reduce the continuous prediction problem to a two-class recognition problem (positive vs. negative or active vs. passive classification; e.g., [66, 92]) or a four-class recognition problem (classification into the quadrants of 2D V-A space; e.g., [8, 26, 29, 47, 106]).

For example, Kleinsmith and Bianchi-Berthouze discriminate between high–low, high–neutral and low–neutral affective dimensions [57], while Wöllmer et al. quantize the V-A dimensions of the SAL database into either 4 or 7 levels, and then use Conditional Random Fields (CRFs) to predict the quantized labels [105]. Attempts for discriminating between more coarse categories, such as positive vs. negative [66], and active vs. passive [8] have also been attempted. Of these, Caridakis et al. [8] uses the SAL database, combining auditive and visual modalities. Nicolaou et al. focus on audiovisual classification of spontaneous affect into negative or positive emotion categories using facial expression, shoulder and audio cues, and utilizing 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to

fuse multiple cues and modalities [66]. Kanluan et al. combine audio and visual cues for affect recognition in V-A space by fusing facial expression and audio cues, using Support Vector Machines for Regression (SVR) and late fusion with a weighted linear combination [50]. The labels used have been discretized on a 5-point scale in the range of $[-1, +1]$ for each emotion dimension. The work presented in [106] utilizes a hierarchical dynamic Bayesian network combined with BLSTM-NN performing regression and quantizing the results into four quadrants (after training).

As far as actual continuous dimensional affect prediction (without quantization) is concerned, there exist a number of methods that deal exclusively with speech (i.e., [33, 105, 106]). The work by Wöllmer et al. uses the SAL Database and Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [105]. Grimm and Kroschel use the Vera am Mittag database [35] and SVRs, and compare their performance to that of the distance-based fuzzy k-Nearest Neighbor and rule-based fuzzy-logic estimators [33]. The work by Espinosa et al. also use the Vera am Mittag database [35] and examine the importance of different groups of speech acoustic features in the estimation of continuous PAD dimensions [19].

Currently, there are also a number of works focusing on dimensional and continuous prediction of emotions from the visual modality [39, 56, 69]. The work by Gunes and Pantic focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject using SVRs [39]. Kipp and Martin in [56] investigated (without performing automatic prediction) how basic gestural form features (e.g., preference for using left/right hand, hand shape, palm orientation, etc.) are related to the single PAD dimensions of emotion. The work by Nicolaou et al. focuses on dimensional and continuous prediction of emotions from naturalistic facial expressions within an Output-Associative Relevance Vector Machine (RVM) regression framework by learning non-linear input and output dependencies inherent in the affective data [69].

More recent works focus on dimensional and continuous prediction of emotions from multiple modalities. For instance, Eyben et al. [21] propose a string-based approach for fusing the behavioral events from visual and auditive modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal audio cues) to predict human affect in a continuous dimensional space (in terms of arousal, expectation, intensity, power and valence dimensions). Although automatic affect analyzers based on physiology end up using multiple signal sources, explicit fusion of multimodal data for continuous modeling of affect utilizing dimensional models of emotion is still relatively unexplored. For instance, Khalili and Moradi propose multimodal fusion of brain and peripheral signals for automatic recognition of three emotion categories (positively excited, negatively excited and calm) [52]. Their results show that, for the task at hand, EEG signals seem to perform better than other physiological signals, and nonlinear features lead to better understanding of the felt emotions. Another representative approach is that of Gilroy et al. [28] that propose a dimensional multimodal fusion scheme based on the power-arousal-PAD space to support detection and integration of spontaneous affective behavior of users (in

terms of audio, video and attention events) experiencing arts and entertainment. Unlike many other multimodal approaches (e.g., [8, 50, 66]), the ground truth in this work is obtained by measuring Galvanic Skin Response (GSR) as an independent measure of arousal.

For further details on the aforementioned systems, as well as on systems that deal with dimensional affect recognition from a single modality or cue, the reader is referred to [37, 38, 109].

10.4.3 Challenges and Prospects

The summary provided in the previous section reflects that automatic dimensional affect recognition is still in its pioneering stage [34, 37, 38, 91, 105]. There are a number of challenges which need to be taken into account when applying a dimensional approach to affect prediction and advancing the current state of the art.

The interpretation accuracy of expressions and physiological responses in terms of continuous emotions is very challenging. While visual signals appear to be better for interpreting valence, audio signals seem to be better for interpreting arousal [33, 68, 100, 105]. A thorough comparison between all modalities would indeed provide a better understanding of which emotion dimensions are better predicted from which modalities (or cues).

Achieving inter-observer agreement is one of the most challenging issues in dimension-based affect modeling and analysis. To date, researchers have mostly chosen to use self-assessments (subjective experiences, e.g. [41]) or the mean (within a predefined range of values) of the observers' ratings (e.g. [57]). Although it is difficult to self-assess arousal, it has been reported that using classes generated from self-assessment of emotions facilitate greater accuracy in recognition (e.g., [9]). This finding results from a study on automatic analysis of physiological signals in terms of A-V emotion space. It remains unclear whether the same holds independently of the utilized modalities and cues. Modeling inter-observer agreement levels within automatic affect analyzers and finding which signals better correlate with self assessment and which ones better correlate with independent observer assessment remain unexplored.

The window size to be used to achieve optimal affect prediction is another issue that the existing literature does not provide a unique answer to. Current affect analyzers employ various window sizes depending on the modality, e.g., 2–6 seconds for speech, 3–15 seconds for bio signals [54]. For instance, when measuring affect from heart rate signals, analysis should not be done on epochs of less than a minute [6]. A time window of 50 s appears to be also necessary to accurately monitor mental stress in realistic settings [83]. There is no consensus on how the efficiency of such a choice should be evaluated. On one hand achieving real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds, e.g. [10]), while on the other hand obtaining a reliable prediction accuracy requires long(er)-term monitoring [6, 83]. For instance, Chanel et al. [10] conducted

short-term analysis of emotions (i.e., time segments of 8 s) in valence and arousal space using EEG and peripheral signals in a self-induction paradigm. They reported large differences in accuracy between the EEG and peripheral features which may be due to the fact that the 8 s length of trials may be too short for a complete activation of peripheral signals while it may be sufficient for EEG signals.

Measuring the intensity of expressed emotion appears to be modality dependent. The way the intensity of an emotion is apparent from physiological data may be different from the way it is apparent from visual data. Moreover, little attention has been paid so far to whether there are definite boundaries along the affect continuum to distinguish between various levels or intensities. Currently intensity is measured by quantizing the affect dimensions into arbitrary number of levels such as neutral, low and high (e.g., [57, 59, 105]). Separate models are then built to discriminate between pairs of affective dimension levels, for instance, low vs. high, low vs. neutral, etc. Generalizing intensity analysis across different subjects is a challenge yet to be researched as different subjects express different levels of emotions in the same situation. Moreover, recent research findings indicate that there also exist significant differences in the intensity levels at which coders from different cultural backgrounds rate emotional behaviors [96].

The Baseline problem is another major challenge in the field. For physiological signals (bio signals) this refers to the problem of finding a condition against which changes in measured physiological signals can be compared (a state of calmness) [65]. For the audio modality this is usually achieved by segmenting the recordings into turns using energy based voice activity detection and processing each turn separately (e.g., [105]). For visual modality the aim is to find a frame in which the subject is expressionless and against which changes in subject's motion, pose, and appearance can be compared. This is achieved by manually segmenting the recordings, or by constraining the recordings to have the first frame containing a neutral expression (see, e.g., [66, 67, 75]). Yet, as pointed out by Levenson in [61], emotion is rarely superimposed upon a prior state of *rest*; instead, emotion occurs most typically when the organism is in some prior activation. Hence, enforcing existence of expressionless state in each recording or manually segmenting recordings so that each segment contains a baseline expression are strong, unrealistic constrains. This remains a great challenge in automatic analysis, which typically relies on existence of a baseline for analysis and processing of affective information.

Generalization capability of automatic affect analyzers across subjects is still a challenge in the field. Kulic and Croft [59] reported that for bio signal based affect measurement, subjects seem to vary not only in terms of response amplitude and duration, but for some modalities, a number of subjects show no response at all. This makes generalization over unseen subjects a very difficult problem. A common way of measuring affect from bio signals is doing it for each participant separately (without computing baseline), e.g. [10]. When it comes to other modalities, most of the works in the field report mainly on subject-dependent dimensional affect measurement and recognition due to limited number of subjects and limited amount of data (e.g., [39, 68, 69, 105]).

Modality fusion refers to combining and integrating all incoming unimodal events into a single representation of the affect expressed by the user. When it

comes to integrating multiple modalities, the major issues are: (i) when to integrate the modalities (at what abstraction level to do the fusion), (ii) how to integrate the modalities (which criteria to use), (iii) how to deal with the increased number of features due to fusion, (iv) how to deal with the asynchrony between the modalities (e.g., if video is recorded at 25 Hz, audio is recorded at 48 kHz while EEG is recorded at 256–512 Hz), and (v) how to proceed with fusion when there is conflicting information conveyed by the modalities. Typically, multimodal data fusion is either done at the feature level (in a maximum likelihood estimation manner) or at the decision level (when most of the joint statistical properties may have been lost). Feature-level fusion is obtained by concatenating all the features from multiple cues into one feature vector which is then fed into a machine learning technique. In the decision-level data fusion, the input coming from each modality/cue is modeled independently, and these single-cue and single-modality based recognition results are combined in the end. Since humans display multi-cue and multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e. mutual correlation between the modalities). Therefore, model-level fusion has been proposed as an alternative approach for fusing multimodal affect data (e.g., [75]). Despite such efforts in the discrete affect recognition field (reviewed in [40, 109]), these issues remain yet to be explored for dimensional and continuous affect prediction.

Machine learning techniques used for dimensional and continuous affect measurement should be able to produce continuous values for the target dimensions. Overall, there is no agreement on how to model dimensional affect space (continuous vs. quantized) and which machine learning technique is better suited for automatic, multimodal, continuous affect analysis using a dimensional representation. Recognition of quantized dimensional labels is obtained via classification while continuous prediction is achieved by regression. Conditional Random Fields (CRF) and Support Vector Machines (SVM) have mostly been used for quantized dimensional affect recognition tasks (e.g., [105]). Some of the schemes that have been explored for the task of prediction are Support Vector Machines for Regression (SVR) (e.g., [39]) and Long Short-Term Memory Recurrent Networks (LSTM-RNN). The design of emotion-specific classification schemes that can handle multimodal and spontaneous data is one of the most important issues in the field. In accordance with this, Kim and Andre propose a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) using the property of the dichotomous categorization in the 2D emotion model and the fact that arousal classification yields a higher correct classification ratio than valence classification (or direct multiclass classification) [55]. They apply this scheme on classification of four emotions (positive/high arousal, negative/high arousal, negative/low arousal and positive/low arousal) from physiological signals recorded while subjects were listening to music. How to create such emotion-specific schemes for dimensional and continuous prediction of emotions from other modalities and cues should be investigated further.

Evaluation measures applicable to categorical affect recognition are not directly applicable to dimensional approaches. Using the Mean Squared Error (MSE) between the predicted and the actual values of arousal and valence, instead of the

recognition rate (i.e., percentage of correctly classified instances) is the most commonly used measure by related work in the literature (e.g., [50, 105]). However, using MSE might not be the best way to evaluate the performance of dimensional approaches to automatic affect measurement and prediction. Therefore, the correlation coefficient that evaluates whether the model has managed to capture patterns inhibited in the data at hand is also employed by several studies (e.g., [50, 67]) together with MSE. Overall, however, how to obtain optimal evaluation metrics for continuous and dimensional emotion prediction remains an open research issue [37]. Generally speaking, the performance of an automatic analyzer can be modeled and evaluated in an *intrinsic* and an *extrinsic* manner (as proposed for face recognition in [103]). The intrinsic performance and its evaluation depend on the intrinsic components such as the dataset chosen for the experiments and the machine learning algorithms (and their parameters) utilized for prediction. The extrinsic performance and evaluation instead depend on the extrinsic factors such as (temporal/spatial) resolution of the multimodal data and recording conditions (e.g., illumination, occlusions, noise, etc.). Future research in continuous affect prediction should analyze the relevance and prospects of the aforementioned performance components, and how they could be applied to continuous prediction of affect.

10.4.4 Applications

Various applications have been using the dimensional (both quantized and continuous) representation and prediction of emotions, ranging from human–computer (e.g., Sensitive Talking Heads [45], Sensitive Artificial Listeners [89, 90], spatial attention analysis [95], arts installations [104]) and human–robot interaction (e.g., humanoid robotics [5, 51]), clinical and biomedical studies (e.g., stress/pain monitoring [36, 64, 101], autism-related assistive technology), learning and driving environments (e.g., episodic learning [22], affect analysis in the car [20]), multimedia (e.g, video content representation and retrieval [53, 98] and personalized affective video retrieval [97]), and entertainment technology (e.g., gaming [80]). These indicate that affective computing has matured enough to have a presence and measurable impact in our lives. There are also spin off companies emerging out of collaborative research at well-known universities (e.g., Affectiva [1] established by R. Picard and colleagues of MIT Media Lab).

10.5 A Representative System: Continuous Analysis of Affect from Voice and Face

The review provided in the previous sections indicates that currently there is a shift toward subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic settings, and toward multimodal analysis and recognition of

human affect. Converging with this shift, in this section we present a representative approach that: (i) fuses facial expression and audio cues for dimensional and continuous prediction of emotions in valence and arousal space, (ii) employs the bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) for the prediction task, and (iii) introduces an output-associative fusion framework that incorporates correlations between the emotion dimensions to further improve continuous prediction of affect.

The section starts with the description of the naturalistic database used in the experimental studies. Next, data pre-processing, audio and facial feature extraction and tracking procedures, as well as the affect prediction process are explained.

10.5.1 Dataset

We use the Sensitive Artificial Listener Database (SAL-DB) [16] that contains spontaneous data collected with the aim of capturing the audiovisual interaction between a human and an operator undertaking the role of a SAL character (e.g., an avatar). The SAL characters intend to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. Each character has its own emotionally defined personality: Poppy is happy, Obadiah is gloomy, Spike is angry, and Prudence is pragmatic. During an interaction, the characters attempt to create an emotional workout for the user by drawing her/him toward their dominant emotion, through a combination of verbal and nonverbal expressions.

The SAL database contains audiovisual sequences recorded at a video rate of 25 fps (352×288 pixels) and at an audio rate of 16 kHz. The recordings were made in a lab setting, using one camera, a uniform background and constant lighting conditions. The SAL data have been annotated manually. Although there are approximately 10 hours of footage available in the SAL database, V-A annotations have only been obtained for two female and two male subjects. We used this portion for our experiments.

10.5.2 Data Pre-processing and Segmentation

The data pre-processing and segmentation stage consists of (i) determining ground truth by maximizing inter-coder agreement, (ii) detecting frames that capture the transition *to* and *from* an emotional state, and (iii) automatic segmentation of spontaneous audiovisual data. We provide a brief summary of these in the following sections. For a detailed description of these procedures the reader is referred to [67].

10.5.2.1 Annotation Pre-processing

The SAL data have been annotated by a set of coders who provided continuous annotations with respect to valence and arousal dimensions using the Feeltrace an-

notation tool [11], as explained in Sect. 10.4.1. Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, +1]$, to rate their impression about the emotional state of the subject.

Annotation pre-processing involves dealing with the issue of missing values (interpolation), grouping the annotations that correspond to one video frame together (binning), determining normalization procedures (normalization) and extracting statistics from the data in order to obtain segments with a baseline and high inter-coder agreement (statistics and metrics).

Interpolation In order to deal with the issue of missing values, similar to other works reporting on data annotated in continuous dimensional spaces (e.g., [105]), we interpolated the actual annotations at hand. We used piecewise cubic interpolation as it preserves the monotonicity and the shape of the data.

Binning Binning refers to grouping and storing the annotations together. As a first step the measurements of each coder c are binned separately. Since we aim at segmenting video files, we generate bins which are equivalent to one video frame f. This is equivalent to a bin of 0.04 seconds (SAL-DB was recorded at a rate of 25 frames/s). The fields with no annotation are assigned a 'not a number' (NaN) identifier.

Normalization The A-V measurements for each coder are not in total agreement, mostly due to the variance in human coders' perception and interpretation of emotional expressions. Thus, in order to deem the annotations comparable, we need to normalize the data. We experimented with various normalization techniques. After extracting the videos and inspecting the superimposed ground-truth plots, we opted for local normalization (normalizing each coder file for each session). This helps us avoid propagating noise in cases where one of the coders is in large disagreement with the rest (where a coder has a very low correlation with respect to the rest of the coders). Locally normalizing to zero mean produces the smallest mean squared error (MSE) both for valence (0.046) and arousal (0.0551) dimensions.

Statistics and Metrics We extract two useful statistics from the annotations: correlation and agreement. We start the analysis by constructing vectors of pairs of coders that correspond to each video session, e.g., when we have a video session where four coders have provided annotations, this gives rise to six pairs. For each of these pairs we extract the correlation coefficient between the valence (*val*) values of each pair, as well as the level of agreement in emotion classification in terms of positive or negative. We define the agreement metric by

$$AGR = \frac{\sum_{f=0}^{n} e(c_i(f).val, c_j(f).val)}{|frames|}, \tag{10.1}$$

where $c_i(f).val$ stands for the valence value annotated by coder c_i at frame f. Function e is defined as

$$e(i,j) = \begin{cases} 1 & \text{if } (sign(i) = sign(j)), \\ 0 & \text{else.} \end{cases}$$

In these calculations we do not consider the NaN values to avoid negatively affecting the results. After these metrics are calculated for each pair, each coder is assigned the average of the results of all pairs that the coder has participated in. We choose the Pearson's Correlation (COR) as the metric to be used in the automatic segmentation process as it appears to be stricter than agreement (AGR) providing better comparison amongst the coders.

10.5.2.2 Automatic Segmentation

The segmentation stage consists of producing negative and positive audiovisual segments with a temporal window that contains an offset before and after (i.e., the baseline) the displayed expression. For instance, for capturing negative emotional states, if we assume that the transition *from* non-negative *to* negative emotional state occurs at time t (in seconds), we would have a window of $[t-1, t, t', t'+1]$ where t' seconds is when the emotional state of the subject turns to non-negative again. The procedure is completely analogous for positive emotional states.

Detecting and Matching Crossovers For an input coder c, the crossing over from one emotional state to the other is detected by examining the valence values and identifying the points where the sign changes. Here a modified version of the sign function is used, it returns 1 for values that are higher than 0 (a value of 0 valence is never encountered in the annotations), -1 for values that are less than zero, and 0 for NaN values. We accumulate all crossover points for each coder, and return the set of crossovers *to-a-positive* and *to-a-negative* emotional state. The set of crossovers is then used for matching crossovers across coders. For instance, if a session has annotations from four coders, the frame (f) where each coder detects the crossover is not the same for all coders (for the session in question). Thus, we have to allow an offset for the matching process. This procedure searches the crossovers detected by the coders and then accepts the matches where there is less than the predefined offset (time) difference between the detections. When a match is found, we remove the matched crossovers and continue with the rest. The existence of different combinations of crossovers which may match using the predefined offset poses an issue. By examining the available datasets, we decided to maximize the number of coders participating in a matched crossover set rather than minimizing the temporal distances between the participating coders. The motivations for this decision are as follows: (i) if more coders agree on the crossover, the reliability of the ground truth produced will be higher, and (ii) the offset amongst the resulting matches is on average quite small (<0.5 s) when considering only the number of participating coders. We disregard cases where only one coder detects a crossover due to lack of agreement between coders.

Segmentation Driven by Matched Crossovers In order to illustrate how the crossover frame decision (for each member of the set) is made, let us assume that for *to-a-negative* transition a coder detects a crossover at frame 2, while the other coder detects a crossover at frame 4. If the frames are averaged to the nearest integer, then we can assume that the crossover happens at frame 3. In this case we have only 2 coders agreeing, we use the *correlation* metric in order to weight their decision and determine the crossover point. This provides a measurement of the relative importance of the annotations for each coder and propagates information from the other two coders not participating in the match. In order to capture 0.5 s before the transition window, the number of frames corresponding to the predefined offset are subtracted from the *start frame*. The ground-truth values for valence are retrieved by incrementing the initial frame number where each crossover was detected by the coders. Again, following the previous example, this means that we consider frame 2 of coder 1 and frame 4 of coder 2 to provide ground-truth values for frame 3 (the average of 2 and 4). This gives us an averaged valence value. Then, the frame 4 valence value (ground truth) would be the combination of frame 3 of coder 1 and frame 5 of coder 2. The procedure of determining combined average values continues until the valence value crosses again to a *non-negative* valence value. The endpoint of the audiovisual segment is then set to the frame including the offset after crossing back to a *non-negative* valence value. The ground truth of the audiovisual segment consists of the arousal and valence (A-V) values calculated.

Typically, an automatically produced segment or clip consists of a single interaction of the subject with the avatar (operator), starting with the final seconds of the avatar speaking, continuing with the subject responding (and thus reacting and expressing an emotional state audiovisually) and concluding where the avatar starts responding.

10.5.3 Feature Extraction

Our audio features include Mel-frequency Cepstrum Coefficients (MFCC) [49] and prosody features (the energy of the signal, the Root Mean Squared Energy and the pitch obtained by using a Praat pitch estimator [74]). Mel-frequency Cepstrum (MFC) is a representation of the spectrum of an audio sample which is mapped onto the nonlinear mel-scale of frequency to better approximate the human auditory system's response. The MFCC coefficients collectively make up the MFC for the specific audio segment. We used six cepstrum coefficients, thus obtaining six MFCC and six MFCC-Delta features for each audio frame. We have essentially used the typical set of features used for automatic affect recognition (e.g., [75]). Along with pitch, energy and RMS energy, we obtained a set of features with dimensionality $d = 15$ per audio frame. Note that we used a 0.04 second window with a 50% overlap (i.e. first frame 0–0.04, second from 0.02–0.06 and so on) in order to obtain a double frame rate for audio (50 Hz) compared to that of video (25 fps). This is an effective and straightforward way to synchronise the audio and video streams (similarly to [75]).

Fig. 10.4 Examples of the data at hand from the SAL database along with the extracted 20 points, used as features for the facial expression cues

To capture the facial motion displayed during a spontaneous expression we track 20 facial feature points (FFP), as illustrated in Fig. 10.4. These points are the corners of the eyebrows (4 points), eyes (8 points), nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the Patras–Pantic particle filtering tracking scheme [73]. For each video segment containing n frames, we obtain a set of n vectors containing 2D coordinates of the 20 points tracked in n frames ($Tr_f = \{Tr_{f1} \ldots Tr_{f20}\}$ with dimensions $n * 20 * 2$).

10.5.4 Dimensional Affect Prediction

This section describes how dimensional affect prediction from voice and face is achieved using the Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN). It first focuses on single-cue prediction from voice or face, and then introduces the model-level and output-associative fusion using the BLSTM-NNs.

10.5.4.1 Bidirectional Long Short-Term Memory Neural Networks

The traditional Recurrent Neural Networks (RNN) are unable to learn temporal dependencies longer than a few time steps due to the vanishing gradient problem [42, 43]. LSTM Neural Networks (LSTM-NNs) were introduced by Graves and Schmidhuber [32] in order to overcome this issue. The LSTM structure introduces recurrently connected memory blocks instead of traditional neural network nodes

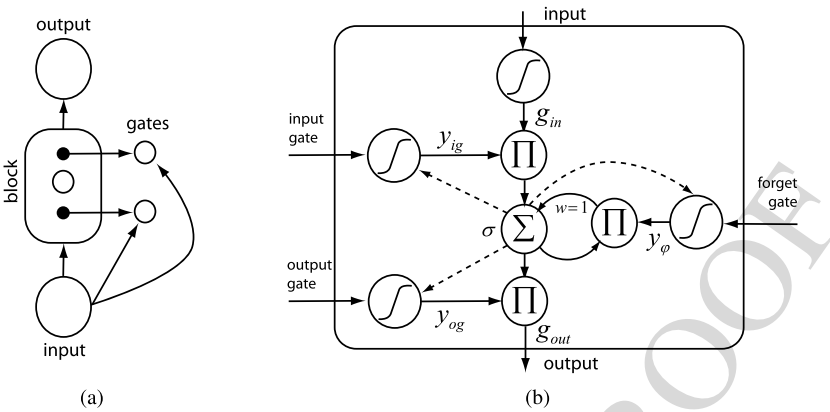


Fig. 10.5 Illustration of (**a**) the simplest LSTM network, with a single input, a single output, and a single memory block in place of the hidden unit, and (**b**) a typical implementation of an LSTM block, with multiplication units (Π), an addition unit (Σ) maintaining the cell state and typically non-linear squashing function units

(Fig. 10.5(a)). Each memory block contains memory cells and a set of multiplicative gates. In its simplest form, a memory block contains one memory cell.

As can be seen from Fig. 10.5(b), there are three types of gates: the input, output and forget gates. These gates are estimated during the training phase of an LSTM-NN.

The input, output and forget gates can be thought of as providing write, read and reset access to what is called a cell state (σ), which represents temporal network information. This can be seen from examining the state updates at time t:

$$\sigma(t) = y_{\phi}(t)\sigma(t-1) + y_{ig}(t)g_{in}(t).$$

The next state $\sigma(t)$ is defined as the sum of the forget gate at time t ($y_{\phi}(t)$) multiplied by the previous state, $\sigma(t-1)$ and the squashed input to the cell $g_{in}(t)$ multiplied by the input gate $y_{ig}(t)$. Thus, the forget gate can reset the state of the network, i.e. when $y_{\phi} \approx 0$ then the next state does not depend on the previous one:

$$\sigma(t) \approx y_{ig}(t)g_{in}(t).$$

This is similar when the input gate is near zero. Then, the next state depends only on the previous state and the forget gate value. The output of the cell is the cell state, as regulated by the value of the output gate (Fig. 10.5(b)). This configuration enforces constant error flow and overcomes the vanishing gradient problem.

In addition, traditional RNNs process input in a temporal order, thus learning input patterns by relating only to past context. Bidirectional RNNs (BRNNs) [3, 94] instead modify the learning procedure to overcome the latter issue of the past and future context: they present each of the training sequences in a forward and a backward order (to two different recurrent networks, respectively, which are connected to a common output layer). In this way, the BRNN is aware of both future and

past events in relation to the current timestep. The concept is directly expanded for LSTMs, referred to as Bidirectional Long Short-Term Memory neural networks (BLSTM-NN). BLSTM-NN have been shown to outperform unidirectional LSTM-NN for speech processing (e.g., [32]) and have been used for many learning tasks. They have been successfully applied to continuous emotion prediction from speech (e.g., [105, 106]) proving that modeling the sequential inputs and long range temporal dependencies appear to be beneficial for the task of automatic emotion prediction.

10.5.4.2 Single-Cue Prediction

The first step in continuous affect prediction task consists of prediction based on single cues. Let $\mathscr{D} = \{V, A\}$ represent the set of emotion dimensions, $\mathscr{C}$ the set of cues consisting of the facial expressions, shoulder movement and audio cues. Given a set of input features $\mathbf{x_c} = [\mathbf{x_{1_c}}, \ldots, \mathbf{x_{n_c}}]$ where n is the training sequence length and $c \in \mathscr{C}$, we train a machine learning technique f_d, in order to predict the relevant dimension output, $\mathbf{y_d} = [y_1, \ldots, y_n]$, $d \in \mathscr{D}$.

$$f_d : \mathbf{x} \mapsto y_d. \tag{10.2}$$

This step provides us with a set of predictions for each machine learning technique, and each relevant dimension employed.

10.5.4.3 Model-Level Fusion

As already explained in Sect. 10.4.2, since humans display multi-cue and multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e. mutual correlation between the modalities). Therefore, we opt for model-level fusion of the continuous predictions as this has the potential of capturing correlations and structures embedded in the continuous output of the predictors/regressors (from different sets of cues). This is illustrated in Fig. 10.6(a).

More specifically, during model-level fusion, a function learns to map predictions to a dimension d from the set of cues as follows:

$$f_{mlf} : f_d(\mathbf{x_1}) \times \cdots \times f_d(\mathbf{x_m}) \mapsto y_d, \tag{10.3}$$

where m is the total number of fused cues.

10.5.4.4 Output-Associative Fusion

In the previous section, we have treated the prediction of valence or arousal as a 1D regression problem. However, psychological evidence shows that valence and

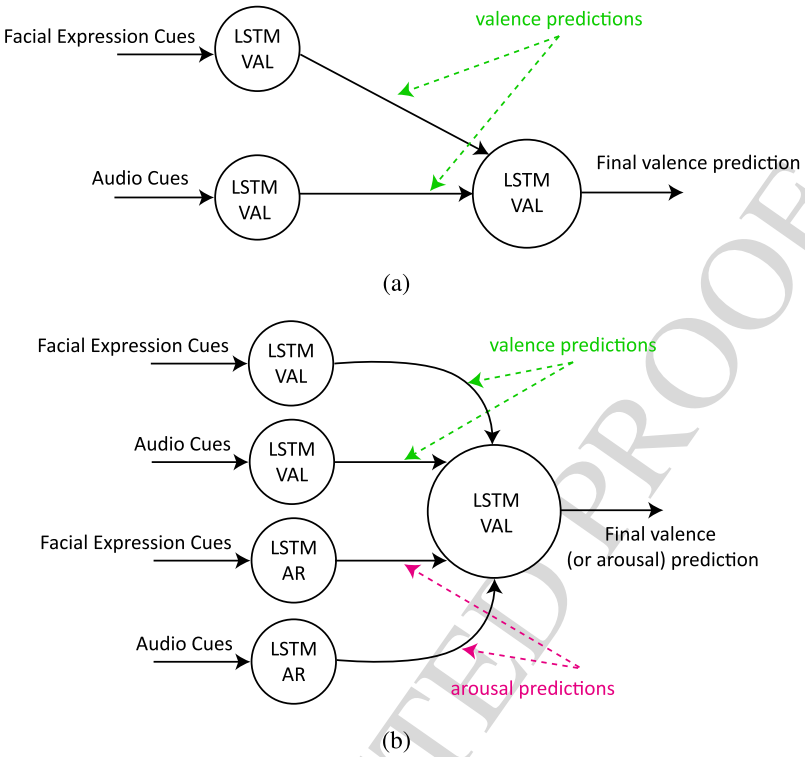


Fig. 10.6 Illustration of (**a**) model-level fusion and (**b**) output-associative fusion using facial expression and audio cues. Model-level fusion combines valence predictions from facial expression and audio cues by using a third network for the final valence prediction. Output-associative fusion combines both valence and arousal values predicted from facial expression and audio cues, again by using a third network, which outputs the final prediction.

arousal dimensions are correlated [2, 70, 107]. In order to exploit these correlations and patterns, we propose a framework capable of learning the dependencies that exist amongst the predicted dimensional values.

Given the setting described in Sect. 10.5.4.2, this framework learns to map the outputs of the intermediate predictors (each BLSTM-NN as defined in (10.2)) onto a higher (and final) level of prediction by incorporating cross-dimensional (output) dependencies (see Fig. 10.6(b)). This method, which we call *output-associative fusion*, can be represented by a function f_{oaf}:

$$f_{oaf} : f_{Ar}(\mathbf{x_1}) \times f_{Val}(\mathbf{x_1}) \times \cdots \times f_{Ar}(\mathbf{x_m}) \times f_{Val}(\mathbf{x_m}) \mapsto y_d . \tag{10.4}$$

As a result, the final output, taking advantage of the temporal and bidirectional characteristics of the regressors (BLSTM-NNs), depends not only on the entire sequence of input features $\mathbf{x_i}$ but also on the entire sequence of intermediate output predictions $\mathbf{f_d}$ of both dimensions (see Fig. 10.6(b)).

Table 10.1 Single-cue prediction results for valence and arousal dimensions

Dimension	Modality	RMSE	COR	SAGR
Arousal	Voice	0.240	0.586	0.764
	Face	0.250	0.493	0.681
Valence	Voice	0.220	0.444	0.648
	Face	0.170	0.712	0.841

10.5.5 Experiments and Analysis

10.5.5.1 Experimental Setup

Prior to experimentation, all features have been normalized to the range of $[-1, +1]$, except for the audio features which have been found to perform better with z-normalization (i.e., normalizing to mean $= 0$ and standard deviation $= 1$).

As the main evaluation metrics we choose to use the root mean squared error (RMSE) that evaluates the root of the prediction by taking into account the squared error of the prediction from the ground truth, the correlation (COR) that provides an evaluation of the linear relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture linear structural patterns inhibited in the data at hand, and the sign agreement metric (SAGR) that measures the agreement level of the prediction with the ground truth by assessing the valence dimension as being positive $(+)$ or negative $(-)$, and the arousal dimension as being active $(+)$ or passive $(-)$.

For validation purposes we use a subset of the SAL-DB that consists of 134 audiovisual segments (a total of 30,042 video frames) obtained by the automatic segmentation procedure (proposed in [67]). As V-A annotations have only been provided for two female and two male subjects, for our experiments we employ *subject-dependent leave-one-sequence-out cross-validation*. More specifically, the evaluation consists of 134 folds where at each fold one sequence is left out for testing and the other 133 sequences are used for training. The prediction results are then averaged over 134 folds.

The parameter optimization for BLSTM-NNs refers to mainly determining the topology of the network along with the number of epochs, momentum and learning rate.

10.5.5.2 Results and Analysis

Single-cue results are presented in Table 10.1, while results obtained from fusion are presented in Table 10.2.

We initiate our analysis with the single-cue results (Table 10.1) and the valence dimension. Various automatic dimensional emotion prediction and recognition studies have shown that arousal can be much better predicted than valence using audio

Table 10.2 Results for output-associative fusion (AOF) and model-level fusion (MLF). The best results are obtained by employing output-associative fusion (shown in bold)

Dimension	OAF			MLF		
	RMSE	COR	SAGR	RMSE	COR	SAGR
Arousal	**0.220**	**0.628**	0.800	0.230	0.605	0.800
Coders	0.145	0.870	0.840	0.145	0.870	0.840
Valence	**0.160**	**0.760**	**0.892**	0.170	0.748	0.856
Coders	0.141	0.850	0.860	0.141	0.850	0.860

cues (e.g., [33, 68, 100, 105]). Our experimental results also support these findings indicating that the visual cues appear more informative for predicting the valence dimension. The facial expression cues provide a higher correlation with the ground truth (COR = 0.71) compared to the audio cues (COR = 0.44). This fact is also confirmed by the RMSE and SAGR metrics. The facial expression cues also provide higher SAGR (0.84), indicating that the predictor was accurate in predicting an emotional state as positive or negative for 84% of the frames. For prediction of the arousal dimension the audio cues appear to be superior to the visual cues. More specifically, audio cues provide COR = 0.59, whereas the facial expression cues provide COR = 0.49.

Fusing facial and audio cues using model-level fusion outperforms the single-cue prediction results. Model-level fusion appears to be much better for predicting the valence dimension rather than the arousal dimension. This is mainly due to the fact that the single-cue predictors for valence dimension perform better, thus containing more correct temporal dependencies and structural characteristics (while the weaker arousal predictors contain fewer of these dependencies). Model-level fusion also reconfirms that visual cues are more informative for valence dimension than the audio cues. Finally, the newly proposed output-associative fusion provides the best results, outperforming both single-cue analysis and model-level fusion results. We denote that the performance increase of output-associative fusion is higher for the arousal dimension (compared to the valence dimension). This could be justified by the fact that the single-cue predictors for valence perform better than for arousal (Table 10.1) and thus, more correct valence patterns are passed onto the output-associative fusion framework. An example of the output-associative valence and arousal prediction from face and audio is shown in Fig. 10.7.

Based on the experimental results provided in Tables 10.1–10.2, we conclude the following.

- Facial expression cues are better suited to the task of continuous valence prediction compared to audio cues. For arousal dimension, instead, the audio cues appear to perform better. This is in accordance with the previous findings in the literature.
- The inherent temporal and structured nature of continuous affective data appears to be highly suitable for predictors that can model temporal dependencies and relate temporally distant events. To evaluate the performance of such frameworks,

the use of not only the RMSE but also the correlation coefficient appears to be very important. Furthermore, the use of other emotion-specific metrics, such as the SAGR (used in this work), is also desirable as they contain valuable information regarding emotion-specific aspects of the predictions.

- As confirmed by the psychological theory, valence and arousal are correlated. Such correlations appear to exist in our data where fusing predictions from both valence and arousal dimensions (output-associative fusion) improves the results compared to using predictions from either valence or arousal dimension alone (as in the model-level fusion case).
- In general, audiovisual data appear to be more useful for predicting valence than for predicting arousal. While arousal is better predicted by using audio features alone, valence is better predicted by using audiovisual data.

Overall, our output-associative fusion framework (i) achieves RMSE $= 0.160$, COR ≈ 0.760 and SAGR ≈ 0.900 for the valence dimension, compared to the human coder (inter-coder) RMSE ≈ 0.141, COR ≈ 0.850, and SAGR ≈ 0.860, and (ii) provides RMSE $= 0.220$, COR ≈ 0.628 and SAGR ≈ 0.800 for the arousal dimension, compared to the human coder (inter-coder) RMSE ≈ 0.145, COR ≈ 0.870 and SAGR ≈ 0.840.

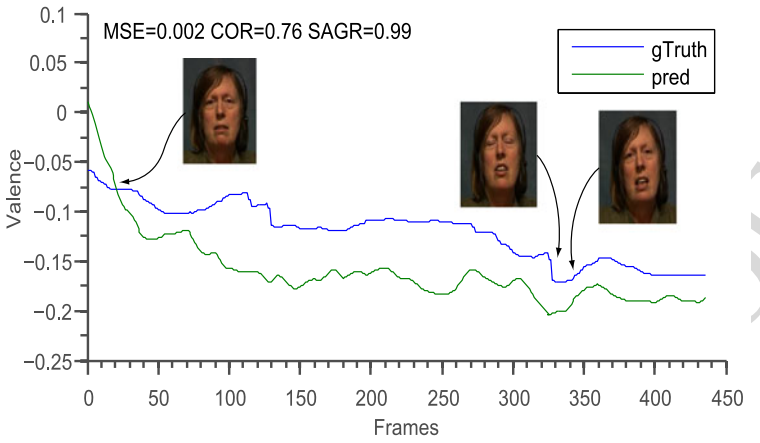
In our experiments we employed a subject-dependent leave-one-sequence-out cross-validation procedure due to the small number of annotated data available. As spontaneous expressions appear to have somewhat person-dependent characteristics, subject-independent experimentation is likely to be more challenging and affect our prediction results.

10.6 Concluding Remarks

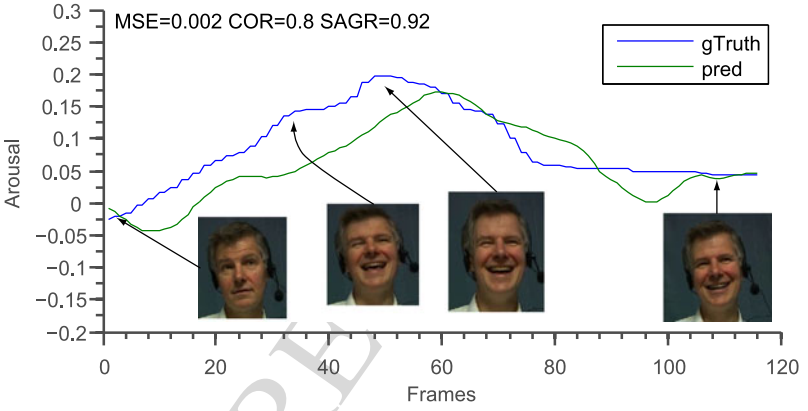
The review provided in this chapter suggests that the automatic affect sensing field has slowly started shifting from categorical (and discrete) affect recognition to dimensional (and continuous) affect prediction to be able to capture the complexity of affect expressed in naturalistic settings. There is a growing research interest driven by various advances and demands (e.g., real-time representation and analysis of naturalistic and continuous human affective behavior for emotion-related disorders like autism), and funded by various research projects (e.g., European Union FP 7, SEMAINE[1]). To date, despite the existence of a number of dimensional emotion models, the two-dimensional model of arousal and valence appears to be the most widely used model in automatic measurement of affect from audio, visual and bio signals.

The current automatic measurement technology has already started dealing with spontaneous data obtained in less-controlled environments using various sensing devices, and exploring a number of machine learning techniques and evaluation measures. However, naturalistic settings pose many challenges to continuous affect

[1] http://www.semaine-project.eu



(a)



(b)

Fig. 10.7 Valence and arousal ground truth (gTruth) compared to predictions (pred) from output-associative fusion of facial expressions and audio cues

sensing and prediction (e.g., when subjects are not restricted in terms of mobility, the level of noise in all recorded signals tends to increase), as well as affect synthesis and generation. As a consequence, a number of issues that should be addressed in order to advance the field remain unclear. These have been summarized and discussed in this chapter.

As summarized in Sect. 10.4.2 and reviewed in [37], to date, only a few systems have actually achieved dimensional affect prediction from multiple modalities. Overall, existing systems use different training/testing datasets (which differ in the way affect is elicited and annotated), they differ in the underlying affect model (i.e., target affect categories), as well as in the employed modality or combination of

modalities, and the applied evaluation method. As a consequence, it remains unclear which recognition and prediction method is suitable for dimensional affect prediction from which modalities and cues. These challenges should be addressed in order to advance the field while identifying the importance, as well as the feasibility, of the following issues:

1. *Among the available remotely observable and remotely unobservable modalities, which ones should be used for automatic dimensional affect prediction? Should we investigate the innate priority among the modalities to be preferred for each affect dimension? Does this depend on the context (who the subject is, where she is, what her current task is, and when the observed behavior has been shown)?*

   Continuous long-term monitoring of bio signals (e.g., autonomic nervous system) appears to be particularly useful and usable for health care applications (e.g., stress and pain monitoring, autism-related assistive technology). Using bio signals for automatic measurement is especially important for applications where people do not easily express themselves outwardly with facial and bodily expressions (e.g., people with autism spectrum disorders) [24]. As stated before, various automatic dimensional emotion prediction and recognition studies have shown that arousal can be much better predicted than valence using audio cues (e.g., [33, 68, 100, 105]). For the valence dimension instead, visual cues (e.g., facial expressions and shoulder movements) appear to perform better [68]. Whether such conclusions hold for different contexts and different data remains to be evaluated. Another significant research finding is that when multiple modalities are available during data annotation, both speed and accuracy of judgments increase when the modalities are expressing the same emotion [15]. How such findings should be incorporated into automatic dimensional affect predictors remains to be researched further.
2. *When labeling emotions, which signals better correlate with self assessment and which ones correlate with independent observer assessment?*

   When acquiring and annotating emotional data, there exist individual differences in emotional response, as well as individual differences in the use of rating scales. We have mentioned some of these differences before, in Sect. 10.4.1. Research also shows that affective state labeling is significantly affected by factors such as familiarity of the person and context of the interaction [44]. Even if the emotive patterns to be labeled are fairly similar, human perception is biased by context and prior experience. Moreover, Feldman presented evidence that when individuals are shown emotional stimulus, they differ in their attention to valence and arousal dimensions [23]. We have also mentioned cross-cultural intensity differences in labeling emotional behaviors [96]. If such issues are ignored and the ratings provided by the human annotators are simply averaged, the measure obtained may be useful in certain experimental contexts but it will be insensitive to individual variations in subjective experience. More specifically, this will imply having a scale that assumes that individual differences are unimportant or nonexistent. An implication of this view is that for an ideal representation of a subject's affective state, labeling schemes and rating scales should be clearly defined (e.g., by making the subjective distances between adjacent numbers on every portion

of the scale equal) and contextualized (e.g., holding the environmental cues constant), both self assessment and external observer assessment (preferably from observers who are familiar with the user to be assessed) should be obtained and used, and culture-related issues should be taken into consideration.

3. *How does the baseline problem affect prediction? Is an objective basis (e.g., a frame with an expressionless display) strictly needed prior to computing the dimensional affect values? If so, how can this be obtained in a fully automatic manner from naturalistic data?*

   Determining the baseline in naturalistic affective displays is challenging even for human observers. This is particularly the case for the visual modality which constitutes of varying head pose and head gestures (like nods and shakes), speech-related facial actions, and blended facial expressions. The implications for automatic analysis can initially be addressed by training predictors that predict baseline (or neutrality) for each cue and modality separately.

4. *How should intensity be modeled for dimensional and continuous affect prediction? Should the aim be personalizing systems for each subject, or creating systems that are expected to generalize across subjects?*

   Modeling the intensity of emotions should be based on the task-dependent environment and target user group. A common way of measuring affect from bio signals is doing it for each participant separately (without computing baseline), e.g., [10]. Similarly to the recent works on automatic affect prediction from the audio or the visual cues (e.g., [69]), better insight may be obtained by comparing subject-dependent vs. subject-independent prediction results. Customizing the automatic predictors to specific user needs is usually desired and advantageous.

5. *In a continuous affect space, how should duration of affect be defined? How can this be incorporated in automated systems? Will focusing on shorter or longer observations affect the accuracy of the measurement process?*

   Similarly to modeling the emotional intensity level, determining the affect duration should be done based on the task-dependent environment and target user group. Focusing on shorter or longer durations appears to have an effect on the prediction accuracy. Achieving real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds, e.g., [10]), while on the other hand obtaining a reliable prediction accuracy requires long(er)-term monitoring [6, 83]. Therefore, analysis duration should be determined as a trade-off between reliable prediction accuracy and real-time requirements of the automatic system.

Finding comprehensive and thorough answers to the questions posed above, and fully exploring the terrain of the dimensional and continuous affect prediction, depends on all relevant research fields (engineering, computer science, psychology, neuroscience, and cognitive sciences) stepping out of their labs, working side-by-side together on real-life applications, and sharing the experience and the insight acquired on the way, to make affect research tangible for realistic settings and lay people [76]. Pioneering projects representing such inter-disciplinary efforts have already started emerging, ranging, for instance, from publishing compiled books of related work (e.g., [30]) and organizing emotion recognition challenges (e.g., INTERSPEECH 2010 Paralinguistic Challenge featuring the affect sub-challenge

with a focus on dimensional affect [93]) to projects as varied as affective human-embodied conversational agent interaction (e.g., European Union FP 7 SEMAINE [89, 90]), and affect sensing for autism (e.g., [76, 78]).

10.7 Summary

Human affective behavior is multimodal, continuous and complex. Despite major advances within the affective computing research field, modeling, analyzing, interpreting and responding to human affective behavior still remains a challenge for automated systems as affect and emotions are complex constructs, with fuzzy boundaries and with substantial individual differences in expression and experience [7]. Therefore, affective and behavioral computing researchers have recently invested increased effort in exploring how to best model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum e.g., from -1 to $+1$) of affective behavior in terms of latent dimensions (e.g., arousal, power and valence) and appraisals, rather than in terms of a small number of discrete emotion categories (e.g., happiness and sadness). This chapter aimed to (i) give a brief overview of the existing efforts and the major accomplishments in modeling and analysis of emotional expressions in dimensional and continuous space while focusing on open issues and new challenges in the field, and (ii) introduce a representative approach for multimodal continuous analysis of affect from voice and face, and provide experimental results using the audiovisual Sensitive Artificial Listener (SAL) Database of natural interactions. The chapter concluded by posing a number of questions that highlight the significant issues in the field, and by extracting potential answers to these questions from the relevant literature.

10.8 Questions

1. What are the major approaches used for affect modeling and representation? How do they differ from each other?
2. Why has the dimensional affect representation gained interest?
3. What are the dimensions used for representing emotions?
4. Affect research scientists usually make a number of assumptions and simplifications while studying emotions. What are these assumptions and simplifications? What implications do they have?
5. How is human affect sensed and measured? What are the signals measured for analyzing human affect?
6. How are affective data acquired and annotated?
7. What is the current state of the art in automatic affect prediction and recognition?
8. What are the challenges faced in automatic dimensional affect recognition?
9. List a number of applications that use the dimensional representation of emotions.

10. What features are extracted to represent an audio-visual affective sequence? How are the audio and video streams synchronized?
11. What is a Bidirectional Long Short-Term Memory Neural Network? How does it differ from a traditional Recurrent Neural Network?
12. What is meant by the statement 'valence and arousal dimensions are correlated'? What implications does this have on automatic affect prediction?
13. What is output-associative fusion? How does it compare to model-level fusion?
14. How are the root mean squared error, correlation, and sign agreement used for evaluating the automatic prediction of emotions?

10.9 Glossary

- *Categorical description of affect.* Hypothesizes that there exist a small number of emotions categories (i.e., anger, disgust, fear, happiness, sadness and surprise) that are basic, hard-wired in our brain, and recognized universally (e.g. [18]).
- *Dimensional description of affect.* Hypothesizes that affective states are not independent from one another; rather, they are related to one another in a systematic manner.
- *Circumplex Model of Affect.* A circular configuration introduced by Russell [82], based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum.
- *PAD emotion space.* The three dimensional description of emotion in terms of pleasure–displeasure, arousal–nonarousal and dominance–submissiveness [63].
- *Dimensional and continuous affect prediction.* Analyzing and inferring the subtlety, complexity and continuity of affective behavior in terms of latent dimensions (e.g., valence and arousal) by representing it along a continuum (e.g., from -1 to $+1$) without discretization.
- *Long Short-Term Memory neural network.* A Bidirectional Recurrent Neural Network that consists of recurrently connected memory blocks, and uses input, output and forget gates to represent and learn the temporal information and dependencies.
- *Output-associative fusion.* A fusion approach that uses multi-layered prediction, i.e. the initial features extracted from each modality are used for intermediate (output) prediction, and these are further used for a higher (and final) level of prediction (by incorporating cross-dimensional dependencies).

Acknowledgements This work has been funded by EU [FP7/2007-2013] Grant agreement No. 211486 (SEMAINE) and the ERC Starting Grant agreement No. ERC-2007-StG-203143 (MAHNOB).

References

1. Affectiva's homepage: http://www.affectiva.com/ (2011)

2. Alvarado, N.: Arousal and valence in the direct scaling of emotional response to film clips. Motiv. Emot. **21**, 323–348 (1997)
3. Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., Soda, G.: Exploiting the past and the future in protein secondary structure prediction. Bioinformatics **15**, 937–946 (1999)
4. Bartneck, C.: Integrating the occ model of emotions in embodied characters. In: Proc. of the Workshop on Virtual Conversational Characters, pp. 39–48 (2002)
5. Beck, A., Canamero, L., Bard, K.A.: Towards an affect space for robots to display emotional body language. In: Proc. IEEE Int. Symp. in Robot and Human Interactive Communication, pp. 464–469 (2010)
6. Berntson, G.G., Bigger, J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H., van der Molen, M.W.: Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology **34**(6), 623 (1997)
7. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
8. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proc. of ACM Int. Conf. on Multimodal Interfaces, pp. 146–154 (2006)
9. Chanel, G., Ansari-Asl, K., Pun, T.: Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 2662–2667, October 2007
10. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum.-Comput. Stud. **67**(8), 607–627 (2009)
11. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of ISCA Workshop on Speech and Emotion, pp. 19–24 (2000)
12. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. **18**, 33–80 (2001)
13. Cowie, R., Gunes, H., McKeown, G., Vaclau-Schneider, L., Armstrong, J., Douglas-Cowie, E.: The emotional and communicative significance of head nods and shakes in a naturalistic database. In: Proc. of LREC Int. Workshop on Emotion, pp. 42–46 (2010)
14. Davitz, J.: Auditory correlates of vocal expression of emotional feeling. In: The Communication of Emotional Meaning, pp. 101–112. McGraw-Hill, New York (1964)
15. de Gelder, B., Vroomen, J.: The perception of emotions by ear and by eye. Cogn. Emot. **23**, 289–311 (2000)
16. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, L., McRorie, M., Martin, L. Jean-Claude, Devillers, J.-C., Abrilian, A., Batliner, S., Noam, A., Karpouzis, K.: The HUMAINE database: addressing the needs of the affective computing community. In: Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction, pp. 488–500 (2007)
17. Ekman, P., Friesen, W.V.: Head and body cues in the judgment of emotion: A reformulation. Percept. Mot. Skills **24**, 711–724 (1967)
18. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice Hall, New Jersey (1975)
19. Espinosa, H.P., Garcia, C.A.R., Pineda, L.V.: Features selection for primitives estimation on emotional speech. In: Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing, pp. 5138–5141 (2010)
20. Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., Nguyen-Thien, N.: Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. Adv. Hum.-Comput. Interact. **2010**, 263593 (2010), 17 pages
21. Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M.: String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)

22. Faghihi, U., Fournier-Viger, P., Nkambou, R., Poirier, P., Mayers, A.: How emotional mechanism helps episodic learning in a cognitive agent. In: Proc. IEEE Symp. on Intelligent Agents, pp. 23–30 (2009)
23. Feldman, L.: Valence focus and arousal focus: Individual differences in the structure of affective experience. J. Pers. Soc. Psychol. **69**, 153–166 (1995)
24. Fletcher, R., Dobson, K., Goodwin, M.S., Eydgahi, H., Wilder-Smith, O., Fernholz, D., Kuboyama, Y., Hedman, E., Poh, M.Z., Picard, R.W.: iCalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. IEEE Tran. on Information Technology in Biomedicine **14**(2), 215
25. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotion is not two-dimensional. Psychol. Sci. **18**, 1050–1057 (2007)
26. Fragopanagos, N., Taylor, J.G.: Emotion recognition in human–computer interaction. Neural Netw. **18**(4), 389–405 (2005)
27. Frijda, N.H.: The Emotions. Cambridge University Press, Cambridge (1986)
28. Gilroy, S.W., Cavazza, M., Niiranen, M., Andre, E., Vogt, T., Urbain, J., Benayoun, M., Seichter, H., Billinghurst, M.: Pad-based multimodal affective fusion. In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction Workshops, pp. 1–8 (2009)
29. Glowinski, D., Camurri, A., Volpe, G., Dael, N., Scherer, K.: Technique for automatic emotion recognition by body gesture analysis. In: Proc. of Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2008)
30. Gokcay, D., Yildirim, G.: Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives. IGI Global, Hershey (2011)
31. Grandjean, D., Sander, D., Scherer, K.R.: Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Conscious. Cogn. **17**(2), 484–495 (2008)
32. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005)
33. Grimm, M., Kroschel, K.: Emotion estimation in speech using a 3d emotion space concept. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 381–385 (2005)
34. Grimm, M., Mower, E., Kroschel, K., Narayanan, S.: Primitives based estimation and evaluation of emotions in speech. Speech Commun. **49**, 787–800 (2007)
35. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: ICME, pp. 865–868. IEEE Press, New York (2008)
36. Grundlehner, B., Brown, L., Penders, J., Gyselinckx, B.: The design and analysis of a real-time, continuous arousal monitor. In: Proc. Int. Workshop on Wearable and Implantable Body Sensor Networks, pp. 156–161 (2009)
37. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emot. **1**(1), 68–99 (2010)
38. Gunes, H., Pantic, M.: Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how. In: Proc. of Measuring Behavior, pp. 122–126 (2010)
39. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Proc. of International Conference on Intelligent Virtual Agents, pp. 371–377 (2010)
40. Gunes, H., Piccardi, M., Pantic, M.: Affective computing: focus on emotion expression, synthesis, and recognition. In: Or, J. (ed.) From the Lab to the Real World: Affect Recognition using Multiple Cues and Modalities, pp. 185–218. I-Tech Education and Publishing, Vienna (2008)
41. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First steps towards an automatic system. In: LNCS, vol. 3068, pp. 36–48 (2004)
42. Hochreiter, S.: Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München (1991)
43. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **6**(2), 107–116 (1998)

44. Hoque, M.E., El Kaliouby, R., Picard, R.W.: When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos. In: Proc. of Intelligent Virtual Agents, pp. 337–343 (2009)
45. Huang, T.S., Hasegawa-Johnson, M.A., Chu, S.M., Zeng, Z., Tang, H.: Sensitive talking heads. IEEE Signal Process. Mag. **26**, 67–72 (2009)
46. Huttar, G.L.: Relations between prosodic variables and emotions in normal American english utterances. J. Speech Hear. Res. **11**, 481–487 (1968)
47. Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., Kollias, S.: Emotion recognition through facial expression analysis based on a neurofuzzy method. Neural Netw. **18**(4), 423–435 (2005)
48. Jia, J., Zhang, S., Meng, F., Wang, Y., Cai, L.: Emotional audio-visual speech synthesis based on PAD. IEEE Trans. Audio Speech Lang. Process. **PP**(9), 1 (2010)
49. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edn. Prentice-Hall, New York (2008)
50. Kanluan, I., Grimm, M., Kroschel, K.: Audio-visual emotion recognition using an emotion recognition space concept. In: Proc. of the 16th European Signal Processing Conference (2008)
51. Karg, M., Schwimmbeck, M., Kühnlenz, K., Buss, M.: Towards mapping emotive gait patterns from human to robot. In: Proc. IEEE Int. Symp. in Robot and Human Interactive Communication, pp. 258–263 (2010)
52. Khalili, Z., Moradi, M.H.: Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of EEG. In: Proc. Int. Joint Conf. on Neural Networks, pp. 1571–1575 (2009)
53. Kierkels, J.J.M., Soleymani, M., Pun, T.: Queries and tags in affect-based multimedia retrieval. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 1436–1439 (2009)
54. Kim, J.: Robust speech recognition and understanding. In: Grimm, M., Kroschel, K. (eds.) Bimodal Emotion Recognition using Speech and Physiological Changes, pp. 265–280. I-Tech Education and Publishing, Vienna (2007)
55. Kim, J., Andre, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2067–2083 (2008)
56. Kipp, M., Martin, J.-C.: Gesture and emotion: Can basic gestural form features discriminate emotions? In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction Workshops, pp. 1–8 (2009)
57. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing affective dimensions from body posture. In: Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction, pp. 48–58 (2007)
58. Kleinsmith, A., De Silva, P.R., Bianchi-Berthouze, N.: Recognizing emotion from postures: Cross–cultural differences in user modeling. In: Proc. of the Conf. on User Modeling, pp. 50–59 (2005)
59. Kulic, D., Croft, E.A.: Affective state estimation for human-robot interaction. IEEE Trans. Robot. **23**(5), 991–1000 (2007)
60. Lang, P.J.: The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders. Erlbaum, Hillside (1985)
61. Levenson, R.: Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In: Social Psychophysiology and Emotion: Theory and Clinical Applications, pp. 17–42 (1988)
62. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Proc. of IEEE Int'l Conf. Multimedia, Expo (ICME'10), pp. 1079–1084, July 2010
63. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Curr. Psychol. **14**, 261–292 (1996)
64. Mihelj, M., Novak, D., Munih, M.: Emotion-aware system for upper extremity rehabilitation. In: Proc. Int. Conf. on Virtual Rehabilitation, pp. 160–165 (2009)

65. Nakasone, A., Prendinger, H., Ishizuka, M.: Emotion recognition from electromyography and skin conductance. In: Proc. of the 5th International Workshop on Biosignal Interpretation, pp. 219–222 (2005)
66. Nicolaou, M.A., Gunes, H., Pantic, M.: Audio-visual classification and fusion of spontaneous affective data in likelihood space. In: Proc. of IEEE Int. Conf. on Pattern Recognition, pp. 3695–3699 (2010)
67. Nicolaou, M.A., Gunes, H., Pantic, M.: Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In: Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 43–48 (2010)
68. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence–arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
69. Nicolaou, M.A., Gunes, H., Pantic, M.: Output-associative RVM regression for dimensional and continuous emotion prediction. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)
70. Oliveira, A.M., Teixeira, M.P., Fonseca, I.B., Oliveira, M.: Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. In: Proc. of the 22nd Annual Meeting of the Int. Society for Psychophysics, pp. 245–250 (2006)
71. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
72. Parkinson, B.: Ideas and Realities of Emotion. Routledge, London (1995)
73. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 97–102 (2004)
74. Paul, B.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences, pp. 97–110 (1993)
75. Petridis, S., Gunes, H., Kaltwang, S., Pantic, M.: Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In: Proc. of ACM Int. Conf. on Multimodal Interfaces, pp. 23–30 (2009)
76. Picard, R.W.: Emotion research by the people, for the people. Emotion Review **2**(3), 250–254
77. Plutchik, R., Conte, H.R.: Circumplex Models of Personality and Emotions. APA, Washington (1997)
78. Poh, M.Z., Swenson, N.C., Picard, R.W.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Trans. Inf. Technol. Biomed. **57**(5), 1243–1252 (2010)
79. Pun, T., Alecu, T.I., Chanel, G., Kronegg, J., Voloshynovskiy, S.: Brain–computer interaction research at the Computer Vision and Multimedia Laboratory, University of Geneva. IEEE Trans. Neural Syst. Rehabil. Eng. **14**, 210–213 (2006)
80. Rehm, M., Wissner, M.: Gamble-a multiuser game with an embodied conversational agent. In: Lecture Notes in Computer Science, vol. 3711, pp. 180–191 (2005)
81. Roseman, I.J.: Cognitive determinants of emotion: A structural theory. In: Shaver, P. (ed.) Review of Personality & Social Psychology, Beverly Hills, CA, vol. 5, pp. 11–36. Sage, Thousand Oaks (1984)
82. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
83. Salahuddin, L., Cho, J., Jeong, M.G., Kim, D.: Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In: Proc. of the IEEE 29th International Conference of the EMBS, pp. 39–48 (2007)
84. Sander, D., Grandjean, D., Scherer, K.R.: A systems approach to appraisal mechanisms in emotion. Neural Netw. **18**(4), 317–352 (2005)
85. Scherer, K.R., Oshinsky, J.S.: Cue utilization in emotion attribution from auditory stimuli. Motiv. Emot. **1**, 331–346 (1977)

86. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford/New York (2001)
87. Schröder, M.: Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis. Ph.D. dissertation, Univ. of Saarland, Germany (2003)
88. Schröder, M., Heylen, D., Poggi, I.: Perception of non-verbal emotional listener feedback. In: Hoffmann, R., Mixdorff, H. (eds.) Speech Prosody, pp. 1–4 (2006)
89. Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Heylen, D., Maat, M., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., Sevin, E., Valstar, M., Wöllmer, M.: A demonstration of audiovisual sensitive artificial listeners. In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, vol. 1, pp. 263–264 (2009)
90. Schröder, M., Pammi, S., Gunes, H., Pantic, M., Valstar, M., Cowie, R., McKeown, G., Heylen, D., ter Maat, M., Eyben, F., Schuller, B., Wöllmer, M., Bevacqua, E., Pelachaud, C., de Sevin, E.: Come and have an emotional workout with sensitive artificial listeners! In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)
91. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. **27**, 1760–1774 (2009)
92. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: Proc. of Automatic Speech Recognition and Understanding Workshop, pp. 552–557 (2009)
93. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH, pp. 2794–2797 (2010)
94. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
95. Shen, X., Fu, X., Xuan, Y.: Do different emotional valences have same effects on spatial attention. In: Proc. of Int. Conf. on Natural Computation, vol. 4, pp. 1989–1993 (2010)
96. Sneddon, I., McKeown, G., McRorie, M., Vukicevic, T.: Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. PLoS ONE **6**, e14679–e14679 (2011)
97. Soleymani, M., Davis, J., Pun, T.: A collaborative personalized affective video retrieval system. In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, pp. 1–2 (2009)
98. Sun, K., Yu, J., Huang, Y., Hu, X.: An improved valence-arousal emotion space for video affective content representation and recognition. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 566–569 (2009)
99. Trouvain, J., Barry, W.J.: The prosody of excitement in horse race commentaries. In: Proc. ISCA Workshop Speech Emotion, pp. 86–91 (2000)
100. Truong, K.P., van Leeuwen, D.A. Neerincx, M.A. de Jong, F.M.G.: Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion. In: Proc. INTERSPEECH, pp. 2027–2030 (2009)
101. Tsai, T.-C., Chen, J.-J., Lo, W.-C.: Design and implementation of mobile personal emotion monitoring system. In: Proc. Int. Conf. on Mobile Data Management: Systems, Services and Middleware, pp. 430–435 (2009)
102. Tsiamyrtzis, P., Dowdall, J., Shastri, D., Pavlidis, I.T., Frank, M.G., Ekman, P.: Imaging facial physiology for the detection of deceit. Int. J. Comput. Vis. (2007)
103. Wang, P., Ji, Q.: Performance modeling and prediction of face recognition systems. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1566–1573 (2006)
104. Wassermann, K.C., Eng, K., Verschure, P.F.M.J.: Live soundscape composition based on synthetic emotions. IEEE Multimed. **10**, 82–90 (2003)
105. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. INTERSPEECH, pp. 597–600 (2008)

106. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. IEEE J. Sel. Top. Signal Process. **4**(5), 867–881 (2010)
107. Yang, Y.-H., Lin, Y.-C., Su, Y.-F., Chen, H.H.: Music emotion classification: A regression approach. In: Proc. of IEEE Int. Conf. on Multimedia and Expo, pp. 208–211 (2007)
108. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: Proc. of 8th Int. Conf. on Spoken Language Processing (2004)
109. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 39–58 (2009)