Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey

Hatice Gunes, Björn Schuller, Maja Pantic and Roddy Cowie

Abstract— **Despite major advances within the affective computing research field, modelling, analysing, interpreting and responding to naturalistic human affective behaviour still remains as a challenge for automated systems as emotions are complex constructs with fuzzy boundaries and with substantial individual variations in expression and experience. Thus, a small number of discrete categories (e.g., happiness and sadness) may not reflect the subtlety and complexity of the affective states conveyed by such rich sources of information. Therefore, affective and behavioural computing researchers have recently invested increased effort in exploring how to best model, analyse, interpret and respond to the subtlety, complexity and continuity (represented along a continuum e.g., from -1 to +1) of affective behaviour in terms of latent dimensions (e.g., arousal, power and valence) and appraisals. Accordingly, this paper aims to present the current state of the art and the new challenges in automatic, dimensional and continuous analysis and synthesis of human emotional behaviour in an interdisciplinary perspective.**

I. Affect in Dimesional Space

Emotions and affect are researched in various scientific disciplines such as neuroscience, psychology and cognitive sciences. Development of automatic affect analysers 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 modelling can be distinguished [1]: *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 recognised universally (e.g., [2]). 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 [3]. Hence, a number of researchers advocate the use of *dimensional description* of human affect, where affective states are not independent from one another; rather, they are related to one another in a systematic manner (e.g., [1], [3], [4], [5]). The most widely used dimensional model is a circular configuration called *Circumplex of Affect* introduced by Russell [3]. 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 polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). Another well-accepted and commonly used dimensional description is the 3D emotional space of pleasure – displeasure, arousal – nonarousal and dominance – submissiveness [4], at times referred to as the *PAD emotion space* [6] or as *emotional primitives* [7]. To guarantee a more complete description of affective colouring, some researchers include expectation (the degree of anticipating or being taken unaware) as the fourth dimension [8], and intensity (how far a person is away from a state of pure, cool rationality) as the fifth dimension (e.g., [9]). Scherer and colleagues introduced another set of psychological models, referred to as *componential models* of emotion, which are based on the appraisal theory [1], [5], [8]. 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) [1], [5], [8], [10]. Despite pioneering efforts of Scherer and colleagues (e.g., [11]), 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 [12]. 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., [13]). Despite the existence of diverse affect models, search for optimal low-dimensional representation of affect, for analysis and synthesis, and for each modality or cue, remains open [8].

Discussion: Affect in Dimesional Space

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 [14]. For instance, speech in general is reported to be less affected by the power dimension [3], than the arousal dimension, etc. A thorough comparison between all modalities would indeed provide a better understanding of which emotion dimensions are better predicted from which modalities (or cues). Although for many practical reasons arousal, valence and power dimensions have been assumed to be independent from each other, researchers reported that these emotion dimensions are correlated [15]. How to better exploit and model the *correlations between emotion dimensions*, for analysis and synthesis, should be investigated further.

II. Modalities and Cues

An individual's inner emotional state may become apparent by subjective experiences (how the person feels), internal/inward expressions (biosignals) and external/outward expressions (audio/visual signals).

A. Biosignals

Biosignals are multichannel recordings from both the central and the autonomic nervous systems. The biosignals used for automatic measurement of affect are galvanic skin response that increases linearly with a person's level of arousal [16], electromyography

H. Gunes is with Imperial College London, UK and Univeristy of Technology Sydney, Australia. M. Pantic is with Imperial College London, UK and EEMCS, Twente University, the Netherlands.

B. Schuller is with the Institute for Human-Machine Communication, Technische Universität München, Germany.

R. Cowie is with Queen's University, Belfast, UK.

This work has been funded by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE) and the grant agreement no 231287 (SSPNet).

(frequency of muscle tension) that is correlated with negatively valenced emotions [17], 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 [16], [17]. Measurements recorded over various parts of the brain including the amygdala also enable observation of the emotions felt [18]. 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. It is also possible to observe the differences between positive and negative emotional stimuli from asymmetrical brain activity [19]. 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 imagery (e.g., [20]).

B. 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 [21], [22]. Scherer and colleagues have also proposed how to judge emotional effects on vocal expression, using the appraisal-based theory [1]. In terms of affect recognition from audio signals the most reliable finding is that pitch appears to be an index into arousal [23]. Another well-accepted finding is that mean of the fundamental frequency (F0), mean intensity, speech rate, as well as pitch range [24], blaring timbre [25] and high-frequency energy [26] are positively correlated with the arousal dimension. Shorter pauses and inter-breath stretches are indicative of higher activation [27]. 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 [26] and longer vowel durations. A detailed literature summary on these can be found in [28] and [29].

C. 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., backwards head bend and arms raised forwards and upwards) and expressions (e.g., head nod), form the widely known and used visual signals for automatic affect analysis and synthesis. Ekman and Friesen [30] considered expressing discrete emotion categories via face, and communicating dimensions of affect via body as more plausible. Studies have shown that there is a relationship between the notion of approach/avoidance via the body movements and affective experiences [31], [32], e.g., as a feedback of positively and negatively valenced emotions [33], postural leaning forwards and backwards in response to affective pictures [34], etc. A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, [3] mapped the facial expressions to various positions on the 2D plane of arousal-valence, while [35] 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 utilised for recording the relationship between affect dimensions and facial feature information [36], and affect dimensions and body posture (e.g., [37], [38]). For instance, Kleinsmith et al. [38] identified that scaling, arousal, valence and action tendency were the affective dimensions used by human observers when discriminating between various body postures.

Discussion: Modalities and Cues

When input from multiple expressive channels is available the affective message conveyed by different modalities might be *congruent* (i.e., agreeing), *incongruent* (i.e., disagreeing), or *masked* (e.g., feeling angry and trying not to express it). A number of studies investigated the issue of cross-modal interactions, i.e. combined perception of human facial and bodily expressions [39], as well as human detection of emotion signs in different modalities (speech, facial expressions, gestures) when they appear to be blended or masked [40]. Overall, further research is needed in multicue and multimodal affect expression and perception in order to explore how *congruency*, *cross-modal interactions* and *blended expressions* affect dimensional affect modeling and recognition.

III. Data

A. Data Acquisition and Annotation

In the affective computing research field, the so-called *data acquisition protocol* consists of context (application domain), subjects (age, gender and cultural background), modalities, number and type of affective states, and type of data to be recorded [41]. Recorded data type can fall into one of the following categories: acted (posed), re-acted (induced via clips) and inter-acted (occurring during an interaction). Acquiring affect data without subjects' knowledge is strongly discouraged and the current trend is to record naturalistic (spontaneous) data in more constrained conditions such as an interview (e.g., [42]) or interaction (e.g., [9]) setting, where subjects are still aware of placement of the sensors and their locations. Currently, there exist a number of annotation tools with different capabilities, used for different purposes in different context. Of these, Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2D emotion space confined to $[-1, 1]$, to rate their impression about the emotional state of the subject. For annotating the internal expressions (biosignals), the level of valence and arousal is usually extracted from subjective experiences (subjects' own responses) (e.g., [18], [43]) 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) [44] is the most widely used means for self assessment. Another tool called the *emotion slider* allows the collection of self-reported valence information from subjects while they interact with a system [45]. When discretised 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 an arbitrary range, e.g., between -50 and $+50$ [46], or between -1 and $+1$ (divided into a number of levels) [14]. The final annotation is usually calculated as the mean of the observers' ratings.

B. Databases

There is a growing body of databases that contain naturalistic multimodal data labelled continuously along the emotion dimensions, and made publicly available for research purposes. The Sensitive Artificial Listener (SAL) Database [47] consists of naturalistic audio-visual data in the form of conversations that took place between a participant and an operator undertaking the role of an avatar with particular personalities, and has been annotated by 4

coders who provided continuous annotations with respect to valence and arousal dimensions. The Vera am Mittag Database consists of 12 hours of audio-visual recordings of a German Talk Show, with emotion labels given on a continuous-valued scale for the PAD primitives [48]. The SEMAINE corpus [9] consists of naturalistic audio-visual conversations taking place between a participant and a number of virtual characters with particular personalities. All recorded conversations have been transcribed and annotated for five affective dimensions (arousal, valence, power, anticipation and intensity) and partially annotated for 27 other dimensions, using trace style continuous ratings [21]. More recent naturalistic affect databases can be found in the upcoming Special Issue of IEEE Tran. on Affective Computing [49].

Discussion: Data Acquisition and Annotation

A major challenge in affective data annotation 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. Development of an easy to use, unambiguous and intuitive annotation scheme that is able to incorporate inter-observer agreement levels will indeed ease the heavy burden of the annotation task. *Obtaining high inter-observer agreement* is another challenge in affect data annotation, especially when (continuous) dimensional approach is adopted. Modelling inter-observer agreement levels within automatic affect analysers, and finding which signals better correlate with self assessments and which ones better correlate with independent observer assessments remain as challenging issues in the field.

IV. Automatic Analysis and Prediction

After affect data has 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 [50]) adopting a categorical approach to affect recognition. For further details on how features are extracted for each communicative modality, and how multicue and multimodal fusion is achieved for affect analysis purposes please see [14], [50], [51].

A. Biosignals

The most commonly employed strategy in automatic dimensional affect recognition from biosignals is to reduce the recognition problem to a two-class problem, e.g., arousal vs. non-arousal and valence vs. non-valence [52]. Interesting concepts such as multi-stage classification and identification of boundaries within the continuous emotion space have also started to emerge. For instance, Khosrowabadi et al. present in [53] an EEG-based emotion recognition system using self-organizing map to identify the boundaries (threshold levels) between separable regions of the arousal and valence dimension (into 4 emotional states). Frantzidis et al. [54] recorded biosignals while subjects viewed affective pictures. The recorded biosignals were first classified along the valence dimension, and then together with gender information, were input to a second layer distance classifier that classifies the data into high and low arousal [54]. 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 should be investigated further.

B. Audio Signals

Similarly to the affect recognition from biosignals, the most commonly employed strategy in automatic dimensional affect recognition from audio signals is to reduce the recognition problem to a two-class problem (positive vs. negative or active vs. passive classification; e.g., [56]) or a four-class problem (classification into the quadrants of 2D arousal-valence (A-V) space; e.g., [57]). As far as actual continuous dimensional affect prediction (without quantisation) is concerned, there exist a number of methods that deal exclusively with speech (i.e., [57], [58], [59]). The work by Wöllmer et al. uses the SAL Database and Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [58]. Grimm and Kroschel use the Vera am Mittag database [48] and SVRs, and compare their performance to that of the distance-based fuzzy k-Nearest Neighbour and rule-based fuzzy-logic estimators [59]. The work by Espinosa et al. also use the Vera am Mittag database [48] and examine the importance of different groups of speech acoustic features in the estimation of continuous PAD dimensions [7]. Another pioneering attempt is that of INTERSPEECH 2010 Paralinguistic Challenge featuring the Affect Sub-challenge with a focus on dimensional affect [60].

C. Visual Signals

The most commonly employed strategy in automatic dimensional affect recognition from visual signals is to reduce the recognition problem to a two-class problem (positive vs. negative or active vs. passive classification; e.g., [61], [62]) or a four-class problem (classification into the quadrants of 2D A-V space; e.g., [63], [64]). Currently, there are also a number of works focusing on dimensional and continuous prediction of emotions from the visual modality [65], [66], [67]. 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 [65]. Kipp and Martin in [66] 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 [67].

A technique to automatically segment emotional clips from long audiovisual interactions is proposed in [68], while extracting emotional segments from video based on the PAD model (assuming independency between the dimensions) is introduced in [69]. Overall, however, there is no agreement on (i) whether continuous prediction should be done without segmentation and (ii) whether segmenting videos into shorter clips is useful for dimensional and continuous emotion recognition.

D. Motion Capture Signals

Motion capture systems have mostly been utilised for recording the relationship between affect dimensions and facial feature information (e.g., [36]), and affect dimensions and body posture (e.g.,

[37], [38]). In general, low-level posture features such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and left shoulder) appear to help in effectively discriminating between the (quantised) affect dimensions [37], [38]. To the best of our knowledge, dimensional and continuous analysis of affect from motion capture data has not been attempted yet.

E. Thermal Imaging Signals

Relatively few efforts have been reported on dimensional affect recognition from thermal imagery. Nhan and Chau [70] focus on recording the thermal infrared signals of the subjects stimulated with images from the International Affective Pictures System. They use the self-reported affect as ground truth and achieve high vs. low classification of arousal and valence using the time-frequency features derived from thermal infrared data. Merla and Romani utilised functional infrared imaging (fIR) to study the facial thermal signatures of three emotional conditions: stress, fear and pleasure arousal [71]. They reported that fIR can be reliably used to assess emotional arousal. Overall, dimensional and continuous analysis of affect from thermal signals has not been attempted yet.

F. Modality Fusion

When it comes to dimensional emotion recognition using multiple modalities the focus has mainly been on discriminating between more coarse categories, such as positive vs. negative [61] and active vs. passive [72]. Of these, Caridakis et al. [72] use the SAL database, combining auditive and visual modalities. Nicolaou et al. focus on audio-visual classification of spontaneous affect into negative or positive emotion categories using facial expression, shoulder and audio cues, and utilising 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities [61]. Kanluan et al. [73] combine audio and visual cues for affect recognition in A-V space by fusing facial expression and audio cues, using SVRs and late fusion with a weighted linear combination with discretised labels (on a 5-point scale in the range of [-1,+1] for each emotion dimension). Wöllmer et al. use multimodal acted data that contain face (obtained from motion capture and video) and audio information, and recognise 3–5 levels of A-V values using various classification techniques [36]. More recent works focus on dimensional and continuous prediction of emotions from multiple modalities. For instance, Eyben et al. [74] propose a string-based approach for fusing the behavioural 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 analysers based on physiology end up using multiple signal sources, explicit fusion of multimodal data for continuous modelling of affect utilising 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) [41]. 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. [75] that propose a dimensional multimodal fusion scheme based on the PAD space to support detection and integration of spontaneous affective behaviour of users (in terms of audio, video and attention events) experiencing arts and entertainment. Unlike many other other multimodal approaches (e.g., [61], [72], [73]), the ground truth in this work is obtained by measuring Galvanic Skin Response (GSR) as an independent measure of arousal. Overall, finding the type of annotation to be used as ground truth (i.e., self assessment, rater assessment or measurement-based assessment) and the best way to fuse the modalities (fusing at feature, event or decision level) for dimensional and continuous emotion recognition remain as an open issues in the field.

Discussion: Automatic Analysis and Prediction

The *window size* to be used to achieve optimal affect recognition is one of the issues that the existing literature does not provide a unique answer to. Current affect recognisers employ various *window sizes* depending on the modality. 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., [42]), while on the other hand obtaining a reliable prediction accuracy requires long(er)-term monitoring [76], [77]. Chanel et al. [42] 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.

The *Baseline problem* is another major challenge in the field. For biosignals this refers to the problem of finding a condition against which changes in measured physiological signals can be compared (a state of calmness) [78]. For audio modality this is usually achieved by segmenting the recordings into turns and processing each turn separately (e.g., [58]). 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 usually achieved by segmenting the recordings (e.g., [61], [79]). This remains a great challenge in automatic analysis, which typically relies on existence of a baseline for analysis and processing of affective information.

Generalisation capability of automatic affect analysers across subjects is still a challenge in the field. Kulic and Croft [43] 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 generalisation over unseen subjects a very difficult problem. A common way of measuring affect from biosignals is doing it for each participant separately (without computing baseline), e.g., [42]. Similarly to the recent works on automatic affect prediction from audio or visual cues (e.g., [67]), better insight may be obtained by comparing subject-dependent vs. subject-independent prediction results.

Classification methods used for dimensional and continuous affect measurement should be able to produce continuous values for the target dimensions. Some of the classification schemes that have been explored for this task are, SVR, RVM, and Long Short-Term Memory Recurrent Networks. Linear Discriminant Analysis, Conditional Random Fields and Support Vector Machines have been used for quantised dimensional affect recognition tasks (e.g., [58]). Overall, there is no agreement on how to model dimensional affect space (continuous vs. quantised) and which classifier is better suited for automatic, multimodal, continuous affect analysis using a dimensional representation.

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 value 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., [58], [73]). However, using MSE might not be the best way to evaluate the performance of dimensional approaches to automatic affect recognition. Therefore, the correlation coefficient is also employed by several studies (e.g., [73]) together with MSE. Overall, however, how to obtain optimal evaluation metrics for continuous and dimensional emotion recognition remains an open research issue [14].

V. Audio-visual Synthesis

A. Speech

There has been a lot of past research on neutral speech synthesis and a number of tools are already freely available for research purposes (see [80] for a recent review). However, despite various methods proposed (e.g., rule-based, data-driven [81], etc.), *expressive speech synthesis* is still a challenging research issue. Expressive speech synthesis utilises acoustic features, such as pitch variables

(F0 level, range, contour and jitter), intensity and speech rate. This is either done in a data-driven manner (using a pre-recorded database and modelling a few well-defined emotional states) or in a model-based manner (using a pre-defined model) [29]. The most commonly employed strategy in expressive speech synthesis is using the categorical representation of emotions and the unit selection method, that concatenates speech segments stored in a database (e.g., [82], [83], [84]). Of these, Bulut et al. propose an interesting approach of emotional speech *re-synthesis* that consists of synthesis, recognition, parameter selection and re-synthesis modules that appears to improve the human evaluation of emotionally synthesised speech [85]. The feasibility and usefulness of such a method should be explored for dimensional and continuous representation in emotional speech synthesis. A representative work for using dimensional representation of emotions is that of Schröder who proposes a model-based expressive speech synthesis technique in [29]. Tao et al. in [86] focus on expressive speech synthesis utilising the unconventional categories of strong, medium, and weak classifications and create a deviation of perceived expressiveness (DPE) measure to evaluate the expressiveness of the output speech. Their evaluation results show that a database with neutral semantic content should be used for emotional speech synthesis. Siliang et al. introduce an emotional speech synthesis process, by adjusting the parameters (XML-tags) used to synthesise emotional speech dynamically, using interactive Genetic Algorithms [87]. For an overview on emotional speech synthesis and its practical applications see [88].

B. Multimodality

Audio-visual expressive speech synthesis is a relatively new research area [89], [90], [91] with a major focus on discrete representation of basic emotions. More recently, a number of works started focusing on dimensional representation of emotions: exploring the facial expressions within a 2D emotion space [92], adopting the activation - evaluation dimensions for synthesising facial expressions of pure and mixed emotions of varying intensities [93], creating an embodied conversational agent by modeling the emotion with positive-negative valence and timing (past, current and future), generating facial expressions that convey the nonverbal information accompanying speech [94], and creating interactive installations that combine both discrete and dimensional models of emotions/moods [95]. Busso et al. [96] found that head motion patterns with neutral speech significantly differ from head motion patterns with emotional speech in terms of motion activation, range and velocity. Their synthesis results also show that head motion modifies the emotional perception (represented in terms of quantised PAD values) of the facial animation especially in the *valence* and *activation* domain. Boukricha et al. in [97] present a facial expression simulation system that has a control architecture for simulating emotional facial expressions with respect to PAD values, and an expressive output component for animating the virtual human's facial muscle actions (Action Units) based on the Facial Action Coding System (FACS). One of the most recent attempts has been introduced by Jia et al. in [6] that uses the 3-D PAD emotion model and proposes a unified model for emotional speech conversion (using Boosting Gaussian Mixture Model) and a facial expression synthesis model. Shen et al. in [98] introduce a system that synthesises the emotional audio-visual speech for a 3-D talking agent by utilising a GMM-based model to predict variation of acoustic features for emotional speech by PAD values, and build a parametric framework of PAD-driven emotional facial expression synthesis.

A further step in audio-visual expressive speech synthesis is the design of virtual humans, commonly known as embodied conversational agents (ECA), that are endowed with (more) elaborate social skills. ECAs have appeared in the scene slightly before the start of this century [99]. The difference between a Spoken Dialogue System and an ECA lies in the latter 'having an identity and a persona' [99]. Therefore, similarly to human-human interaction, an ECA is at times expected to play the role of a *speaker* (producer of the communicative behaviour), while at other times the role of a *listener* (recipient of the communicative behaviour), with appropriate turn-taking behaviour (a turn occurs when an interlocutor resumes the primary speaker role). Most of the past research focused on creating ECA systems based on static input parameters, rather than dynamically changing the behaviour of the virtual agent based on the behaviour of the user during interaction [99]. One of the pioneering attempts in creating expressive and adaptive virtual agents is the SEMAINE system [100], a publicly available, fully autonomous Sensitive Artificial Listeners (SAL) system that consists of virtual dialogue partners based on audiovisual analysis and synthesis [100]. The system runs in real-time, and combines incremental analysis of user behaviour, dialogue management, and synthesis of speaker and listener behaviour of a SAL character displayed as a virtual agent. Endowing ECAs with *multimodal expressive and adaptive behaviour* is an ongoing research topic within the virtual agent research community [101], [102]. The ECAs of today are endowed with functions and capabilities in various areas such as task and content, control and social-affective behaviour (e.g., displaying friendliness, being able to motivate people and give confidence, and being polite, showing rapport, empathy, or engagement). The interplay of these areas and multiple behavioural cues challenge the mapping between signal/behaviour and meaning/function [99]. Although perception studies have attempted to unveil the significance of these, they strip off the context from the displayed behaviour, and therefore provide only limited insight to the issue. This is in a way the chicken or the egg causality dilemma, where current video recordings contain only somewhat artificial interactions used for the design of current generation virtual agents, that in turn only allow for somewhat artificial interactions, that are judged negatively by the

Discussion: Synthesis and ECAs

Creating virtual agents that can interactively align to their interaction partners in their verbal and nonverbal behaviour is challenging. A virtual agent can be made capable of showing alignment behaviour by picking up the behaviour of the dialogue partner's syntactic structures and lexical items in its subsequent utterances, and generating gestures that will concord with such linguistic content (e.g., [101]). Currently, linguistic features are predominant in mediating alignment.

The *ECA systems of today assume independency* between parsing (analysing), interpreting (understanding) and generating (producing) speech. As each of these tasks faces different challenges, they are modeled (and optimised) independently. Currently the aim is to retain the empirically observed correlations between speech and other modalities (e.g., gesture) while keeping the predominant role of the linguistic choices (lexical or syntactic choices). However, *modelling cross-modal aspects* (modality alignment and coupling) *of an interaction* remains a complex task, and calls for appropriate understanding of natural language.

The *speaker role of an ECA*, by actively generating verbal and nonverbal signals, *has received much more emphasis than its listener role* [103]. Endowing ECAs with appropriate listening behaviuor is more challenging as it depends on the context and the state of the user, and involves generating feedback signals and back-channels accordingly.

Most ECA systems are concerned with congruent adaptation and coupling of modalities (obtained via empirical observation). *When and how to model incongruency* remains as an interesting phenomenon to be explored for ECAs.

human interlocutors. Further research is needed to understand how various communicative signals work together in different content, context and conditions.

VI. Frameworks and Tools

In the last five years, various research groups have created publicly available frameworks and tools to be used for researching dimensional and continuous analysis and synthesis of emotions (e.g., [104], [105], [106]). Of these, the SEMAINE API introduced in [104] is an open source framework for building emotion-oriented systems, using standard representation formats and providing a Java and C++ wrapper around a message-oriented middleware. The OpenSmile library is written in C++ and enables extraction of large audio feature spaces in real-time [105]. A middleware solution to aid the design and development of healthcare applications with affective information is introduced in [106].

VII. Applications

Various applications have been using the dimensional (both quantised and continuous) representation and prediction of emotions, ranging from human-computer (e.g., Sensitive Talking Heads [107], Sensitive Artificial Listeners [100], spatial attention analysis [98], arts installations [95]) and human-robot interaction (e.g., humanoid robotics [108], [109]), clinical and biomedical studies (e.g., stress/pain monitoring [110], [111], [112], autism-related assistive technology), learning and driving environments (e.g., episodic learning [113], affect analysis in the car [114]), multimedia (e.g, video content representation and retrieval [115], [116] and personalised affective video retrieval [117]), and entertainment technology (e.g., gaming [118]). 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 [119] established R. Picard and colleagues of MIT Media Lab). These advances, in turn, have triggered further issues such as *ethics*. As has been emphasised in the IEEE Tran. on Affective Computing's Special Issue on ethics [121], 'If machines are going to be turned loose on their own to kill and heal, explore and decide, the need for designing them to be moral becomes pressing'. The rapid progress in affective computing points to the fact that ethics is not merely science fiction and should be taken into serious consideration [120].

VIII. Concluding Remarks

Human affect representation, analysis and synthesis based on dimensional approaches is still in its infancy. However, 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 behaviour for emotion-related disorders like autism), and funded by various research projects (e.g., European Union FP 7, SEMAINE). 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. 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 sensing and recognition (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 summarised in this paper in the Discussion tables provided in each section. Despite encouraging efforts and major progress in affect analysis and synthesis, highlighted in this paper, in general, (dynamic) social-affect capabilities, that by definition require understanding of the needs, desires, goals and emotional state of the user, of most virtual agents are rather limited [99]. Overall, affect analysis and affect synthesis appear to be detached from each other even in multi-party and multi-disciplinary projects such as SEMAINE [100]. Although the overall perception and acceptability of an automated system depends on the (complex) interplay of these two domains, analysis and synthesis are treated as independent problems and only linked in the final stage. Investigating how to inter-relate these in earlier stages will indeed provide valuable insight into the nature of both areas that play a crucial role for the realisation of multimodal, dimensional and continuous affective computing.

References

[1] D. Grandjean, D. Sander, & K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness & Cognition*, vol. 17, no. 2, pp. 484 – 495, 2008, Social Cognition, Emotion, & Self-Consciousness.

[2] P. Ekman & W.V. Friesen, *Unmasking The Face: A Guide To Recognizing Emotions From Facial Clues*, Prentice Hall, New Jersey, 1975.

[3] J. A. Russell, "A circumplex model of affect," *J. of Personality & Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[4] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology: Developmental, Learning, Personality, Social*, vol. 14, pp. 261–292, 1996.

[5] K.R. Scherer, A. Schorr, & T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford Univ. Press, Oxford/New York, 2001.

[6] J. Jia et al., "Emotional audio-visual speech synthesis based on pad," *IEEE Trans. on Audio, Speech, & Language Processing*, vol. PP, no. 9, pp. 1–1, 2010.

[7] H.P. Espinosa, C.A.R. Garcia, & L.V. Pineda, "Features selection for primitives estimation on emotional speech," in *Proc. IEEE ICASSP*, 2010, pp. 5138–5141.

[8] J. R. Fontaine et al., "The world of emotion is not two-dimensional," *Psychological Science*, vol. 18, pp. 1050–1057, 2007.

[9] G. McKeown et al., "The semaine corpus of emotionally coloured character interactions," in *Proc. IEEE ICME*, 2010, pp. 1079–1084.

[10] N. H. Frijda, *The emotions*, Cambridge Univ. Press, 1986.

[11] D. Sander, D. Grandjean, & K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Netw.*, vol. 18, no. 4, pp. 317–352, 2005.

[12] A. Ortony, G. L. Clore, & A. Collins, *The cognitive structure of emotions*, Cambridge Univ. Press, Oxford, 1988.

[13] C. Bartneck, "Integrating the occ model of emotions in embodied characters," 2002, pp. 39–48.

[14] H. Gunes & M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[15] R. Dietz & A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *Proc. Cognitive Technology*, 1999.

[16] G. Chanel, K. Ansari-Asl, & T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," in *Proc. IEEE SMC*, Oct. 2007, pp. 2662–2667.

[17] A. Haag et al., "Emotion recognition using bio-sensors: First steps towards an automatic system," in *LNCS 3068*, 2004, pp. 36–48.

[18] T. Pun et al., "Brain–computer interaction research at the computer vision and multimedia laboratory, univ. of geneva," *IEEE Tran. on Neural Systems & Rehabilitation Engineering*, vol. 14, pp. 210–213, 2006.

[19] R.J. Davidson & N.A. Fox, "Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants," *Science*, vol. 218, pp. 1235–1237, 1982.

[20] P. Tsiamyrtzis et al., "Imaging facial physiology for the detection of deceit," *Int. J. of Computer Vision*, 2007.

[21] R. Cowie et al., "Feeltrace: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop on Speech & Emotion*, 2000, pp. 19–24.

[22] R. Cowie et al., "Emotion recognition in human–computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 33–80, 2001.

[23] R.A. Calvo & S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[24] G. L. Huttar, "Relations between prosodic variables and emotions in normal american english utterances," *J. Speech Hearing Res.*, vol. 11, pp. 481–487, 1968.

[25] J. Davitz, *The Communication of Emotional Meaning*, chapter Auditory correlates of vocal expression of emotional feeling, pp. 101–112, McGraw-Hill, 1964.

[26] K. R. Scherer & J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation Emotion*, vol. 1, pp. 331–346, 1977.

[27] J. Trouvain & W. J. Barry, "The prosody of excitement in horse race commentaries," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 86–91.

[28] M. Schröder, *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.

[29] M. Schröder, D. Heylen, & I. Poggi, "Perception of non-verbal emotional listener feedback," in *Speech Prosody*, R. Hoffmann & H. Mixdorff, Eds., 2006, pp. 1–4.

[30] P. Ekman & W. V. Friesen, "Head and body cues in the judgment of emotion: A reformulation," *Perceptual & Motor Skills*, vol. 24, pp. 711–724, 1967.

[31] M. Chen & J. A. Bargh, "Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus," *Personality & Social Psychology Bulletin*, vol. 25, pp. 215–224, 1999.

[32] J. Forster & F. Strack, "Influence of overt head movements on memory for valenced words: A case of conceptual–motor compatibility," *J. of Personality & Social Psychology*, vol. 71, pp. 421–430, 1996.

[33] C.S. Carver, "Pleasure as a sign you can attend to something else: Placing positive feelings within a general model of affect," *Cognition & Emotion*, vol. 17, pp. 241–261, 2003.

[34] C.H. Hillman, K.S. Rosengren, & D.P. Smith, "Emotion and motivated behavior: postural adjustments to affective picture viewing," *Biological Psychology*, vol. 66, pp. 51–62, 2004.

[35] R. Cowie et al., "The emotional and communicative significance of head nods and shakes in a naturalistic database," in *Proc. LREC Int. Workshop on Emotion*, 2010, pp. 42–46.

[36] M. Wöllmer et al., "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH*, 2010, pp. 2362–2365.

[37] A. Kleinsmith & N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proc. ACII*, 2007, pp. 48–58.

[38] A. Kleinsmith, P. R. De Silva, & N. Bianchi-Berthouze, "Recognizing emotion from postures: Cross–cultural differences in user modeling," in *Proc. the Conf. on User Modeling*, 2005, pp. 50–59.

[39] H. K. Meeren, C. C. Van Heijnsbergen, & B. De Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proc. the National Academy of Sciences*, vol. 102, pp. 16518–16523, 2005.

[40] S. Buisine et al., "Perception of blended emotions: From video corpus to expressive agent," in *Proc. IVA*, 2006, pp. 93–106.

[41] Z. Khalili & M.H. Moradi, "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*, 2009, pp. 1571–1575.

[42] G. Chanel et al., "Short-term emotion assessment in a recall paradigm," *Int. J. of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[43] D. Kulic & E. A. Croft, "Affective state estimation for human-robot interaction.," *IEEE Trans. on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[44] P.J. Lang, *The cognitive psychophysiology of Emotion: Anxiety and the anxiety disorders*, Lawrence Erlbaum, NJ, 1985.

[45] G. Laurans, P. Desmet, & P. Hekkert, "The emotion slider: A self-report device for the continuous measurement of emotion," in *Proc. ACII Workshops*, 2009, pp. 1–6.

[46] J.M. Kessens et al., "Perception of synthetic emotion expressions in speech: Categorical and dimensional annotations," in *Proc. ACII Workshops*, 2009, pp. 1–5.

[47] E. Douglas-Cowie et al., "The humaine database: addressing the needs of the affective computing community," in *Proc. ACII*, 2007, pp. 488–500.

[48] M. Grimm, K. Kroschel, & S. Narayanan, "The vera am mittag german audio-visual emotional speech database.," in *Proc. IEEE ICME*, 2008, pp. 865–868.

[49] B. Schuller et al., "Special issue on naturalistic affect resources for system building and evaluation," *IEEE Tran. on Affective Computing*, 2011.

[50] H. Gunes, M. Piccardi, & M. Pantic, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, Ch. From the Lab to the Real World: Affect Recognition using Multiple Cues and Modalities, pp. 185–218, I-Tech Educ. & Publishing, 2008.

[51] Z. Zeng et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Tran. on Pattern Analysis & Machine Intelligence*, vol. 31, pp. 39–58, 2009.

[52] Y. Gu et al., "Emotion-aware technologies for consumer electronics," in *Proc. IEEE Int. Symp. on Consumer Electronics*, 2008, pp. 1–4.

[53] R. Khosrowabadi et al., "Eeg-based emotion recognition using self-organizing map for boundary detection," in *Proc. ICPR*, 2010, pp. 4242–4245.

[54] C.A. Frantzidis et al., "On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications," *IEEE Trans. on Information Techn. in Biomedicine*, vol. 14, no. 2, pp. 309–318, 2010.

[55] J. Kim & E. Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[56] B. Schuller et al., "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE ASRU*, 2009.

[57] M. Wöllmer et al., "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE J. of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867 –881, Oct. 2010.

[58] M. Wöllmer et al., "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Interspeech*, 2008, pp. 597–600.

[59] M. Grimm & K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," in *Proc. IEEE Automatic Speech Recog. & Understanding Workshop*, 2005, pp. 381–385.

[60] B. Schuller et al., "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.

[61] M.A. Nicolaou, H. Gunes, & M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Proc. Int. Conf. on Pattern Recognition*, 2010, pp. 3695–3699.

[62] D. McDuff et al., "Affect valence inference from facial action unit spectrograms," in *Proc. IEEE CVPR Workshops*, 2010, pp. 17–24.

[63] D. Glowinski et al., "Technique for automatic emotion recognition by body gesture analysis," in *Proc. IEEE CVPR Workshops*, 2008, pp. 1–6.

[64] S. Ioannou et al., "Emotion recognition through facial expression analysis based on a neurofuzzy method," *J. of Neural Networks*, vol. 18, pp. 423–435, 2005.

[65] H. Gunes & M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. of IVA*, 2010, pp. 371–377.

[66] M. Kipp & J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?," in *Proc. ACII Workshops*, 2009, pp. 1–8.

[67] M.A. Nicolaou, H. Gunes, & M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," in *Proc. of IEEE FG*, 2011.

[68] M.A. Nicolaou, H. Gunes, & M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," in *Proc. LREC Int. Workshop on Multimodal Corpora*, 2010, pp. 43–48.

[69] S. Arifin & P. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Trans. on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008.

[70] B.R. Nhan & T. Chau, "Classifying affective states using thermal infrared imaging of the human face," *IEEE Tran. on Biomedical Engineering*, vol. 57, no. 4, pp. 979–987, 2010.

[71] A. Merla & G.L. Romani, "Thermal signatures of emotional arousal: a functional infrared imaging study," in *Proc. Int. Conf. of the IEEE Engineering in Medicine & Biology Society*, 2007, pp. 247–249.

[72] G. Caridakis et al., "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proc. ACM ICMI*, 2006, pp. 146–154.

[73] I. Kanluan, M. Grimm, & K. Kroschel, "Audio-visual emotion recognition using an emotion recognition space concept," *Proc. European Signal Processing Conf.*, 2008.

[74] F. Eyben et al, "String-based Audiovisual Fusion of Behavioural Events for the Assessment of Dimensional Affect," in *Proc. IEEE FG*, 2011.

[75] S.W. Gilroy et al., "Pad-based multimodal affective fusion," in *Proc. ACII Workshops*, 2009, pp. 1–8.

[76] G.G. Berntson et al., "Heart rate variability:origins, methods, and interpretive caveats," *Psychophysiology*, vol. 34, no. 6, pp. 623.

[77] L. Salahuddin et al., "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *Proc. IEEE Int. Conf. of EMBS*, 2007, pp. 39–48.

[78] A. Nakasone, H. Prendinger, & M. Ishizuka, "Emotion recognition from electromyography and skin conductance," in *Proc. Int. Workshop on Biosignal Interpretation*, 2005, pp. 219–222.

[79] S. Petridis et al., "Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities," in *Proc. ACM ICMI*, 2009, pp. 23–30.

[80] S. Pammi, M. Charfuelan, & M. Schröder, "Multilingual voice creation toolkit for the mary tts platform," in *Proc. LREC*, 2010.

[81] E. Eide et al., "A corpus-based approach to ¡ahem/¿ expressive speech synthesis," in *Proc. IEEE Speech Synthesis Workshop*, 2002, pp. 79–84.

[82] S. Mori, T. Moriyama, & S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," in *Proc. IEEE ICME*, 2006, pp. 1093–1096.

[83] M. Bulut, Sungbok Lee, & S. Narayanan, "A statistical approach for modeling prosody features using pos tags for emotional speech synthesis," in *Proc. IEEE ICASSP*, 2007, vol. 4, pp. 1237–1240.

[84] Y. Qin, X. Zhang, & H. Ying, "A hmm-based fuzzy affective model for emotional speech synthesis," in *Proc. Int. Conf. on Signal Processing Systems*, 2010, vol. 3, pp. 525–528.

[85] M. Bulut, Sungbok Lee, & S. Narayanan, "Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech," in *Proc. IEEE ICASSP*, 2008, pp. 4629–4632.

[86] J. Tao, Y. Kang, & A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. on Audio, Speech, & Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[87] S. Lv, S. Wang, & X. Wang, "Emotional speech synthesis by xml file using interactive genetic algorithms," in *Proc. ACM/SIGEVO Summit on Genetic & Evolutionary Computation*, 2009.

[88] F. Burkhardt & J. Stegmann, "Emotional speech synthesis: Applications, history and possible future," in *Proc. ESSV*, 2009.

[89] E. Bevacqua & C. Pelachaud, "Expressive audio-visual speech," *Computer Animation & Virtual Worlds*, vol. 15, no. 3-4, pp. 297–304, 2004.

[90] M. Schröder, "Emotional speech synthesis:a review," in *Proc. Eurospeech*, 2001, vol. 1, pp. 561–564.

[91] H. Tang et al., "Humanoid audiovisual avatar with emotive text-to-speech synthesis," *IEEE Trans. on Multimedia*, vol. 10, no. 6, pp. 969–981, 2008.

[92] Z. Ruttkay, H. Noot, & P. Hagen, "Emotion disc and emotion squares: Tools to explore the facial expression space," *Computer Graphics Forum*, vol. 22, pp. 49–53, 2003.

[93] I. Albrecht et al., "Mixed feelings: expression of non-basic emotions in a musclebased talking head," *Virtual Reality*, vol. 8, pp. 201–212, 2005.

[94] C. Pelachaud & M. Bilvi, "Computational model of believable conversational agents," in *LNCS 2650*, 2003, pp. 300–317.

[95] K. C. Wassermann, K. Eng, & P. F. M. J. Verschure, "Live soundscape composition based on synthetic emotions," *IEEE Multimedia*, vol. 10, pp. 82–90, 2003.

[96] C. Busso et al., "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Trans. on Audio, Speech, & Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2009.

[97] H. Boukricha et al., "Pleasure-arousal-dominance driven facial expression simulation," in *Proc. ACII Workshops*, 2009, pp. 1–7.

[98] X. Shen, X. Fu, & Y. Xuan, "Do different emotional valences have same effects on spatial attention?," in *Proc. of Int. Conf. on Natural Computation*, 2010, vol. 4, pp. 1989–1993.

[99] D. Heylen et al., "Social agents: The first generations," in *Proc. ACII Workshops*, 2009, pp. 1–7.

[100] M. Schröder et al., "A demonstration of audiovisual sensitive artificial listeners," in *Proc. ACII*, 2009, vol. 1, pp. 263–264.

[101] H. Buschmeier, K. Bergmann, & S. Kopp, "Adaptive expressiveness virtual conversational agents that can align to their interaction partner," in *Proc. Int. Conf. on Autonomous Agents & Multiagent Systems*, 2010, pp. 91–98.

[102] C. Pelachaud, "Modelling multimodal expression of emotion in a virtual agent," *Philosophical Trans. of Royal Society B Biological Science*, vol. 364, pp. 3539–3548, 2009.

[103] E. Bevacqua et al., "Facial feedback signals for ecas," in *Proc. Artificial & Ambient Intelligence*, 2007, pp. 1–7.

[104] M. Schröder, "The semaine api: Towards a standards-based framework for building emotion-oriented systems," *Advances in Human-Machine Interaction*, vol. 2010, pp. 1–21, 2010.

[105] F. Eyben, M. Wöllmer, & B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[106] T. Taleb, D. Bottazzi, & N. Nasser, "A novel middleware solution to improve ubiquitous healthcare systems aided by affective information," *IEEE Trans. on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 335–349, 2010.

[107] T.S. Huang et al., "Sensitive talking heads," *IEEE Signal Processing Magazine*, vol. 26, pp. 67–72, 2009.

[108] A. Beck, L. Canamero, & K.A. Bard, "Towards an affect space for robots to display emotional body language," in *Proc. IEEE RO-MAN*, 2010, pp. 464–469.

[109] M. Karg et al., "Towards mapping emotive gait patterns from human to robot," in *Proc. IEEE RO-MAN*, 2010, pp. 258–263.

[110] M. Mihelj, D. Novak, & M. Munih, "Emotion-aware system for upper extremity rehabilitation," in *Proc. Int. Conf. on Virtual Rehabilitation*, 2009, pp. 160–165.

[111] T.-C. Tsai, J.-J. Chen, & W.-C. Lo, "Design and implementation of mobile personal emotion monitoring system," in *Proc. Int. Conf. on Mobile Data Management*, 2009, pp. 430–435.

[112] B. Grundlehner et al., "The design and analysis of a real-time, continuous arousal monitor," in *Proc. Int. Workshop on Wearable & Implantable Body Sensor Networks*, 2009, pp. 156–161.

[113] U. Faghihi et al., "How emotional mechanism helps episodic learning in a cognitive agent," in *Proc. IEEE Symp. on Intelligent Agents*, 2009, pp. 23–30.

[114] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, & N. Nguyen-Thien, "Emotion on the road - necessity, acceptance, and feasibility of affective computing in the car," *Advances in Human-Machine Interaction*, vol. 2010, pp. 1–17, 2010.

[115] K. Sun et al., "An improved valence-arousal emotion space for video affective content representation and recognition," in *Proc. IEEE ICME*, 2009, pp. 566–569.

[116] J.J.M. Kierkels, M. Soleymani, & T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Proc. IEEE ICME*, 2009, pp. 1436–1439.

[117] M. Soleymani, J. Davis, & T. Pun, "A collaborative personalized affective video retrieval system," in *Proc. ACII Workshops*, 2009, pp. 1–2.

[118] M. Rehm & M. Wissner, "Gamble a multiuser game with an embodied conversational agent," in *LNCS 3711*, 2005, pp. 180–191.

[119] "Affectiva's homepage: http://www.affectiva.com/," 2010.

[120] C. Allen, W. Wallach, & I. Smit, "Why machine ethics?," *IEEE Intelligent Systems*, vol. 21, pp. 12–17, 2006.

[121] A. Beavers, "Special issue on ethics & affective computing ," *IEEE Trans. on Affective Computing*, 2011.