Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing

Alessandro Vinciarelli, *Member*, *IEEE*, Maja Pantic, *Fellow*, *IEEE*, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schröder

Abstract—Social Signal Processing is the research domain aimed at bridging the social intelligence gap between humans and machines. This paper is the first survey of the domain that jointly considers its three major aspects, namely, modeling, analysis, and synthesis of social behavior. Modeling investigates laws and principles underlying social interaction, analysis explores approaches for automatic understanding of social exchanges recorded with different sensors, and synthesis studies techniques for the generation of social behavior via various forms of embodiment. For each of the above aspects, the paper includes an extensive survey of the literature, points to the most important publicly available resources, and outlines the most fundamental challenges ahead.

Index Terms—Social signal processing, nonverbal behavior analysis and synthesis, social interactions understanding.

1 Introduction

Following one of the most famous statements of Western philosophy (Aristotle, Politika ca. 328 BC)[1]:

> *Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human.*

Almost 25 centuries after these words were written for the first time, several disciplines confirm the intuition of Aristotle by grounding the social nature of humans into measurable and observable aspects of human biology, psychology, and behavior. Neuroscientists have identified brain structures, called *mirror neurons* [1], that seem to have no other goal than improving our awareness of others, whether this means to share their feelings [2] or to learn through imitation [3]. Biologists and physiologists have shown that our ears are tuned to human voices more than to any other sound [4], that the only facial muscles present in every human being (the others can be absent) are those we use to communicate the six basic emotions [5], and, more generally, that evolution has shaped our body and senses around social contacts. Furthermore, human sciences (psychology, anthropology, sociology, etc.) have shown how social interactions dominate our perception of the world [6] and shape our daily behavior by attaching social meaning to acts as simple and spontaneous as gestures, facial expressions, intonations, etc. [7].

The computing community could not remain immune to this wave of interest for the "*social animal*." Nowadays, computers are leaving their original role of improved versions of old tools and moving toward a new, human-centered vision of computing [8] where intelligent machines seamlessly integrate and support human-human interactions [9], embody natural modes of human communication for interacting with their users [10], and are the platform through which large scale social activities take place online [11]. In such a new context, the gap between social animal and unsocial machine is no longer acceptable and socially adept computers become a crucial need and challenge for the future of computing [12].

Social Signal Processing (SSP) [13] is the new, emerging domain addressing such a challenge by providing computers with social intelligence [14], the facet of our cognitive abilities that guides us through our everyday social interactions, whether these require us to be a respected colleague in the workplace, a careful parent at home, a leader in our community, or simply a person others like to have around in a moment of relaxation. At its heart, social intelligence aims at correct perception, accurate interpretation, and appropriate display of social signals [15], [16].

1. At the time this paper is being written, the sentence "*Man is by nature a social animal*" returns 1.6 million documents when submitted to Google as a query (only documents including the whole statement are counted).

- *A. Vinciarelli is with the University of Glasgow, Sir A. Williams Building, Glasgow G12 8QQ, United Kingdom, and with the Idiap Research Institute, CP592, 1920 Martigny, Switzerland. E-mail: vincia@dcs.gla.ac.uk.*
- *M. Pantic is with the Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom, and wit the University of Twente, Human Media Interaction, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: m.pantic@imperial.ac.uk.*
- *D. Heylen is with the University of Twente, Human Media Interaction, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: d.k.j.heylen@utwente.nl.*
- *C. Pelachaud is with CNRS—LTCI UMR 5141, Institut TELECOM—TELECOM ParisTech, 37 rue Dareau, Paris 75014, France. E-mail: catherine.pelachaud@telecom-paristech.fr.*
- *I. Poggi and F. D'Errico are with the University Roma Tre, Via del Castro Pretorio, 20, Roma 00185, Italy. E-mail: {poggi, fderrico}@uniroma3.it.*
- *M. Schröder is with DFKI GmbH, Campus D3_2, Stuhlsatzenhausweg 3, D-66123, Saarbrücken, Germany. E-mail: marc.schroeder@dfki.de.*

Manuscript received 25 Aug. 2010; revised 1 Feb. 2011; accepted 6 July 2011; published online 4 Aug. 2011.
Recommended for acceptance by J. Gratch.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2010-08-0063.
Digital Object Identifier no. DOI: 10.1109/T-AFFC.2011.27.

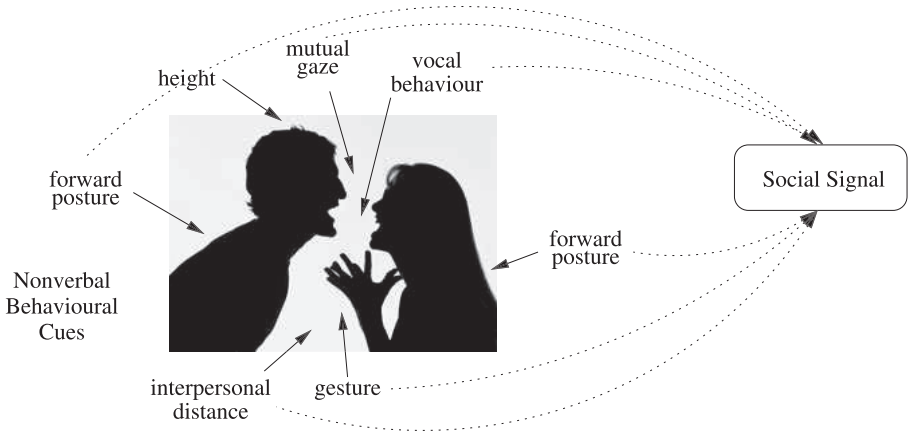

Fig. 1. Nonverbal behavioral cues and social signals. With no more than two silhouettes at disposition, it is not difficult for most people to guess that the picture portrays a couple involved in a fight. Nonverbal behavioral cues allow one to understand that the social signals being exchange are disagreement, hostility, aggressiveness, etc., and that the two people have a tight relationship.

These are relational attitudes through which we express, often unconsciously, our actual feelings toward interactions and social contexts. Social signals include interest, empathy, hostility, agreement, flirting, dominance, superiority, etc.

One of the most important aspects of social signals is that they can take the form of complex constellations of nonverbal behavioral cues (facial expressions, prosody, gestures, postures, etc.) that accompany any human-human [17] and human-machine [18] interaction (see Fig. 1). Several decades of human sciences have shown that we are surprisingly effective at understanding social signals underlying the rich variety of nonverbal behaviors displayed by people around us [19]. This leads to the two core questions addressed by Social Signal Processing:

- Is it possible to automatically infer social signals from nonverbal behavioral cues detected through sensors like microphones and cameras?
- Is it possible to synthesize nonverbal behavioral cues conveying desired social signals for embodiment of social behaviors in artificial agents, robots, or other manufacts?

Most SSP works revolve around these questions and involve a tight, multidisciplinary collaboration between human sciences (psychology, anthropology, sociology, etc.) on one hand and computing sciences (computer vision, speech analysis and synthesis, machine learning, signal processing, etc.) on the other hand.

This paper is the first survey of SSP that includes the three major aspects of the domain (and the most important challenges they involve), namely, modeling, analysis, and synthesis of nonverbal behavior in social interactions. The modeling problem relates to studying laws and principles underlying social interactions and the role that nonverbal behavior plays in these (see Section 2). The analysis problem investigates the development of automatic approaches for extraction and interpretation of nonverbal behavioral cues in data captured with microphones, cameras, and any other suitable sensor (see Section 3). The synthesis problem addresses the automatic generation of appropriate nonverbal behavior via different forms of embodiment like conversational agents, robots, etc. (see Section 4). The survey considers as well the application domains where SSP has played, or is likely to play, a major role (see Section 5).

2 Nonverbal Communication and Social Signals

Investigation of nonverbal communication flourished in the early 1950s, stimulated by interest in semiotics as a field broader than mere linguistics [20], and supported by development of recording techniques. After a first pioneering study [21], the whole repertoire of nonverbal behaviors was surveyed in [22], and "*kinemes*" were proposed as an analog of phonemes and morphemes in linguistics to analyze body behavior [23]. Up to the 1970s, research was mainly devoted to unimodal communicative systems, in particular facial expressions [24], gaze [25], posture [26], and gestures [21].

Research on sign languages of the deaf showed that any sign can be described in terms of a small number of parameters (handshape, location, orientation, and movement) with respect to which it may assume specific values, analogous to phonemes of verbal languages [27]. This finding has been extended to symbolic gestures of hearing people [17], and to the systems of touch [17], gaze (eyebrows and eyelids positions and movements, eye direction, pupil dilation, eye humidity, and reddening) [17], head movements [28], and facial expressions [24]. A set of parameters cutting across different modalities, even if they have been so far only analyzed for gestures, concerns the "*expressivity*" and includes amplitude, fluidity, power, acceleration, and repetition [29]. These studies show how verbal and nonverbal communication systems share a similar structure and have, in some cases, a comparable degree of sophistication. The only aspect that seems to pertain mainly to verbal communication is syntax.

On the semantic side, while some scholars are cautious to attribute specific meanings to specific nonverbal signals, by maintaining that nonverbal behavior is more polysemous and context dependent than verbal language, others have identified recurrent correspondences between nonverbal signals and meanings. This has resulted in detailed semantic analyses of single gestures [17], and in dictionaries

of symbolic gestures of many cultures [30], but also in "*lexicons*" of touch and gaze [17].

At the end of the 1980s the importance of integrating multiple modalities in face-to-face communication came to be fully highlighted [31]. The growth point perspective [32] shows, e.g., that words and gestures constitute a bimodal integrated unit and that they are planned together, while experiments in [33] show how meaning is distributed across the two modalities and that gestures can be a cue to a speaker's cognitive processes.

Several studies tackle the issue of multimodality from the point of view of message production: Is gesture useful for the speaker, the listener, or both? According to [34], gestures are specially important for the speaker because they help conceptual processing and lexical retrieval; in [35] gestures are considered important mainly for the listener because they help comprehension and memory; in [17] gestures are considered useful to both speaker and listener. These works also put forward hypotheses on multimodal planning: some by relying on the model in [36], others by providing a computational model.

On the other hand, multimodality is studied from the point of view of its "*fusion*" in the perceiver's mind. How are signals in the different modalities perceived and integrated together? After the explosion of studies on the perception illusions like the McGurk effect, caused by an interference of a phoneme and an incongruent viseme [37], cognitive research has investigated multimodal integration (or fusion) of data [38], in particular face and voice [39]. Recent results [40] pointed out the cognitive and neural mechanisms involved in two types of integration of voice and facial expressions: complementary and correlated. In the former the visual signal gives information that is not the same as the auditory one, and adds to it, while in the latter the visual is a duplication of the auditory. Besides the process of meaning understanding, the multimodal (and crossmodal) studies identified neural processes involved in emotional expression, in particular of fear; stronger activation of left amygdala when face and voice are congruent has been observed in [41].

2.1 Toward a General Definition of Signals

Notions pertinent to define *signals* have been proposed in various disciplines. In the domain of "signal processing," a *signal* [42] is simply an analog or digital electrical representations of physical quantities varying in space or time; Information Theory defines *information* as a change in the probability value of some event [43]; in Linguistics and Semiotics a *sign* is an entity of two faces, a signifier (an acoustic or visual image) and a signified (a concept); in Ethology, a *cue* is any feature of the world or property of an organism that influences an animal's behavior [44]; in Psychology, a trait or state works as an *indicator*, encompassing single cues that, once received as a *percept*, are attributed information through a decoding process [45].

In a cognitive perspective [46], to plan and perform adequate actions humans need information that they draw, through perception, signification, and communication from other individuals' actions and properties, and from objects and events in the world. In *perception* a *physical stimulus* (a pattern of physical energy) is received by an individual's sensory apparatus and then dynamically modeled through the effects of gestalt laws [47] becomes a *percept*: It is information, but information, in a sense, only similar to itself; if I see a "tree branch on the ground" I only know there is a branch on the ground. *Signification* is the attribution of some "meaning" to some percept. If branch means "someone passed on the wood path," this is a second information I draw from "branch on the ground." In *communication*, signification is used by someone to convey meanings to others: My friend cut and dropped a "branch on the ground" to signal "he passed on the wood path." A *signal* is then information (a simple or complex percept produced by one or more *physical stimuli*) from which a Receiver can draw more information (*meaning*).

In communication, but sometimes also in bare signification, the meaning of a signal is reconstructed by relying on stable connections, "codified" on a cultural or biological basis, that are stored in the Receiver's long-term memory (and in communication are supposed to be shared by Senders and Receivers); in these "lexicons," both for verbal languages and items of some body communication systems, each signal is meaningful, yet polysemic, since it corresponds to a small set of possible meanings, out of which the most plausible one can be selected based on information coming from context.

Suppose two women are facing each other across a table, with one mirroring the other's head movements. From your past experience, this similarity of movements recalls either an instinctive way of showing empathy and affinity or a hypocritical manipulative way to comply with the other. To find which interpretation is more plausible you may resort to contextual knowledge: If the two often laugh and show a relaxed posture, they might be two close friends, but if they are in a luxurious office and one, older than the other, is sitting in a large armchair, this might be a job interview, with a young woman attempting an ingratiation strategy on the interviewer. Like in a mosaic, where adding new pieces makes the global picture clearer, context contributes to both disambiguate and enrich incoming information. Moreover, each signal, beyond its literal meaning—one drawn from the lexicon plus contextual interpretation—may have one or more indirect meanings: further implications, presuppositions, or other kind of inferences that can be drawn, again, through reasoning by the interaction of the literal meaning with previous knowledge.

We may thus distinguish *informative* and *communicative* signals: A signal is "communicative" if it is produced by a Sender to convey meaning to others (the speech acts perspective [17]); it is "informative" if the Receiver draws some meaning from a signal even if the one who produces it does not intend to convey a meaning (see the Semiotic perspective [20]). Not only humans, but also animals can emit both informative and communicative signals, with information conveyed, of course, at different levels of sophistication, depending on their evolutionary level.

Some informative signals, including ethology's cues, are not even, strictly speaking, "emitted" by anybody. That a video on YouTube is seen by millions of people "means" it is a very popular video, but this meaning is not drawn on the basis of each single click, rather of the *combination* of

many clicks on the video. In mimicry, where two people conversing mimic one anothers movements, a third observer might tell they are in syntony, but not from one or the other persons movements by themselves, rather from their both doing the same movement simultaneously or in close sequence; the signal is not their action but the *similarity of actions*. These are "honest signals" [48] since they cannot be faked or simulated: a particular type of informative signals. Also "honest" are the communicative signals that are not under conscious control, which leaves room for considering those governed by an unconscious goal of communicating, like mimicry—if seen on the part of the one who mimics one's interlocutor—and those regulated by biological goals of communicating, like the stickleback's reddening abdomen, which signals readiness to mate, or pupil dilation, a signal of sexual excitement that one cannot perform (or refrain from performing) on purpose.

"Honest signals" are thus an intersection of informative and communicative signals: "Honest" communicative signals are actions or morphological features determined by unconscious goals or biological functions; "honest" informative signals are not actions ascribable to some specific agent, but events, for example, combinations of simultaneous or sequential actions of different agents.

2.2 Social Signals and Social Facts

Taking these notions into account, social signals can be defined as follows: A Social signal is a communicative or informative signal that, either directly or indirectly, provides information about social facts, namely, social interactions, social emotions, social attitudes, or social relations.

2.2.1 Social Actions and Social Interaction

To define social interactions, a notion of "*social action*" must be defined first. An action of an Agent A can be defined as a social action if it is performed by A in relation to some Agent B (i.e., Agent B is mentioned in A's mental representation of that action) and if, while doing that action, A views B not as an object but as a self-regulated Agent, one having and pursuing goals of one's own [46]. Social interaction is a simple or complex event in which an agent A performs some social actions addressed at another agent that is actually or virtually present [13]. Face-to-face interactions include two main kinds of social actions, namely, those related to turn taking and back channel. Both kinds of actions aim at synchronization, i.e., at inducing mutual reactions between interaction participants and at negotiating each participants role in the conversation as that of a speaker or of a listener.

The turn taking system that governs a conversation is conveyed by nonverbal signals like mouth opening, gaze direction [49], or variation in vocal intensity; but the very structure of turns in a specific conversation is also a signal in itself in that it tells how friendly or competitive an interaction is. Overlapping speech and interruptions may be a cue of conflict [50], while the number and length of turns may inform about dominance patterns [51]. Backchannel is an even more effective means of synchronization as it informs the speaker whether others are listening, following, understanding [52], possibly believing, finding interesting, and agreeing with what someone says [17]. This helps interactants to adjust exposition and to take into account interlocutor's needs, thoughts, and points of view. Backchannel research has mainly considered hesitations, interjections, fillers, affect bursts [53], head movements [28], and smiles [54].

2.2.2 Social Attitudes

An attitude is a set of beliefs, evaluations, social emotions, social dispositions, tendencies to act, that together determine (and are determined by) preferences and intentions [13]. On the side of attitude expression, research has investigated the verbal and nonverbal signals that communicate one's evaluation of people, and hence the disposition to behave toward them. This includes, e.g., the expression of emotions like contempt [55] or signals of dominance [56]. In the same vein, laughter has been seen as a social signal of superiority and negative evaluation [57], [58]; in irony, it is not only the ironic smile but sometimes exaggerated body language or incongruence between different modalities signal teasing intent [58]. Yet, evaluation can also be conveyed in indirect ways, e.g., expressions of compassion or tenderness may be a cue to negative evaluation and overprotective attitude [59]. The literature on self-presentation and impression management [60] has investigated trust-inducing gestures and postures showing, e.g., which gestures, faces, and types of gaze can be aimed at persuading [61] or be effective at persuading [62]. Coming to the addressees response, typical signals of agreement and disagreement (corresponding to positive and negative evaluations, respectively) are, in one-to-many interaction, applause [63], in everyday conversation, head nods and head shakes, smiles, lip wipes, crossed arms, hand wagging, etc. [64].

2.2.3 Social Emotions

Social emotions are those related to social relations [65], for instance, pride, shame, or embarrassment or those felt toward someone else, like hate, envy, contempt, admiration [66], [67]. Some expressions of social emotions are social signals in their own right because they establish a specific relation to others. This is the case with cues displaying contempt [55], shame [68], but also of laughter [58], and smile, especially if it is not simply the result of a person being in a positive state, like in the more individual view of smile maintained by [69], but an attempt to show that the positive state depends on the presence of others, like in the more social view in [70].

2.2.4 Social Relations

A social relation is a relation between two (or more) people that have common or related goals, that is, in which the pursuit, achievement, or thwarting of a goal of one of these persons determines or is determined in some way by the pursuit, achievement, or thwarting of a goal of the other involved person [13], [71].

Different typologies of relations have been proposed in terms of criteria like public versus private, cooperation versus competition, presence versus absence of sexual relations, social-emotional support oriented versus task oriented [72]. Within group relations, some studies concern the definition and description of mechanisms of power, dominance, and leverage [65], [73], including the allocation, change, and enhancement of power relations (e.g., through alliance,

influence, and reputation [74]), the interaction between gender and power relations, and the nature of leadership.

Various types of relations exist, and different classes of signals convey different types of relations. Typical signals revealing social relations include the manner of greeting (saying "*hello*" signals the wish for a positive social relation, saluting signals belonging to a specific group like the army), the manner of conversing (e.g., formal allocutives like addressing someone as "*professor*" to signal submission), mirroring (signaling the wish to have a positive social relation, or displaying "*typical*" group's behavior), spatial positioning and gaze direction (e.g., making a circle around a certain person, or gazing at her more frequently distinguishes that person as the group leader [25]), physical contact (touching another person may indicate an affective relation [17]). For group relationships, both deliberate and unconscious signals, like regional accent, the manner of dressing or cutting one's hair, and mirroring, are typical signals revealing whether a person (feels to) belong to a specific group or not. The emblems on the clothes, how elaborate is a hair style or a crown, and the spatial arrangement of the members of a group typically reveal the rank (i.e., power relations) of different members in the group [75].

3 Automatic Analysis of Social Signals via Nonverbal Communication

In its most general form [13], an automatic approach for the analysis of social signals includes several steps. The first is *data capture*, performed in various settings and using different equipments, from simple laptop webcams to fully equipped smart meeting rooms [76], [77], and wearable devices [78], [79]. In most cases, the process of data capture results in signals (audio, video, etc.) that portray more than one person. This makes it necessary to perform *person detection*, i.e., to identify which segments of the captured data portray which person. This is the second step of the process and it involves technologies like face detection [80], speaker segmentation [81], tracking [82], etc. The data segments isolated during person detection carry information about the behavior of each interactant and it is from them that nonverbal behavioral cues are extracted. This is the third step of the process (*behavioral cues extraction*) and requires technologies like facial expression analysis [83], prosody extraction [84], gesture and posture recognition [85], etc. (see [13] for an extensive survey of techniques applied to all processing steps). At the end of the process, the automatically extracted behavioral cues are used in the last step (*social interaction interpretation*) to infer social signals. This is the aspect of the problem most specific to SSP and the rest of this section focuses on it.

3.1 State of the Art

The attention of the computing community toward automatic analysis of social signals has significantly increased during the last few years [13], [86] and many socially relevant phenomena have been investigated in a technological perspective (e.g., boredom, interest, understanding, confusion, engagement, leadership, etc.). The rest of this section focuses on some of those that, to the best of our knowledge, have received the widest attention, namely, the analysis of social relations (especially when it comes to the recognition of roles) and social attitudes (in particular dominance and personality as well as their effects in terms of interaction outcomes and conflict). The main details of the works discussed in the rest of this section are presented in Table 1.

3.1.1 Analysis of Social Relations: Role Recognition

Roles are a key aspect of social interactions: "*[...] interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability*" [87]. Two main approaches have been used for the recognition of roles, the analysis of speaking activity, and the modeling of lexical choices. In a few cases, the two approaches have been combined and some works propose movement-based features (fidgeting) as well, resulting in multimodal approaches based on both audio and video analysis. Turn-taking has been used in [88], [89], where temporal proximity of speakers is used to build social networks and extract features fed to Bayesian classifiers based on discrete distributions. Temporal proximity, and duration of interventions are used in [90], [91], [92], [93] as well, where they are combined with the distribution of words in speech transcriptions. Role recognition is based on BoosTexter (a text categorization approach) in [90], on the combination of Bayesian classifiers (working on turn taking) and Support Vector Machines (working on term distributions) in [92], and on probabilistic sequential approaches (Hidden Markov Models and Maximum Entropy Classifiers) in [91], [93]. An approach based on C4.5 decision trees [94] and empirical features (number of speaker changes, number of speakers talking in a given time interval, number of overlapping speech intervals, etc.) is proposed in [95]. A similar approach is proposed in [96], where the features are the probability that someone starts speaking when everybody is silent or when someone else is speaking. Role recognition is performed with a Bayesian classifier based on Gaussian distributions. The only multimodal approaches are proposed in [97], [98], where features accounting for speaking activity and fidgeting are recognized using Support Vector Machines first [97], then replaced with influence models to exploit dependencies across roles [98]. Even if they use fidgeting features, these two works still suggest that audio-based features are the most effective for the recognition of roles.

3.1.2 Analysis of Social Emotions

It is interesting to note that while the state of the art in machine analysis of basic emotions such as happiness, anger, fear, and disgust is fairly advanced, especially when it comes to analysis of acted displays recorded in constrained lab settings [83], machine analysis of social emotions such as empathy, envy, admiration, etc., is yet to be attempted. Although some of the social emotions could be arguably represented in terms of affect dimensions—valence, arousal, and dominance—and pioneering efforts toward automatic dimensional and continuous emotion recognition have been recently proposed [99], a number of crucial issues need to be addressed first if these approaches to automatic dimensional and continuous emotion recognition are to be used with freely moving subjects in real-world

TABLE 1
The Table Provides the Following Details about the Analysis Works Discussed in This Survey: Data Set Used for Experiments (Including Its Length Whenever Such an Information Is Available) and Performances Achieved

Article	Data	Performance
Role Recognition		
[89]	Broadcast+AMI (90h)	80% frame accuracy
[90]	Broadcast (17h)	80.0% story accuracy
[91]	Broadcast (17h)	77.0% story accuracy
[92]	AMI (45h)	67.9% frame accuracy
[95]	Meetings (45m)	53.0% analysis segments accuracy
[96]	AMI (45h)	53% frame accuracy
[98]	MSC (4h.30m)	75% role assignment accuracy
[121]	MSC (4h.30m)	90% analysis segment accuracy
Personality		
[103]	AMI-40	70% dominance level recognition rate
[104]	AMI and M4 subset (95m)	75% dominance level recognition rate
[105]	AMI subset (5h)	80% dominant person recognition rate
[108]	Smart badge data (3096h)	correlation between features and personality traits
[109]	640 audo clips (330 identities, 1h46m)	between 57% and 76% correct assessment prediction depending on trait
Analysis of (Dis-)Agreement		
[118]	Canal9 (43h)	66% (dis-)agreement recognition rate
[119]	ICSI subset (8094 talk spurts)	78% (dis-)agreement recognition rate
[120]	ICSI subset	86.9% (dis-)agreement recognition rate
Group Dynamics		
[122]	AMIsubset+broadcast data (4h20m)	100% conversational dynamics recognition rate
[123]	38 conversations	75% conversation setting recognition rate
Negotiation outcome and coordination		
[112]	dyadic interactions	70% correct interaction outcome prediction
[116]	7 human-robot conversations	∼95% head nod and shake correct detection
[117]	10 cellular phone conversations	measures of gait alignment

scenarios like patient-doctor discussions, talk shows, job interviews, etc. In particular, published techniques revolve around the emotional expressions of a single subject rather than around the dynamics of the emotional feedback exchange between two subjects, which is the crux in the analysis of any social emotions. Moreover, the state-of-the-art techniques are still unable to handle natural scenarios such as incomplete information due to occlusions, large and sudden changes in head pose, and other temporal dynamics typical of natural facial expressions [83], which must be expected in human-human interaction scenarios in which social emotions occur.

However, social emotions have attracted attention in the Natural Language Processing (NLP) community, where they have been the subject of sentiment analysis, the domain aimed at recognition of opinions and subjective feelings in written texts, often collected from social media such as mailing lists, blogs, forums, social networking sites, etc. Proposed approaches make typically use of technologies originally developed in other NLP domains (e.g., information retrieval and text categorization) and, to the best of our knowledge, do not involve nonverbal behavioral aspects [100].

3.1.3 Analysis of Social Attitudes: Dominance, Personality, and Their Effects

In every social context, there are people that tend to have higher impact on development and outcomes of interactions [101]. These people are said to be *dominant* and several automatic approaches have been aimed at their identification [102], [103], [104], [105]. Speaking activity (speaking time, number of turns, interruptions, etc.) and Support Vector Machines have been used to map people into three dominance classes (low, normal, and high) in [103], [104]. The same speaking-related features and gaze behavior (who looks at whom) have been modeled in [102] with a Dynamic Bayesian Network. Another multimodal approach has been proposed in [105], where speaking activity (e.g., number of turns, histograms of turn durations, successful interruptions, etc.) and motion-based features (e.g., time during which a person moves, number of time intervals during which a person moves, etc.) have been fed to Support Vector Machines to identify the most dominant person in meetings. Like in the case of role recognition, speaking activity related features appear to be the most effective, though some improvements are always obtained when they

are combined with vision-based features like movement or gaze.

Dominance depends, to a large extent, on the social role an individual is playing. However, it can be considered as well as one aspect of *personality*, the latent construct accounting for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms—hidden or not—behind those patterns*" [106]. In order to make computers capable of dealing with a wider spectrum of situations, the literature proposes several approaches for the inference of personality traits from nonverbal communication. In [107], SVMs classify audio and video feature vectors (including mean of pitch, energy and spectral entropy, fidgeting, etc.) into classes accounting for two personality traits (extraversion and locus of control), and [108] estimates the correlation between the same traits and features captured via wearable sensors like movement, proximity with others, speaking activity (energy, amount of speaking time, etc.), centrality, and betweenness in social networks. In both works personality traits are self-assessed, i.e., judged by the same people whose personality is being measured. In contrast, the works in [109], [110] predict the way prosodic features influence the perception of personality, namely, the way traits are perceived by others. Both works use machine learning algorithms (e.g., SVMs) to map basic prosodic features (e.g., pitch and energy) into personality assessments made in terms of the Big Five, the most important and common personality model [111].

Social attitudes have an important effect on interaction outcomes and the emergence of conflicts. Both phenomena have been addressed in the literature. The prediction of negotiation outcomes has been proposed in [112], where features accounting for speaking activity, consistency (stability of speaking features), influence (statistical dependence of a speaker on the other one), and mimicry (see below) predict with an accuracy of 70 percent the result of salary negotiations, hiring interviews, and speed dating conversations. Mimicry and coordination play a major role in establishing (and accounting for) a good quality of rapport [113], [114], [115], and several approaches aim at their measurement and detection. Coordination is used in [116] to improve the recognition of head gestures of people interacting with virtual agents and robots. Automatic measurements of coordination are performed in [117], where the gait alignment of people talking via cellular phones is measured (using oscillation theory) through accelerometers embedded in phones.

Given the impact that conflicts can have on the life of a group [101], a topic that is attracting increasing interest is the detection of agreement and disagreement [64]. In [118], a Markov Model captures the tendency that people have to react to one another when they disagree and reconstructs the fronts opposing one another in political debates. A similar approach is applied in [119], [120], where pairs of talk spurts (short turns) are first modeled in terms of lexical (which words are uttered), durational (length, overlapping, etc.), and structural (spurts per speaker, spurts between two speakers, etc.) features and then classified as expressions of agreement or disagreement with a Maximum Entropy Model.

3.2 Data and Resources

A relatively large number of databases is available for SSP purposes (see [86] for an extensive survey of this aspect). However, the collection of appropriate SSP corpora faces three main problems:

- The domain is still in its early stages [13] and no major efforts have been done yet for the collection of data *specifically* aimed at the analysis of social phenomena. Most of the works in the literature use data originally aimed at different purposes (e.g., broadcast material collected for Information Retrieval) and annotated ad hoc for analyzing some specific social phenomena (e.g., the subset of the AMI Meeting Corpus annotated in terms of dominance while originally aimed at speech recognition and computer vision goals [103], [124]). This negatively influences the ecological validity of the data and limits the spectrum of social phenomena that can be investigated.
- Social interactions involve a large variety of aspects and no standard annotation or data collection protocol seems to be possible. In other words, each social phenomenon seems to require the collection of a specific corpus. This effect can be limited by designing scenarios where several social phenomena take place at the same time, but a large number of corpora will still be necessary to cover all possible aspects of social interaction.
- The annotation of behavioral data in terms of social signals should be performed by a number of assessors sufficiently large to ensure replicability, i.e., to ensure that the agreement between independent assessors is statistically significant (at least 10 annotators, following a commonly applied thumb rule) [125]. This makes the collection of corpora suitable for rigorous scientific research expensive and time consuming. Furthermore, not all of the corpora currently used in the literature actually respect the requirement above.

Table 2 summarizes the main characteristics of the most important data corpora described in the literature so far [86], including the number of subjects involved in recorded interactions (figures in bold are average values), the number of recordings, and the total duration of all recording constituting the corpora in question. Furthermore, the table reports the availability of main data modalities (Audio and Video) and the available data annotation in terms of speaker segmentation, speech transcripts, roles, dominance, and personality. Some of the listed data corpora also include other modalities and annotations. For example, ICSI [126] includes annotation in terms of dialogue acts and interest level, M4 [127] is annotated in terms of interest and turn taking types, AMI [128] provides slides, hand-written and whiteboard notes, and it is annotated in terms of dialogue acts as well, the AMI-12 [124] includes information about subjects focus of attention and about who addresses whom, VACE [129] is annotated in terms of visual focus of attention, ATR [130] includes annotation in terms of body movements and turn taking types, and the Canal9 corpus

TABLE 2
This Table Reports the Characteristics of the Most Important Data Collections Currently Used for Analysis of Social Interactions

Corpus	Group Size	Items	time	Audio	Video	Speaker Segmentation	Speech Transcripts	Roles	Dominance	Personality
ISL [132]	**6.4**	104	103h	√	√	√	√			
ICSI [126]	**6**	75	72h	√		√	√			
M4 [127]	4	60	5h	√	√	√	√			
NIST [133]	**5.4**	19	15h	√	√	√	√			
AMI [128]	4	167	100h	√	√	√	√	√		
AMI-12 [124]	4	12	6h	√	√	√	√	√	√	
AMI-40 [103]	4	40	20h	√		√	√	√	√	
NTT [102]	4	4	22m	√	√	√				
VACE [129]	5	NA	NA	√	√	√	√		√	
ATR [130]	4-9	10	10h	√	√	√		√		
MSC-1 [121]	4	11	3h 45 m	√	√	√		√		
MSC-2 [134]	4	13	7h	√	√	√		√		√
Canal9 [131]	5	70	43h	√	√	√		√		

The figures of this table, as well as those reported in this section, are courtesy of Daniel Gatica-Perez and are available in [86]).

[131] provides information about shot segmentation, shot types, agreement and disagreement, and identity of speakers at each turn. An important collaborative effort toward the collection of resources useful for research in social signal processing is being done by the European project titled the Social Signal Processing Network of Excellence (SSPNet). The projects web-portal (www.sspnet.eu) provides three kinds of resources: *Knowledge* (an extensive bibliography covering various aspects of Social Signal Processing), *Data* (a large variety of publicly available corpora directly accessible through the portal), and *Tools* (a collection of publicly available software packages addressing a wide range of needs in Social Signal Processing) [135].

3.3 Challenges

Nonverbal behaviors like social signals cannot be read like words in a book [136], [7]; they are not always unequivocally associated to a specific meaning, although according to someone they generally are [17], and their appearance can depend on factors that have nothing to do with social behavior. For example, some postures correspond to certain social attitudes, but sometimes they are simply comfortable [137]. Similarly, physical distances typically account for social distances, but sometimes they are simply the effect of physical constraints [138]. Moreover, the same signal can correspond to different social behavior interpretations depending on context and culture [139], although many advocate that social signals are natural rather than cultural [140]. In other words, social signals are intrinsically ambiguous, high-level semantic events, which typically include interactions with the environment and causal relationships.

An important distinction between the analysis of high-level semantic events and the analysis of low-level semantic events like the occurrence of an individual behavioral cue like the blink is the degree to which the context, different modalities, and time must be explicitly represented and manipulated, ranging from simple spatial reasoning to context-constrained reasoning about multimodal events shown in temporal intervals. However, most of the present approaches to machine analysis of human behavior are neither multimodal nor context sensitive nor suitable for handling longer time scales [8], [13], [83], [141]. Hence, the focus of future research efforts in the field should be primarily on tackling the problem of context-constrained analysis of multimodal behavioral signals shown in temporal intervals. As suggested in [8], [141], this problem should be treated as one complex problem rather than a number of detached problems in human sensing, context sensing, and human behavior understanding.

More specifically, there are a number of scientific and technical challenges that we consider essential for advancing the state of the art in machine analysis of human behavior like social signals.

3.3.1 Modalities

Which behavioral channels, such as the face, the body, and the tone of the voice, are minimally needed for realization of robust and accurate human behavior analysis? Does this hold independently of the target communicative intention (e.g., social interactions/emotions/relations) to be recognized? No comprehensive study on the topic is available yet. What we know for sure, however, is that integration of multiple modalities (at least facial and vocal) produces superior results in human behavior analysis when compared to single-modal approaches. Numerous studies have theoretically and empirically demonstrated this (e.g., see the literature overview in [142] for such studies in psychology, and the literature overview in [83] for such studies in automatic analysis of human behavior). It is therefore not surprising that some of the most successful works in SSP so far use features extracted from multiple modalities (for an extensive overview of the past works, see [13]). However, other issues listed above are yet to be investigated. Also note that some studies in the field indicate that the relative contributions of different modalities and the related behavioral cues to judgment of displayed behavior depend on the targeted behavioral category and the context in which the behavior occurs [142].

3.3.2 Fusion

How should we model temporal multimodal fusion which will take into account temporal correlations within and between different modalities? What is the optimal level of integrating these different streams? Does this depend on the time scale at which the fusion is achieved? What is the optimal function for the integration? More specifically, most of the present audiovisual and multimodal systems in the field perform decision-level data fusion (i.e., classifier fusion) in which the input coming from each modality is modeled independently and these single-modal recognition results are combined at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the assumption of conditional independence between audio and visual data streams in decision-level fusion is incorrect and results in the loss of information of mutual correlation between the two modalities. To address this problem, a number of model-level fusion methods were proposed that make use of the correlation between audio and visual data streams and relax the

requirement of synchronization of these streams [83]. However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is yet to be explored.

3.3.3 Fusion and Context

Does context-dependent fusion of modalities and discordance handling, which are typical for fusion of sensory neurons in humans, pertain in machine context sensing? Note that context-dependent fusion and discordance handling were never attempted within an automated system. Also note that while W4 (where, what, when, who) is dealing only with the apparent perceptual aspect of the context in which the observed human behavior is shown, human behavior understanding is about W5+ (where, what, when, who, why, how), where the why and how are directly related to recognizing communicative intention including social signals, affect, and cognitive states of the observed person. Hence, SSP is about W5+. However, since the problem of context sensing is extremely difficult to solve, especially for a general case (i.e., general purpose W4 technology does not exist yet [8]), answering the why and how questions in a W4-context-sensitive manner when analyzing human behavior is a virtually unexplored area of research. Having said that, it is not surprising that context-dependent fusion is truly a blue-sky research topic.

4 Synthesis of Social Behavior

Synthetic social behavior is a crucial need for artificial agents, robots, intelligent interfaces, and any other kind of device capable of interacting with users like people interact with one another. Agents and robots should be able to show, through their artificial bodies, a similar range of verbal and nonverbal signals of their emotional state and their social stance as humans through their voice, face, and other body parts. The synthesis of the actual behaviors is driven by the artificial mind, the AI, of the agent. There are many aspects of social behaviors that have been studied and implemented over the years. Starting with the work on social talk by the REA agent [143], the field has developed to look at a broader range of social skills such as empathy, rapport, and politeness.

4.1 State of the Art

Several works aimed at automatic generation of social actions (e.g., turn taking, backchanneling), social emotions, and social attitudes (e.g., politeness) have been recently proposed in the literature.

4.1.1 Synthesis of Social Actions

As conversation is considered the "primordial site of human sociality and social life" [144], the research community has focused in particular on the synthesis of social actions in talk-in-interaction. One of the most salient aspects of conversations is turn taking and the literature proposes many works aimed at the synthesis of social actions related to this phenomenon (for the important role of turn taking in analysis of social interactions see Section 3). The model proposed in [145] predicts a change of gaze direction during the transitions from one turn to the following. Furthermore, the same work has shown the correlation between posture changes and intonational structure by analyzing a collection of videos. The approach in [146] integrates a perception-action loop to generate real-time turn taking mechanisms. A later version of the same approach [147] uses parallel neural networks to select actions to be performed. This approach looks at turn taking as a process of coordination between two parties. Another model based on a similar view simulates turn taking behaviors using an imitation model [148].

During conversations, virtual agents act not only as speakers, but also as listeners and they should not freeze when they do not talk. Social actions related to back-channel (e.g., head nodding and utterances like "ah-ah," "yeah," etc.) are important not only to make the agent look more alive in a conversation, but they are also cues to the level of engagement of the listener; they signal its attitude toward what the speaker is saying and they allow the creation of rapport between interactants. Most models are based on the acoustic and/or visual analysis of speaker's movement and voice.

They are either rule-based, specifying when a back-channel is triggered depending on the signals emitted by the speaker [149], or stochastically computed [150]. In the latter case, learning algorithms have been applied to extract predictive models of the correlation between speakers visual and/or acoustic cues and backchannel productions [150]. Other approaches also take into account a semantic analysis of what the speaker is saying. When coupled with a model of the agents mental state, these models ensure that the agent displays coherent and appropriate backchannel signals [151].

During their turns, virtual agents are expected not only to talk, but also to generate basic social actions such as laughs, sighs, or expressive feedback utterances. Various approaches have been proposed for synthesizing laughter, including unit selection [152] and models inspired from physics [153]. By collecting dialogues between a speaker and a synthetic voice, [154] prepared the generation of a richer set of vocalizations together with synthetic speech. Only limited evidence is available to date regarding the suitable *use* of such vocalizations, however: In [155], it is shown that laughs are perceived to differ in suitability for a given synthetic dialogue; some laughs were considered to be completely inappropriate for the given context. Schröder et al. [53] investigated the acceptability of using affect bursts [156], [157] as listener feedback in a dialogue, and explained suitability ratings in terms of socio-cultural *display rules* [158].

Last, but not least, social actions include the use of space and mutual position as a social cue. Several synthesis approaches take into account this aspect, especially when it comes to the simulation of group interactions. The formation of a group follows specific patterns [159] that dynamically evolve to include (or not) newcomers and to adapt after one or more interaction participants leave. The notion of human territory [160] and the *F-Formation* [159] are applied in [161] to animate groups of agents in a virtual world. In [162], the dynamic formation of multiparty interactions is simulated using the model of social force field that has been developed for human crowds modeling

[163]. These different models implement proxemics between virtual agents. They act not only on the distance between agents and their body orientation toward one another but also on the gaze patterns of the participants.

4.1.2 Synthesis of Social Emotions

Emotion has been synthesized using both artificial voices and faces. In the first case, the most common approaches are based on explicit rules mapping state to be expressed (the emotion) into expressive parameters (voice prosody). In [164], such a rule-based approach is combined with Multiband Resynthesis Overlap Add (MBROLA) diphone synthesis to generate a synthetic voice with a degree of emotionality according to the emotion dimensions activation, evaluation, and power. In [165], the same combination lends a personality to voice-enabled products in a futuristic shopping scenario by means of rules representing the intended vocal correlates of the various personalities. Other rule-based approaches such as formant synthesis also use explicit rules for realizing different emotional expressions (e.g., in [166]). However, the substantially higher quality of data-driven synthesis technologies (see the paragraph on expressive speech below) has sidelined rule-based expressive synthesis research. It remains to be seen whether statistical models can be combined with rule-based methods so as to combine a high-quality baseline with the control needed for the intended expressivity.

4.1.3 Synthesis of Social Attitudes: Politeness and Expressive Speech

To the best of our knowledge, the first attempt at implementing politeness strategies in virtual agents was made in [167], with a recent follow-up in [168]. In these works, the desired level of politeness of an utterance depends on the social distance between the dialogue participants, the power one has over the other, and the estimated face threat posed by a speech act. Similar works [169], [170], [171] aim at generating tutoring responses, based on the politeness theory presented in [172]. These systems synthesize politeness based on static input parameters, rather than on dynamic user models updated during interaction. This problem is overcome in [173] where a "*Virtual Guide*" is equipped with an adaptive politeness model (based again on the theory in [172]) that dynamically determines the user's politeness level during the dialogue and lets the "*Virtual Guide*" adapt its politeness level accordingly: A politely worded request for information will result in a polite answer, while a rudely phrased question will result in a less polite reaction.

The politeness theory in [172] has been extended to communicative gestures in [174] and to facial expressions in [175]. In these works, video corpora were analyzed and annotated in terms of politeness strategies and multimodal behaviors, such as gesture types (iconic, metaphoric, etc.) [174] and facial expressions (of felt, inhibited, masked, fake emotions) [175]. An approach proposed in [176] models social role awareness and introduces a set of procedures, called "*social filter programs*," that take as input parameters politeness strategies, personality of the interlocutors, and their emotions. The output of the filter is the intensity of the facial expressions of emotions to be displayed by the agents.

Research on the generation of vocal social signals has mostly focused on generating high-quality expressive speech in a flexible way [177]. High-quality speech output can be obtained using *unit selection* synthesis techniques [178], which generate arbitrary speech output by resequencing small snippets of speech recordings according to a linguistically defined target utterance. Since the expressivity in the recordings is preserved during resequencing, it is possible to generate any expressive style for which a sufficiently large speech database can be recorded. Examples include an expressive tone suitable for presenting *good news* versus *bad news* [179]; in a military scenario, *commands* versus *conversation* [180]; or a creaky voice suitable for a poker player game character [181]. The major downside of this approach is the lack of flexibility: For every expression to be generated, a full speech synthesis corpus must be recorded, which is time consuming and costly. Therefore, alternatives are being investigated which, while staying in a data-driven paradigm, increase the flexibility. By using signal modification techniques such as pitch-synchronous overlap-add (PSOLA), it is possible to change the prosody of a synthesized utterance [182], however, at the cost of degraded quality. Voice conversion techniques can be used to change the expression of synthesis output [183], e.g., from a neutral to an expressive speaking style. Intermediate expressions, such as a medium intensity of anger, can be generated by interpolating between a neutral and an expressive rendition of a given target utterance [184].

In statistical-parametric speech synthesis, statistical models trained on speech synthesis recordings are used to predict context-dependent acoustic parameters for a target utterance, and a vocoder is used to generate the corresponding audio [185]. Style-specific voice databases can be trained in a similar way to style-specific unit selection voices [186]. In addition, by introducing a *style control technique* [187], it is possible to interpolate between styles, and even to exaggerate a speaking style [188]. Model adaptation is a method to reduce the amount of expressive speech material required [189], compared to traditional training of a voice on the expressive material.

Both approaches, unit selection and statistical-parametric synthesis, rely on training data to yield a certain expression. In both cases, expressivity is solely determined by the speech material used, and is global throughout the speech. Local effects such as emphasis on an individual word are not easy to generate in data-driven synthesis, and, again, seem to depend on suitable training data. Acoustic models of emphasis were trained in [190] using a partially annotated database and the models were used to extend the annotation to the unannotated part of the data. In a listening test, they obtained a very moderate degree of preference for emphasized over nonemphasized test sentences. With carefully designed and recorded training material, [191] obtained higher preference rates.

4.2 Data and Resources

To the best of our knowledge, no databases have been created for the synthesis of social signals, with the only exception being SEMMEL [174], a corpus aimed at the study of nonverbal behaviors in relation with politeness strategies. In contrast, several Embodied Conversational

Agent platforms and voice synthesizers are publicly available and constitute an important resource for the development of behavior synthesis approaches. The system *RUTH* [192] allows the control of a talking head through a precise behavior language and prosodic parameters. The agent systems *Cadia* [193], *Greta* [194], *SmartBody* [195] are *SAIBA* compliant [196]. SAIBA is a three-stages agent platform specification: The first stage corresponds to intention planning of the agent, the second instantiates the intention into multimodal behaviors, and the third computes the synchronized acoustic and visual signals the agent displays. Communicative and emotional data are encoded with *Function Markup Language* (FML) [197] and behavior specifications are encoded with *behavior Markup Language* (BML) [196], [198]. Cadia and SmartBody work with BML. These agent platforms allow the animation of a virtual agent and can be plugged in interactive applications where users dialog with agents. An extensible, standards-based framework for building such applications is the open-source SEMAINE API [199].

Several speech synthesizers are publicly available as well, e.g., *Festival* [200], *OpenMary* [201], and *Euler* [202].

4.3 Challenges

Automatic synthesis of social signals targets a human observer's or listener's perception of socially relevant information. While it may be true that much of social behavior goes unnoticed [203], it appears that social signals still have an effect in terms of unconscious perception [204]; without being able to say exactly why, we either consider a person trustworthy, competent, polite, etc., or not. In automatic behavior synthesis, the aim is thus to create this perception by timely generating suitable signals and behaviors in synthetic voices, facial expressions, and gestures of an Embodied Conversational Agent (ECA). This faces two major problems:

- Too little is known about the types of socially relevant information conveyed in everyday human-to-human interactions, as well as about the signals and behaviors that humans naturally use to convey them. A first step in this direction would be to acknowledge the complexity of the phenomena, as has been done for emotion-related communication [205]. Then, different contexts and effects could be studied based on suitable data, and the findings could be described in terms of explicit markup language or in terms of statistical, data-driven models [206].
- It is not self-evident that synthetic agents should behave in the same way as humans do or that they should exhibit faithful copy of human social behaviors. On the contrary, evidence from the cartoon industry [207] suggests that, in order to be believable, cartoon characters need to show strongly exaggerated behavior. This suggests further that a tradeoff between the degree of naturalness and the type of (exaggerated) gestural and vocal expression may be necessary for modeling a believable ECA's behavior.

Certain aspects of social signals are particularly relevant and challenging when it comes to synthesis of human-like behavior.

4.3.1 Effect of Unconscious Processes

One of the main problems facing the synthesis of social signals is the lack of knowledge about the way social information is conveyed in everyday interactions. One of the reasons is that much of the Social Signal Processing in humans is done automatically and unconsciously and not accessible to introspection [208], [209]. This also poses a methodological problem in the design of behaviors for synthetic agents and the evaluation of the behaviors through perception studies. A subtle difference in timing that goes unnoticed may result in a different effect. One of the challenges for social signal synthesis is to design and carry out experiments that bypass these pitfalls.

4.3.2 Continuity

Unlike traditional dialogue systems in which verbal and nonverbal behavior is exhibited only when the system has the "turn," socially aware systems tend to be continuous in terms of nonverbal behavior to be exhibited. In any socially relevant situation, social signals are continuously displayed, and lack of such displays in an automatic conversational system is interpreted as social ignorance [13].

4.3.3 Complexity and Context

Relationships between social signals and their meaning are intrinsically complex. First, the meaning of various signals is often not additive: When signals with meanings x and y are shown at the same time, the meaning of this complex signal may not be derivable from x and y alone. In addition, context plays a crucial role for the choice and interpretation of social signals. For example, environmental aspects such as the level of visibility and noise influence the choice of signals to be shown. On the other hand, societal aspects such as the formality of the situation and previously established roles and relations of the persons involved, and individual aspects such as the personality and affective state influence not only the choice of signals to be shown but the interpretation of the observed signals as well. Hence, context-sensitive synthesis of human behavior is needed but it still represents an entirely blue-sky research topic.

4.3.4 Timing

Social signals are not only characterized by the verbal and nonverbal cues by means of which they are displayed but also by their timing, that is, when and for how long they were displayed in relation to the signals displayed by other communicators involved in the interaction. Thus, the social signals of an ECA need to be produced in anticipation, synchrony, or response to the actions of the human user with whom the character engages in the social interaction. This requires complex feedback loops between action and perception in real-time systems.

4.3.5 Consistency

In general, it appears that human users are very critical when it comes to the consistency of a virtual character [210]. This relates to the challenge of multimodal synchronization, that is, to timing between facial expression, gesture, and voice conveying a coherent and appropriate message. Research on this aspect is still ongoing. There is no

consensus on whether multimodal cues need to be fully synchronized, whether the redundancy of information coming from multiple cues is required, or whether it is also possible for one modality to compensate for the lack of expressiveness in other modalities (e.g., [211]). Consistency may also play a role in Mori's notion of an "*uncanny valley*" [212]—a robot that looks like a human but does not behave like one is perceived as unfamiliar and "strange." Similarly, behavior that may be consistent with a photo-realistic character may not be perceived as natural for a cartoon-like character and vice versa.

Even when it is clear what signals and behaviors to generate, a practical challenge remains: Current technology still lacks flexible models of expressivity and it usually does not operate in real time. Expressive synthetic speech, for example, is a research topic that despite two decades of active research is still somewhat in its infancy [177]. Existing approaches are either capable of domain-specific natural-sounding vocal expressivity for a small number of possible expressions, or they achieve more flexible control over expressivity but of lower quality. Similarly, fully naturalistic movements of virtual agents can be attained when human movements recorded using motion capture technology are played back [213], but movements generated based on behavior markup language [196] tend to look less natural [214]. These problems are not specific to synthesis of social signals, and they do not form insurmountable obstacles to research; however, they slow down the research by making it substantially more time consuming to create high-quality examples of the targeted expressions. Given the above-mentioned importance of timing, the lack of real-time systems impedes the realization of timely appropriate social behaviors. Even a slight delay in the analysis and synthesis of signals hinders dynamic adaptation and synchrony that are crucial in social interaction. Furthermore, the technological limitations pose serious difficulties for exploitation of research results in end-user applications, where fast adaptation to new domains is an important requirement. Therefore, enhancing the existing technology remains an important challenge facing the researchers in the field, independently of whether the aim is to develop socially adapt ECAs or robots with no need of social awareness.

5 Applications of SSP

The business community has recently recognized the urge for automatic systems dealing with social signals. The pioneering contributions described in [48], for instance, are identified as a breakthrough that will change management practices as deeply as the microscope has changed medicine and biology few centuries ago [215]. The reason is that social signals reveal the invisible aspects of social interaction, i.e., those aspects that are perceived and elaborated outside conscious awareness, but still influence human behaviors as much as the visible aspects, i.e., meaning and reasoning, erroneously believed to be the only important factors in social exchanges. In the same vein, one of the main applications of SSP so far, *Reality Mining* [78], has been identified as one of the 10 technologies likely to change the world in the near future [216].

However, the spectrum of application domains that can benefit from socially intelligent machines is still wide and applications based on SSP are just at the beginning of their history. Some application domains have already tried to introduce a social intelligence component (e.g., human-robot interaction) while others recognize it as need, but still lack it in their mainstream approaches (e.g., multimedia indexing).

The rest of this section presents some application domains where socially intelligent machines can play an important role and SSP is likely to have a significant impact in the next years.

5.1 Multimedia Indexing

Social interaction is one of the main channels through which we access reality [6] and, not surprisingly, information about people is one of the elements we retain most in multimedia data we consume (pictures, videos, e-mails, etc.) [217]. Thus, to represent (i.e., to index) the content of multimedia material in terms of the social interactions they portray means to bring information retrieval systems closer to our social intelligence, with potentially high improvements in terms of retrieval performance. Some attempts of indexing multimedia data in these terms have already been made (see, e.g., [218]), but extensive evaluations of how this impacts the retrieval performance are still missing. As it aims at social interaction understanding, SSP is likely to have a significant impact on this application domain.

5.2 Implicit Human-Centered Tagging

One of the most promising frontiers in multimedia retrieval is the use of nonverbal behavioral feedback as a source of information about the content of the data people consume (e.g., videos eliciting laughter should be categorized as *funny* or *comedy*) [219]. Several approaches based on such an idea have been recently proposed (see, e.g., [220]), but they are all at a rather early stage. In particular, the data sets used in these works are way too small to be considered representative of real-world application environments. SSP is likely to play a major role in this emerging domain as it involves behavior analysis as one of its major components.

5.3 Mobile Social Interactions

Cellular phones are among the most pervasive technologies (the large majority of individuals in developed countries carry their phone during the whole day) and they are reshaping the way people interact with one another [221]. So far, cellular phones have been used to perform macroanalysis of large social networks [222], but micro-analysis approaches for conversations taking place through cellular phones are still in a pioneering stage. Some works have shown that people talking through cellular phones tend to coordinate their gait [117] and interact in virtual spaces with the help of location devices embedded in their phones [223], but SSP-inspired approaches can certainly extend the spectrum of social phenomena that can be automatically analyzed in mobile scenarios. This is expected to have a major impact on the design of cellular phones and, more generally, portable devices [79].

5.4 Computer Mediated Communication

Remote communication is still far less natural than face-to-face interaction. Current video-conferencing systems do not take into account social phenomena [9], and the research in the domain has focused mainly on the creation of shared workspaces, while considering only gaze contact a cue important enough to be transmitted [224]. More recently,

there have been attempts to use virtual characters embodying social behaviors [225]. SSP can improve current technologies by improving understanding of ongoing social interactions and by synthesizing social behaviors at distance to guarantee quality of rapport in remote interactions.

5.5 Human-Computer Interaction

A large body of evidence shows not only that we display the same nonverbal behavioral cues whether we interact with other humans or with machines, but also that we unconsciously attribute human characteristics (e.g., personality, intentions, relational attitudes, etc.) to machines we interact with [18]. As SSP aims at automatic understanding and generating nonverbal behavioral cues, it is likely that it will have a major impact on the design of computer interfaces (and, in general, human-machine interaction) expected to accommodate human natural modes of interaction and to be socially adept when interacting with users [8].

5.6 Marketing

Nonverbal communication plays a major role in customer-seller interactions. The customer's perception of the sales person's personality, motivations, and trustworthiness influences significantly the decisions of customers [226]. In a similar way, nonverbal aspects of people portrayed in advertisement are known to have an impact on consumer behavior (see, e.g., [227]). Furthermore, self-presentation issues tend to influence nonverbal behavior of consumers in focus groups (one of the most important instruments in marketing) and bias the responses consumers provide toward expectations of focus group organizers [228]. These are but a few evidences of the importance of nonverbal behavior in marketing and thus of the potential impact automatic approaches for its understanding and generation (the goal of SSP) can have in this domain.

5.7 Social Signals and Social Simulations

In virtual worlds such as the well-known Second Life, people interact through embodied representations of themselves. Just as in real life, being able to communicate the proper social signals through body language is important in these mediated forms of interaction. Several people have started to investigate the automatic generation of proper nonverbal behavior in such worlds [229]. Interactions with avatars and virtual humans is not restricted to entertainment sites such as Second Life, but can also be found in serious games for language and culture training or for training other social skills [230].

5.8 Human-Virtual Agents Interaction

Virtual agents can play different roles. They can be a companion, mentor, coach, tutor, etc. In each of these roles, it is important for the agent to display appropriate social cues as well as to perceive them from the user. Through social cues, the agent can display its engagement with the user. These signals should evolve dynamically as the interaction evolves through time. To build a long-term relationship, the agent will need to display cues of strong ties and friendship.

5.9 Social Robots

Since the very early development of robots, researchers have been interested in robots endowed with social intelligence. Such robots ought to be endowed with the capacity to perceive and interpret their surroundings, and to communicate and engage with humans [231]. In this overview, we will deal exclusively with the synthesis of social signals for agents and ignore the synthesis of such signals in robots. Although, to a great extent, the problems and solutions are similar in both cases, one should note that there are also important differences. The physicality of the robot, its presence in space leads to other affordances and another type of interaction, which are an important factor in studies on proxemics, for instance. Several papers have been dedicated to surveys on socially interactive robots [231].

6 Conclusions

There is no doubt that automatic analysis and synthesis of human behavior has attracted major interest in the computing community for at least 20 years. However, the meaning attached to the word "*behavior*" has changed significantly between the earliest works dedicated to the problem, dating back to the early 1990s, and the latest approaches proposed recently. In the earliest works, "behavior" usually defined simple actions that can be performed by a person and analyzed or synthesized by a computer, e.g., talking on the phone, taking written notes, uttering words, etc. In the latest works, "behavior" accounts for social, affective, and, more generally, psychological aspects of human actions.

This survey has focused on the later approaches to analysis and synthesis of human behavior and, in particular, on the social meaning attached to behavioral cues such as gestures, postures, vocalizations, facial expressions, etc. This research domain, coined Social Signal Processing, aimed at bringing social intelligence in computers, has been surveyed in this paper. To the best of our knowledge, this is the first such paper covering the three fundamental problems of SSP: modeling, analysis, and synthesis of nonverbal behavior in social interactions.

In all the above-mentioned subfields of SSP, the state of the art is in a pioneering stage but constantly evolving and maturing thanks to a vibrant community. However, a number of challenges still need to be addressed before bringing SSP to full maturity, including the actual correspondence (at least in probabilistic terms) between observable behavioral cues and social phenomena, the limited availability of data based on realistic settings and scenarios, the fusion of multiple modalities corresponding to phenomena taking place at different time scales (e.g., vocalizations and facial expressions), the need for real-time systems for testing socially oriented Human-Computer Interaction approaches, etc.

Given the potential outcome in terms of new applications and substantial improvement of existing ones, the community is doing significant efforts toward a solution, at least partial, of the above problems. Evident signs of interest are the increasing number of individual researchers and groups that include SSP among their interests, the growing number of scientific gatherings (workshops, special sessions, etc.) dedicated to human behavior, and large-scale international collaborations such as *SSPNet*[2] that aim at providing the scientific community with basic resources such as annotated data, tools, and extensive bibliographies.

2. Social Signal Processing Network, www.sspnet.eu.

ACKNOWLEDGMENTS

The research that has led to this work has been supported in part by the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

REFERENCES

[1] G. Rizzolatti and L. Craighero, "The Mirror-Neuron System," *Ann. Rev. Neuroscience,* vol. 27, pp. 169-192, 2004.
[2] M. Iacoboni, *Mirroring People: The Science of Empathy and How We Connect with Others.* Picador, 2009.
[3] C. Frith and U. Frith, "Social Cognition in Humans," *Current Biology,* vol. 17, no. 16, pp. 724-732, 2007.
[4] J. Pickles, *An Introduction to the Physiology of Hearing.* Academic Press, 1982.
[5] B. Waller, J. Cray, and A. Burrows, "Selection for Universal Facial Emotion," *Emotion,* vol. 8, no. 3, pp. 435-439, 2008.
[6] Z. Kunda, *Social Cognition.* MIT Press, 1999.
[7] V. Richmond and J. McCroskey, *Nonverbal Behaviors in Interpersonal Relations.* Allyn and Bacon, 1995.
[8] M. Pantic, A. Nijholt, A. Pentland, and T. Huang, "Human-Centred Intelligent Human-Computer Interaction (HCI^2): How Far Are We from Attaining It?" *Int'l J. Autonomous and Adaptive Comm. Systems,* vol. 1, no. 2, pp. 168-187, 2008.
[9] J. Crowley, "Social Perception," *ACM Queue,* vol. 4, no. 6, pp. 34-43, 2006.
[10] T. Bickmore and J. Cassell, "Social Dialogue with Embodied Conversational Agents," *Advances in Natural, Multimodal, Dialogue Systems,* J. van Kuppevelt, L. Dybkjaer, and N. Bernsen, eds., pp. 23-54, Kluwer, 2005.
[11] F. Wang, K. Carley, D. Zeng, and W. Mao, "Social Computing: From Social Informatics to Social Intelligence," *IEEE Intelligent Systems,* vol. 22, no. 2, pp. 79-83, Mar. 2007.
[12] A. Pentland, "Socially Aware Computation and Communication," *Computer,* vol. 38, no. 3, pp. 33-40, Mar. 2005.
[13] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an Emerging Domain," *Image and Vision Computing J.,* vol. 27, no. 12, pp. 1743-1759, 2009.
[14] K. Albrecht, *Social Intelligence: The New Science of Success.* John Wiley & Sons Ltd., 2005.
[15] I. Poggi and F. D'Errico, "Cognitive Modelling of Human Social Signals," *Proc. IEEE Workshop Social Signal Processing,* pp. 21-26, 2010.
[16] P. Brunet, H. Donnan, G. McKeown, E. Douglas-Cowie, and R. Cowie, "Social Signal Processing: What are the Relevant Variables? and in What Ways Do They Relate?" *Proc. IEEE Workshop Social Signal Processing,* pp. 1-6, 2009.
[17] I. Poggi, *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication.* Weidler Buchverlag, 2007.
[18] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship.* MIT Press, 2005.
[19] N. Ambady, F. Bernieri, and J. Richeson, "Towards a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of Behavior," *Proc. Advances in Experimental Social Psychology,* pp. 201-272, 2000.
[20] U. Eco, *Trattato di Semiotica Generale.* Bompiani, 1975.
[21] D. Efron, *Gesture and Environment.* King's Crown Press, 1941.
[22] P. Ekman and W. Friesen, "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding," *Semiotica,* vol. 1, no. 1, pp. 49-98, 1969.
[23] R. Birdwhistell, *Introduction to Kinesics, An Annotation System for Analysis of Body Motion and Gesture.* Univ. of Louisville, 1952.
[24] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS): Manual.* A Human Face, 2002.
[25] M. Argyle and M. Cook, *Gaze and Mutual Gaze.* Cambridge Univ. Press, 1976.
[26] W. Condon and W. Ogston, "A Segmentation of Behavior," *J. Psychiatric Research,* vol. 5, no. 3, pp. 221-235, 1967.
[27] E. Klima and U. Bellugi, *The Signs of Language.* Harvard Univ. Press, 1979.
[28] L. Cerrato, "Communicative Feedback Phenomena across Languages and Modalities," PhD dissertation, KTH Stockolm, 2007.
[29] B. Hartmann, M. Mancini, and C. Pelachaud, "Implementing Expressive Gesture Synthesis for Embodied Conversational Agents," *Proc. Seventh Int'l Gesture Workshop,* pp. 188-199, 2006.
[30] G. Kreidlin, "The Dictionary of Russian Gestures," *Semantics and Pragmatics of Everyday Gestures.* C. Mueller and R. Posner, eds. Weidler, 2004.
[31] A. Kendon, "On Gesture: Its Complementary Relationship with Speech," *Nonverbal Behavior and Comm.,* pp. 65-97, Erlbaum, 1997.
[32] D. McNeill, *Hand and Mind.* Univ. of Chicago Press, 1992.
[33] *Pointing,* S. Kita, ed. Erlbaum, 2003.
[34] R.M. Krauss, Y. Chen, and P. Chawla, "Nonverbal Behavior and Nonverbal Communication: What Do Conversational Hand Gestures Tell Us?" *Advances in Experimental Social Psychology,* vol. 28, pp. 389-450, 1996.
[35] G. Merola, "The Effects of the Gesture Viewpoint on the Students' Memory of Words and Stories," *Proc. Gesture Workshop,* pp. 272-281, 2007.
[36] W. Levelt, *Speaking from Intention to Articulation.* MIT Press, 1989.
[37] H. McGurk and J. McDonalds, "Hearing Lips and Seeing Voices," *Nature,* vol. 264, pp. 746-748, 1976.
[38] M. Meredith, "On the Neural Basis for Multisensory Convergence: A Brief Overview," *Cognitive Brain Research,* vol. 14, no. 1, pp. 31-40, 2002.
[39] S. Campanella and P. Belin, "Integrating Face and Voice in Person Perception," *Trends in Cognitive Sciences,* vol. 11, no. 12, pp. 535-543, 2007.
[40] R. Campbell, "The Processing of Audio-Visual Speech: Empirical and Neural Bases," *Philosophical Trans. Royal Soc. London—B Biological Sciences,* vol. 363, no. 1493, pp. 1001-1010, 2007.
[41] R.E.A. Dolan, "Crossmodal Binding of Fear in Voice and Face," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 10006-10010, 2001.
[42] A. Oppenheim and R. Schafer, *Digital Signal Processing.* Prentice-Hall, 1975.
[43] C. Shannon and R. Weaver, *The Mathematical Theory of Information.* Univ. of Illinois Press, 1949.
[44] O. Hasson, "Cheating Signals," *J. Theoretical Biology,* vol. 167, no. 3, pp. 223-238, 1994.
[45] K. Scherer, "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Comm.,* vol. 40, nos. 1/2, pp. 227-256, 2003.
[46] R. Conte and C. Castelfranchi, *Cognitive and Social Action.* Univ. College London, 1995.
[47] M. Wertheimer, "Laws of Organization in Perceptual Forms," *A Source Book of Gestalt Psychology,* W. Ellis, ed., pp. 71-88, Routledge & Kegan Paul, 1938.
[48] A. Pentland, *Honest Signals: How They Shape Our World.* MIT Press, 2008.
[49] E. Ahlsén, J. Lund, and J. Sundqvist, "Multimodality in Own Communication Management," *Proc. Second Nordic Conf. Multimodal Comm.,* pp. 43-62, 2005.
[50] C. Bazzanella, *Le Facce del Parlare.* La Nuova Italia, 1994.
[51] J.K. Burgoon and N.E. Dunbar, "Interpersonal Dominance as a Situationally, Interactionally, and Relationally Contingent Social Skill," *Comm. Monographs,* vol. 67, no. 1, pp. 96-121, 2000.
[52] D. Heylen, "Challenges Ahead: Head Movements and Other Social Acts in Conversations," *Proc. Joint Symp. Virtual Social Agents,* pp. 45-52, 2005.
[53] M. Schröder, D. Heylen, and I. Poggi, "Perception of Non-Verbal Emotional Listener Feedback," *Proc. Speech Prosody,* 2006.
[54] N. Chovil, "Discourse-Oriented Facial Displays in Conversation," *Research on Language and Social Interaction,* vol. 25, pp. 163-194, 1992.
[55] P. Garotti, "Disprezzo," *Introduzione alla Psicologia delle Emozioni,* V. D'Urso and R. Trentin, eds., Laterza, 1998.
[56] M. Argyle, *Bodily Communication.* Methuen, 1988.
[57] A. Valitutti, O. Stock, and C. Strapparava, "GRAPHLAUGH: A Tool for the Interactive Generation of Humorous Puns," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 592-593, 2009.
[58] I. Poggi, F. Cavicchio, and E. Magno Caldognetto, "Irony in a Judicial Debate, Analyzing the Subtleties of Irony while Testing the Subtleties of an Annotation Scheme," *J. Language Resources and Evaluation,* vol. 41, nos. 3/4, pp. 215-232, 2008.
[59] I. Poggi and F. D'Errico, "Social Signals and the Action—Cognition Loop, the Case of Overhelp and Evaluation," *Proc. IEEE Conf. Affective Computing and Intelligent Interaction,* pp. 106-113, 2009.

[60] B.M. De Paulo, "Nonverbal Behavior and Self-Presentation," *Psychological Bull.,* vol. 111, no. 2, pp. 203-243, 1992.
[61] I. Poggi and L. Vincze, "Persuasive Gaze in Political Discourse," *Proc. Symp. Persuasive Agents,* 2008.
[62] J.K. Burgoon, T. Birk, and M. Pfau, "Nonverbal Behaviors, Persuasion, and Credibility," *Human Comm. Research,* vol. 17, no. 1, pp. 140-169, 1990.
[63] J. Atkinson, "Refusing Invited Applause: Preliminary Observations from a Case Study of Charismatic Oratory," *Handbook of Discourse Analysis,* T. van Dijk, ed., vol. III, pp. 161-181, Academic Press, 1985.
[64] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting Agreement and Disagreement: A Survey of Nonverbal Audiovisual Cues and Tools," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* vol. II, pp. 121-129, 2009.
[65] M. Lewis, "Self-Conscious Emotions: Embarrassment, Pride, Shame, and Guilt," *Handbook of Emotions,* M. Lewis and J. Haviland-Jones, eds., pp. 623-636, Guilford Press, 2000.
[66] *Itinerari del Rancore,* R. Rizzi ed. Bollati Boringhieri, 2007.
[67] I. Poggi and V. Zuccaro, "Admiration," *Proc. AFFINE Workshop,* 2008.
[68] D. Keltner, "Signs of Appeasement: Evidence for the Distinct Displays of Embarrassment, Amusement, and Shame," *J. Personality and Social Psychology,* vol. 68, no. 3, pp. 441-454, 1995.
[69] M.G. Frank, P. Ekman, and W.V. Friesen, "Behavioral Markers and Recognizability of the Smile of Enjoyment," *J. Personality and Social Psychology,* vol. 64, no. 1, pp. 83-93, 1993.
[70] A. Fridlund and A. Gilbert, "Emotions and Facial Expressions," *Science,* vol. 230, pp. 607-608, 1985.
[71] D. Byrne, *The Attraction Paradigm.* Academic Press, 1971.
[72] E. Berscheid and H. Reiss, "Attraction and Close Relationships," *Handbook of Social Psychology,* D. Gilbert, S. Fiske, and G. Lindzey, eds., pp. 193-281, McGraw Hill, 1997.
[73] C. Castelfranchi, "Social Power: A Missed Point in DAI, MA and HCI," *Decentralized AI,* Y. Demazeau and J. Mueller, eds., pp. 49-62, Elsevier, 1990.
[74] R. Conte and M. Paolucci, *Reputation in Artificial Societies. Social Beliefs for Social Order.* Kluwer, 2002.
[75] M. Halliday, *Il Linguaggio Come Semiotica Sociale.* Zanichelli, 1983.
[76] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling Human Interaction in Meetings," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 748-751, 2003.
[77] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, and R. Stiefelhagen, "SMaRT: The Smart Meeting Room Task at ISL," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 752-755, 2003.
[78] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Signals," *J. Personal and Ubiquitous Computing,* vol. 10, no. 4, pp. 255-268, 2006.
[79] R. Murray-Smith, "Empowering People Rather Than Connecting Them," *Int'l J. Mobile Human Computer Interaction,* vol. 3, 2009.
[80] M. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 1, pp. 34-58, Jan. 2002.
[81] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, no. 5, pp. 1557-1565, Sept. 2006.
[82] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational Studies of Human Motion Part 1: Tracking and Motion Synthesis," *Foundations and Trends in Computer Graphics and Vision,* vol. 1, no. 2, pp. 77-254, 2006.
[83] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, Jan. 2009.
[84] D. Crystal, *Prosodic Systems and Intonation in English.* Cambridge Univ. Press, 1969.
[85] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Rev.,* vol. 37, no. 3, pp. 311-324, May 2007.
[86] D. Gatica-Perez, "Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review," *Image and Vision Computing,* vol. 27, no. 12, pp. 1775-1787, 2009.
[87] H. Tischler, *Introduction to Sociology.* Harcourt Brace College Publishers, 1990.
[88] A. Vinciarelli, "Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling," *IEEE Trans. Multimedia,* vol. 9, no. 9, pp. 1215-1226, Oct. 2007.
[89] H. Salamin, S. Favre, and A. Vinciarelli, "Automatic Role Recognition in Multiparty Recordings: Using Social Affiliation Networks for Feature Extraction," *IEEE Trans. Multimedia,* vol. 11, no. 7, pp. 1373-1380, Nov. 2009.
[90] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The Rules Behind the Roles: Identifying Speaker Roles in Radio Broadcasts," *Proc. 17th Nat'l Conf. Artificial Intelligence,* pp. 679-684, 2000.
[91] Y. Liu, "Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech," *Proc. Human Language Technology Conf. NAACL, Companion Volume: Short Papers,* pp. 81-84, June 2006.
[92] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli, "Role Recognition for Meeting Participants: An Approach Based on Lexical Information and Social Network Analysis," *Proc. ACM Int'l Conf. Multimedia,* pp. 693-696, 2008.
[93] S. Favre, A. Dielmann, and A. Vinciarelli, "Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models," *Proc. ACM Int'l Conf. Multimedia,* pp. 585-588, 2009.
[94] J. Quinlan, *C4. 5: Programs for Machine Learning.* Morgan Kaufmann, 1993.
[95] S. Banerjee and A. Rudnicky, "Using Simple Speech Based Features to Detect the State of a Meeting and the Roles of the Meeting Participants," *Proc. Int'l Conf. Spoken Language Processing,* pp. 221-231, 2004.
[96] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation," *Proc. Ninth ISCA/ACL SIGdial Workshop Discourse and Dialogue,* pp. 148-155, June 2008.
[97] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic Detection of Group Functional Roles in Face to Face Interactions," *Proc. Int'l Conf. Mutlimodal Interfaces,* pp. 47-54, 2006.
[98] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro, "Using the Influence Model to Recognize Functional Roles in Meetings," *Proc. Ninth Int'l Conf. Multimodal Interfaces,* pp. 271-278, Nov. 2007.
[99] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int'l J. Synthetic Emotion,* vol. 1, no. 1, pp. 68-99, 2010.
[100] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* vol. 2, nos. 1/2, pp. 1-135, 2008.
[101] J. Levine and R. Moreland, "Small Groups," *Handbook of Social Psychology,* D. Gilbert and G. Lindzey, eds., vol. 2, pp. 415-469, Oxford Univ. Press, 1998.
[102] K. Otsuka, Y. Takemae, and J. Yamato, "A Probabilistic Inference of Multiparty-Conversation Structure Based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances," *Proc. ACM Int'l Conf. Multimodal Interfaces,* pp. 191-198, 2005.
[103] R. Rienks, D. Zhang, and D. Gatica-Perez, "Detection and Application of Influence Rankings in Small Group Meetings," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 257-264, 2006.
[104] R. Rienks and D. Heylen, "Dominance Detection in Meetings Using Easily Obtainable Features," *Proc. Machine Learning for Multimodal Interaction,* pp. 76-86, 2006.
[105] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling Dominance in Group Conversations from Non-Verbal Activity Cues," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 17, no. 3, pp. 501-513, Mar. 2009.
[106] D. Funder, "Personality," *Ann. Rev. Psychology,* vol. 52, pp. 197-221, 2001.
[107] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal Support to Group Dynamics," *Personal and Ubiquitous Computing,* vol. 12, no. 3, pp. 181-195, 2008.
[108] D. Olguin-Olguin, P. Gloor, and A. Pentland, "Capturing Individual and Group Behavior with Wearable Sensors," *Proc. AAAI Spring Symp. Human Behavior Modeling,* 2009.
[109] G. Mohammadi, M. Mortillaro, and A. Vinciarelli, "The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions," *Proc. Int'l Workshop Social Signal Processing,* pp. 17-20, 2010.

[110] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artificial Intelligence Research,* vol. 30, pp. 457-500, 2007.
[111] *The Five-Factor Model of Personality,* J. Wiggins, ed. Guilford, 1996.
[112] J. Curhan and A. Pentland, "Thin Slices of Negotiation: Predicting Outcomes from Conversational Dynamics within the First 5 Minutes," *J. Applied Psychology,* vol. 92, no. 3, pp. 802-811, 2007.
[113] J. Burgoon, L. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns.* Cambridge Univ. Press, 1995.
[114] T. Chartrand and J. Bargh, "The Chameleon Effect: The Perception-Behavior Link and Social Interaction," *J. Personality and Social Psychology,* vol. 76, no. 6, pp. 893-910, 1999.
[115] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand, "The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry," *J. Nonverbal Behavior,* vol. 27, no. 3, pp. 145-162, 2003.
[116] L. Morency, C. Sidner, C. Lee, and T. Darrell, "Head Gestures for Perceptual Interfaces: The Role of Context in Improving Recognition," *Artificial Intelligence,* vol. 171, nos. 8/9, pp. 568-585, 2007.
[117] R. Murray-Smith, A. Ramsay, S. Garrod, M. Jackson, and B. Musizza, "Gait Alignment in Mobile Phone Conversations," *Proc. Int'l Conf. Human-Computer Interaction with Mobile Devices and Services,* pp. 214-221, 2007.
[118] A. Vinciarelli, "Capturing Order in Social Interactions," *IEEE Signal Processing Magazine,* vol. 26, no. 5, pp. 133-137, Sept. 2009.
[119] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data," *Proc. North Am. Chapter of the Assoc. for Computational Linguistics Human Language Technology,* 2003.
[120] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies," *Proc. Meeting Assoc. for Computational Linguistics,* pp. 669-676, 2004.
[121] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "A Multimodal Annotated Corpus of Consensus Decision Making Meetings," *The J. Language Resources and Evaluation,* vol. 41, nos. 3/4, pp. 409-429, 2008.
[122] D. Jayagopi, B. Raducanu, and D. Gatica-Perez, "Characterizing Conversational Group Dynamics Using Nonverbal Behaviour," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 370-373, 2009.
[123] M. Cristani, A. Pesarin, C. Drioli, A. Perina, A. Tavano, and V. Murino, "Auditory Dialog Analysis and Understanding by Generative Modelling of Interactional Dynamics," *Proc. Int'l Workshop Computer Vision and Pattern Recognition for Human Behavior,* pp. 103-109, 2009.
[124] N. Jovanovic, R. op den Akker, and A. Nijholt, "A Corpus for Studying Addressing Behaviour in Multi-Party Dialogues," *Language Resources and Evaluation,* vol. 40, no. 1, pp. 5-23, 2006.
[125] R. Bakeman and J. Gottman, *Observing Interaction: An Introduction to Sequential Analysis.* Cambridge Univ. Press, 1986.
[126] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. I, pp. 364-367, 2003.
[127] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic Analysis of Multimodal Group Actions in Meetings," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 305-317, Mar. 2005.
[128] J. Carletta et al., "The AMI Meeting Corpus: A Pre-Announcement," *Proc. Second Int'l Conf. Machine Learning for Multimodal Interaction,* pp. 28-39, 2005.
[129] L. Chen, R. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, D. Tuttle, and T. Huang, "VACE Multimodal Meeting Corpus," *Proc. Second Int'l Conf. Machine Learning for Multimodal Interaction,* pp. 40-51, 2006.
[130] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps, "A Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow," *Proc. Conf. Language and Resources Evaluation,* pp. 391-394, 2006.
[131] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A Database of Political Debates for Analysis of Social Interactions," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* vol. 2, pp. 96-99, 2009.
[132] S. Burger, V. MacLaren, and H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style," *Proc. Int'l Conf. Spoken Language Processing,* pp. 301-304, 2002.
[133] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi, "The NIST Meeting Room Pilot Corpus," *Proc. Language Resource and Evaluation Conf.,* 2004.
[134] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection," *Proc. Workshop Tagging, Mining and Retrieval of Human Related Activity Information,* pp. 9-14, 2007.
[135] A. Vinciarelli and M. Pantic, "Techware: www.sspnet.eu, a Web Portal for Social Signal Processing," *IEEE Signal Processing Magazine,* vol. 27, no. 4, pp. 142 -144, July 2010.
[136] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction.* Harcourt Brace College Publishers, 1972.
[137] A. Scheflen, "The Significance of Posture in Communication Systems," *Psychiatry,* vol. 27, pp. 316-331, 1964.
[138] E. Hall, *The Silent Language.* Doubleday, 1959.
[139] H. Triandis, *Culture and Social Behavior.* McGraw-Hill, 1994.
[140] *Nonverbal Communication: Where Nature Meets Culture,* U. Segerstrale and P. Molnar, eds. Lawrence Erlbaum Assoc., 1997.
[141] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human Computing and Machine Understanding of Human Behavior: A Survey," *Proc. Eighth Int'l Conf. Multimodal Interfaces,* vol. 4451, pp. 47-71, 2007.
[142] J. Russell, J. Bachorowski, and J. Fernandez-Dols, "Facial and Vocal Expressions of Emotion," *Ann. Rev. Psychology,* vol. 54, no. 1, pp. 329-349, 2003.
[143] J. Cassell, T.W. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H.H. Vilhjalmsson, and H. Yan, "Embodiment in Conversational Interfaces: Rea," *Proc. SIGCHI Conf. Human Factors in Computing Systems,* pp. 520-527, 1999.
[144] E. Schegloff, "Analyzing Single Episodes of Interaction: An Exercise in Conversation Analysis," *Social Psychology Quarterly,* vol. 50, no. 2, pp. 101-114, 1987.
[145] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents," *Proc. 21st Ann. Conf. Computer Graphics Interactive Techniques,* pp. 413-420, 1994.
[146] K. Thórisson, "Natural Turn-Taking Needs No Manual," *Multimodality in Language and Speech Systems,* I.K.B. Granström and D. House, ed., pp. 173-207, Kluwer Academic Publishers, 2002.
[147] J. Bonaiuto and K.R. Thórisson, "Towards a Neurocognitive Model of Realtime Turntaking in Face-to-Face Dialogue," *Embodied Comm. in Humans And Machines,* G.K.I. Wachsmuth and M. Lenzen, ed., pp. 451-484, Oxford Univ. Press, 2008.
[148] K. Prepin and A. Revel, "Human-Machine Interaction as a Model of Machine-Machine Interaction: How to Make Machines Interact as Humans Do," *Advanced Robotics,* vol. 21, no. 15, pp. 1709-1723, 2007.
[149] R.M. Maatman, J. Gratch, and S. Marsella, "Natural Behavior of a Listening Agent," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 25-36, 2005.
[150] L. Morency, I. de Kok, and J. Gratch, "Predicting Listener Backchannels: A Probabilistic Multimodal Approach," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 176-190, 2008.
[151] S. Kopp, T. Stocksmeier, and D. Gibbon, "Incremental Multimodal Feedback for Conversational Agents," *Proc. Eighth Int'l Conf. Intelligent Virtual Agents,* pp. 139-146, 2007.
[152] J. Urbain, S. Dupont, T. Dutoit, R. Niewiadomski, and C. Pelachaud, "Towards a Virtual Agent Using Similarity-Based Laughter Production," *Proc. Interdisciplinary Workshop Laughter and Other Interactional Vocalisations in Speech,* 2009.
[153] S. Sundaram and S. Narayanan, "Automatic Acoustic Synthesis of Human-Like Laughter," *J. Acoustical Soc. Am.,* vol. 121, no. 1, pp. 527-535, 2007.
[154] S. Pammi and M. Schröder, "Annotating Meaning of Listener Vocalizations for Speech Synthesis," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 453-458, 2009.

[155] J. Trouvain and M. Schröder, "How (Not) to Add Laughter to Synthetic Speech," *Proc. Workshop Affective Dialogue Systems,* pp. 229-232, 2004.
[156] M. Schröder, "Experimental Study of Affect Bursts," *Speech Comm.* Special issue speech and emotion, vol. 40, nos. 1/2, pp. 99-116, 2003.
[157] K.R. Scherer, "Affect Bursts," *Emotions: Essays on Emotion Theory,* S.H.M. van Goozen, N.E. van de Poll, and J.A. Sergeant, eds., pp. 161-193, Lawrence Erlbaum, 1994.
[158] P. Ekman, "Biological and Cultural Contributions to Body and Facial Movement," *Anthropology of the Body,* J. Blacking, ed., pp. 39-84, Academic Press, 1977.
[159] A. Kendon, *Conducting Interaction: Pattern of Behavior in Focused Encounter.* Cambridge Univ. Press, 1990.
[160] A. Scheflen, *Body Language and Social Order.* Prentice-Hall, Inc., 1973.
[161] C. Pedica and H.H. Vilhjálmsson, "Spontaneous Avatar Behavior for Human Territoriality," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 344-357, 2009.
[162] D. Jan and D.R. Traum, "Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversations," *Proc. Sixth Int'l Joint Conf. Autonomous Agents and Multiagent Systems,* 2007.
[163] D. Helbing and P. Molnár, "Social Force Model for Pedestrian Dynamics," *Physical Rev. E,* vol. 51, no. 5, pp. 4282-4287, 1995.
[164] M. Schröder, "Expressing Degree of Activation in Synthetic Speech," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1128-1136, July 2006.
[165] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W.J. Barry, "Modelling Personality Features by Changing Prosody in Synthetic Speech," *Proc. Speech Prosody,* 2006.
[166] F. Burkhardt and W.F. Sendlmeier, "Verification of Acoustical Correlates of Emotional Speech Using Formant Synthesis," *Proc. ISCA Workshop Speech and Emotion,* pp. 151-156, 2000.
[167] M. Walker, J. Cahn, and S. Whittaker, "Linguistic Style Improvisation for Lifelike Computer Characters," *AAAI Workshop Entertainment and AI/A-Life,* aAAI Technical Report WS-96-03, 1996.
[168] S. Gupta, M.A. Walker, and D.M. Romano, "Generating Politeness in Task Based Interaction: An Evaluation of the Effect of Linguistic form and Culture," *Proc. 11th European Workshop Natural Language Generation,* pp. 57-64, 2007.
[169] E. André, M. Rehm, W. Minker, and D. Buhler, "Endowing Spoken Language Dialogue Systems with Emotional Intelligence," *Proc. Affective Dialogue Systems,* pp. 178-187, 2004.
[170] L. Johnson, P. Rizzo, W. Bosma, M. Ghijsen, and H. van Welbergen, "Generating Socially Appropriate Tutorial Dialog," *Proc. Affective Dialogue Systems,* pp. 254-264, 2004.
[171] K. Porayska-Pomsta and C. Mellish, "Modelling Politeness in Natural Language Generation," *Proc. Int'l Conf. Natural Language Generation,* pp. 141-150, 2004.
[172] P. Brown and S.C. Levinson, *Politeness—Some Universals in Language Usage.* Cambridge Univ. Press, 1987.
[173] M. de Jong, M. Theune, and D. Hofs, "Politeness and Alignment in Dialogues with a Virtual Guide," *Proc. Seventh Int'l Conf. Autonomous Agents and Multiagent Systems,* pp. 207-214, 2008.
[174] M. Rehm and E. André, "Informing the Design of Embodied Conversational Agents by Analysing Multimodal Politeness Behaviors in Human-Human Communication," *Proc. Workshop Conversational Informatics for Supporting Social Intelligence and Interaction,* 2005.
[175] R. Niewiadomski and C. Pelachaud, "Model of Facial Expressions Management for an Embodied Conversational Agent," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 12-23.
[176] H. Prendinger and M. Ishizuka, "Social Role Awareness in Animated Agents," *Proc. Int'l Conf. Autonomous Agents,* pp. 270-277, 2001.
[177] M. Schröder, "Expressive Speech Synthesis: Past, Present, and Possible Futures," *Affective Information Processing,* J. Tao and T. Tan, eds., pp. 111-126, Springer, 2009.
[178] A. Hunt and A.W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," *Proc. IEEE Int'l Conf. Audio, Speech, and Signal Processing,* vol. 1, pp. 373-376, 1996.
[179] J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M.A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1099-1108, July 2006.
[180] W.L. Johnson, S.S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited Domain Synthesis of Expressive Military Speech for Animated Characters," *Proc. IEEE Workshop Speech Synthesis,* pp. 163-166, 2002.
[181] P. Gebhard, M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, and O. Türk, "IDEAS4Games: Building Expressive Virtual Characters for Computer Games," *Proc. Intelligent Virtual Agents,* pp. 426-440, 2008.
[182] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards Emotional Speech Synthesis: A Rule Based Approach," *Proc. ISCA Speech Synthesis Workshop,* pp. 219-220, 2004.
[183] O. Türk and M. Schröder, "A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis," *Proc. Interspeech,* 2008.
[184] M. Schröder, "Interpolating Expressions in Unit Selection," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 718-720, 2007.
[185] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrumc Pitch and Duration in HMM-Based Speech Synthesis," *Proc. Eurospeech,* 1999.
[186] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis," *Proc. Eurospeech,* pp. 2461-2464, 2003.
[187] T. Nose, J. Yamagishi, and T. Kobayashi, "A Style Control Technique for Speech Synthesis Using Multiple Regression HSMM," *Proc. Interspeech,* pp. 1324-1327, 2006.
[188] K. Miyanaga, T. Masuko, and T. Kobayashi, "A Style Control Technique for HMM-Based Speech Synthesis," *Proc. Eighth Int'l Conf. Spoken Language Processing,* pp. 1437-1440, 2004.
[189] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model Adaptation Approach to Speech Synthesis with Diverse Voices and Styles," *Proc. Int'l Conf. Audio, Speech and Signal Processing,* vol. IV, pp. 1233-1236, 2007.
[190] R. Fernandez and B. Ramabhadran, "Automatic Exploration of Corpus-Specific Properties for Expressive Text-to-Speech: A Case Study in Emphasis," *Proc. Sixth ISCA Workshop Speech Synthesis,* pp. 34-39, 2007.
[191] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling Prominence and Emphasis Improves Unit-Selection Synthesis," *Proc. Interspeech,* pp. 1282-1285, 2007.
[192] Ruth, http://www.cs.rutgers.edu/village/ruth/, 2012.
[193] Cadia, http://cadia.ru.is/projects/bmlr/, 2012.
[194] Greta, http://www.tsi.enst.fr/~pelachau/greta/, 2012.
[195] SmartBody, http://www.smartbody-anim.org/, 2012.
[196] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N.E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A.N. Marshall, C. Pelachaud, Z. Ruttkay, K.R. Thórisson, H. van Welbergen, and R. van der Werf, "The Behavior Markup Language: Recent Developments and Challenges," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 99-111, Sept. 2007.
[197] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson, "Why Conversational Agents Do What They Do? Functional Representations for Generating Conversational Agent Behavior," *Proc. Seventh Int'l Conf. Autonomous Agents and Multiagent Systems,* 2008.
[198] BML, http://wiki.mindmakers.org/projects:bml:main, 2012.
[199] M. Schröder, "The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-oriented Systems," *Advances in Human-Computer Interaction,* pp. 319-406, 2010.
[200] Festival, http://www.cstr.ed.ac.uk/projects/festival/, 2012.
[201] OPENMARY, http://mary.dfki.de/, 2012.
[202] EULER, http://tcts.fpms.ac.be/synthesis/euler/, 2012.
[203] M. Gladwell, *Blink: The Power of Thinking without Thinking.* Little Brown & Company, 2005.
[204] S.E. Hyman, "A New Image for Fear and Emotion," *Nature,* vol. 393, pp. 417-418, 1998.
[205] E. Douglas-Cowie, L. Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal Databases of Everyday Emotion: Facing Up To Complexity," *Proc. Interspeech,* pp. 813-816, 2005.
[206] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-Driven Lip Motion Generation with a Trajectory HMM," *Proc. Interspeech,* pp. 2314-2317, 2008.
[207] J. Bates, "The Role of Emotion in Believable Agents," *Comm. ACM,* vol. 37, no. 7, pp. 122-125, 1994.

[208] J.S. Uleman, L.S. Newman, and G.B. Moskowitz, "People as Flexible Interpreters: Evidence and Issues from Spontaneous Trait Inference," *Advances in Experimental Social Psychology,* M.P. Zanna, ed., vol. 28, pp. 211-279, Elsevier, 1996.

[209] J.S. Uleman, S.A. Saribay, and C.M. Gonzalez, "Spontaneous Inferences, Implicit Impressions, and Implicit Theories," *Ann. Rev. Psychology,* vol. 59, pp. 329-360, 2008.

[210] K. Isbister and C. Nass, "Consistency of Personality in Interactive Characters: Verbal Cues, Non-Verbal Cues, and User Characteristics," *Int'l J. Human-Computers Studies,* vol. 53, no. 2, pp. 251-267, 2000.

[211] B. de Gelder and J. Vroomen, "The Perception of Emotions by Ear and by Eye," *Cognition and Emotion,* vol. 14, no. 3, pp. 289-311, 2000.

[212] M. Mori, "The Uncanny Valley," *Energy,* vol. 7, no. 4, pp. 33-35, 1970.

[213] T.B. Moeslund, A. Hilton, and V. Krüger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis," *Computer Vision and Image Understanding,* vol. 104, nos. 2/3, pp. 90-126, 2006.

[214] M.E. Foster, "Comparing Rule-Based and Data-Driven Selection of Facial Displays," *Proc. Workshop Embodied Language Processing,* pp. 1-8, 2007.

[215] M. Buchanan, "The Science of Subtle Signals," *Strategy+Business,* vol. 48, pp. 68-77, 2007.

[216] K. Greene, "10 Emerging Technologies 2008," *MIT Technology Rev.,* Feb. 2008.

[217] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins, "Stuff I've Seen: A System for Personal Information Retrieval and Re-Use," *Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval,* pp. 72-79, 2003.

[218] C. Weng, W. Chu, and J. Wu, "Rolenet: Movie Analysis from the Perspective of Social Networks," *IEEE Trans. Multimedia,* vol. 11, no. 2, pp. 256-271, Feb. 2009.

[219] M. Pantic and A. Vinciarelli, "Implicit Human Centered Tagging," *IEEE Signal Processing Magazine,* vol. 26, no. 6, pp. 173-180, Nov. 2009.

[220] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. Jose, "Integrating Facial Expressions Into User Profiling for the Improvement of a Multimodal Recommender System," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1440-1443, 2009.

[221] I. Arminen and A. Weilenmann, "Mobile Presence and Intimacy—Reshaping Social Actions in Mobile Contextual Configuration," *J. Pragmatics,* vol. 41, no. 10, pp. 1905-1923, 2009.

[222] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones: An Emerging Tool for Social Scientists," *Sociological Methods and Research,* vol. 37, no. 3, pp. 426-454, 2009.

[223] S. Strachan and R. Murray-Smith, "Nonvisual, Distal Tracking of Remote Agents in Geosocial Interaction," *Proc. Symp. Location and Context Awareness,* 2009.

[224] H. Ishii and M. Kobayashi, "ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact," *Proc. SIGCHI Conf. Human Factors in Computing Systems,* pp. 525-532, 1992.

[225] J. Bailenson and N. Yee, "Virtual Interpersonal Touch and Digital Chameleons," *J. Nonverbal Behavior,* vol. 31, no. 4, pp. 225-242, 2007.

[226] N. Ambady, M. Krabbenhoft, and D. Hogan, "The 30-Sec Sale: Using Thin-Slice Judgments to Evaluate Sales Effectiveness," *J. Consumer Psychology,* vol. 16, no. 1, pp. 4-13, 2006.

[227] A. Chattopadhyay, D. Dahl, R. Ritchie, and K. Shahin, "Hearing Voices: The Impact of Announcer Speech Characteristics on Consumer Next Term Response to Broadcast Advertising," *J. Consumer Psychology,* vol. 13, no. 3, pp. 198-204, 2003.

[228] D. Wooten and A. Reed II, "A Conceptual Overview of the Self-Presentational Concerns and Response Tendencies of Focus Group Participants," *J. Consumer Psychology,* vol. 9, no. 3, pp. 141-153, 2000.

[229] W. Breitfuss, H. Prendinger, and M. Ishizuka, "Automatic Generation of Non-Verbal Behavior for Agents in Virtual Worlds: A System for Supporting Multimodal Conversations of Bots and Avatars," *Proc. 3D Int'l Conf. Online Communities and Social Computing,* pp. 153-161, 2009.

[230] A. Sagae, B. Wetzel, A. Valente, and W.L. Johnson, "Culture-Driven Response Strategies for Virtual Human Behavior in Training Systems," *Proc. Speech and Language Technologies in Education,* 2009.

[231] K. Dautenhahn, "Socially Intelligent Robots: Dimensions of Humanrobot Interaction," *Philosophical Trans. Royal Soc. B,* vol. 362, pp. 679-704, 2007.

Alessandro Vinciarelli is a lecturer at the University of Glasgow and a senior researcher in the Idiap Research Institute. His main research interests include Social Signal Processing, the new domain aimed at bringing social intelligence in computers. He is coordinator of the FP7 Network of Excellence SSPNet (www.sspnet. eu) and is, or has been, a principal investigator for several national and international projects. He has authored and coauthored more than 60 publications, including one book and 20 journal papers. He has organized a large number of international workshops, he is the cochair of the IEEE Technical Committee on SSP, and he is an associate editor of the *IEEE Signal Processing Magazine* for the social sciences. He is founder of a knowledge management company (Klewel) recognized with several national and international prizes (www.klewel.com). He is a member of the IEEE.

Maja Pantic is a professor in affective and behavioral computing at Imperial College London, Department of Computing, United Kingdom, and at the University of Twente, Department of Computer Science, The Netherlands. She has published more than 150 articles and received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship and the Roger Needham Award 2011. She currently serves as the editor-in-chief of *Image and Vision Computing Journal* and as an associate editor for both the *IEEE Transactions on Systems, Man, and Cybernetics Part B* and the *IEEE Transactions on Multimedia*. She is a fellow of the IEEE.

Dirk Heylen is an associate professor in the Human Media Interaction Group at the University of Twente, where his research involves modeling conversational and cognitive functions of embodied conversational agents. His work on the analysis and synthesis of nonverbal communication in (multiparty) conversations has been concerned with gaze, and head movements in particular. Besides looking at how to analyze and synthesize communicative behaviors, his research deals with building models of communicative agents. This includes building models of affective and social interaction.

Catherine Pelachaud is the director of Research at CNRS in the LTCI laboratory, Telecom ParisTech. Her research interests include embodied conversational agents, representation language for agent, nonverbal communication, expressive behaviors, and multimodal interfaces. She has been involved in or is still involved in several European projects related to believable embodied conversational agents (IST-MagiCster), emotion (FP5 NoE Humaine, FP6 IP CALLAS, FP7 STREP SEMAINE), and social behaviors (FP7 NoE SSPNet). She has published more than 80 papers, including three edited books. She has coorganized several international workshops and conferences.

Isabella Poggi is a full professor of general psychology and psychology of communication on the Faculty of Education of the University Roma Tre. She has been working since 1976 on a general model of verbal and multimodal communication, and she has published books and papers in teaching Italian as a first language, pragmatics (interjections, deception, persuasion), emotions (shame, guilt, compassion, enthusiasm, empathy, contagion), multimodal communication (gesture, gaze, touch, music, interaction between modalities) by exploiting methods of conceptual analysis, observation, empirical research, and simulation on Embodied Agents. Within the LAOC of Roma Tre (Laboratory of Organization Learning and Communication) she is responsible for research on emotions, communication, and social relations at work (power relations, mobbing).

Francesca D'Errico is a postdoctoral researcher at the University Roma Tre. Her research interests include cognitive modeling of social signals and nonverbal communication. She is an author or coauthor of several articles and book chapters. Furthermore, she is the guest editor of a special issue on social signal processing for *Cognitive Processing*.

Marc Schröder received the maitrise in language science from the University of Grenoble 3 in 1998 and the PhD degree in phonetics from Saarland University in 2003. He is a senior researcher at DFKI and the leader of the DFKI speech group. Since 1998, he has been responsible at DFKI for building up technology and research in TTS. Within the FP6 NoE HUMAINE, he has built up the scientific portal emotion-research.net, which won the Grand Prize for the best IST project website 2006. He is chair of the W3C Emotion Incubator group, coordinator of the FP7 STReP SEMAINE, and project leader of the national funded basic research project PAVOQUE. He is an author of more than 45 scientific publications and a program committee member for many conferences and workshops.