Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain

Alessandro Vinciarelli[1,2], Maja Pantic[3,4], Hervé Bourlard[1,2] and Alex Pentland[5]
[1]IDIAP Research Institute, CP592 - 1920 Martigny (CH)
[2]Ecole Polytechnique Federale de Lausanne - 1015 Lausanne (CH)
[3]Computing, Imperial College London, 180 Queen's Gate - London SW7 2AZ (U.K.)
[4]EEMCS, University of Twente, Drienerlolaan 5 - 7522 NB Enschede (NL)
[5]The MIT Media Laboratory, 20 Ames St. - Cambridge, MA 01239 (USA)
vincia@idiap.ch, m.pantic@imperial.ac.uk, bourlard@idiap.ch,
pentland@media.mit.edu

ABSTRACT

The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence. Social intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for success in life. This paper argues that next-generation computing needs to include the essence of social intelligence – the ability to recognize human social signals and social behaviours like politeness, and disagreement – in order to become more effective and more efficient. Although each one of us understands the importance of social signals in everyday life situations, and in spite of recent advances in machine analysis of relevant behavioural cues like blinks, smiles, crossed arms, laughter, and similar, design and development of automated systems for Social Signal Processing (SSP) are rather difficult. This paper surveys the past efforts in solving these problems by a computer, it summarizes the relevant findings in social psychology, and it proposes a set of recommendations for enabling the development of the next generation of socially-aware computing.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Vision and Scene Understanding; J.4 [**Social and behavioral Sciences**]: Psychology

General Terms

Algorithms

1. INTRODUCTION

The exploration of how human beings react to the world and interact with it and each other remains one of the greatest scientific challenges. Perceiving, learning, and adapting to the world are commonly labelled as intelligent behaviour. But what does it mean being intelligent? Is IQ a good measure of human intelligence and the best predictor of somebody's success in life? There is now a growing research in cognitive sciences, which argues that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called *social intelligence* [2][3][10] and includes the ability to express and recognise social signals and social behaviours like agreement, politeness, and empathy, coupled with the ability to manage them in order to get along well with others while winning their cooperation. Social signals are the expression of one's attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, and vocal outbursts like laughter (see Figure 1). These typically last for a short time (milliseconds, like a blink or gaze shift, to minutes, like a posture), compared to the actual social signals and social behaviours that last longer (seconds, like agreement, to minutes, like politeness, to hours or days, like empathy) and are expressed as temporal patterns of non-verbal behavioural cues. The skills of social intelligence have been argued to be indispensable and perhaps the most important for success in life [2].

When it comes to computers, however, they are socially ignorant [66]. Current computing devices do not account for the fact that human-human communication is always socially situated and that discussions are not just facts but part of a larger social interplay. However, not all computers will need social intelligence and none will need all of the related skills humans have. The current-state-of-the-art categorical computing works well and will always work well for context-independent tasks like making plane reservations and buying and selling stocks. However, this kind of computing is utterly inappropriate for virtual reality applications as well as for interacting with each of the (possibly hundreds) computer systems diffused throughout future smart environments (predicted as the future of computing by several visionaries such as Mark Weiser) and aimed at improving the quality of life by anticipating the users needs. Computer systems and devices capable of sensing agreement, inattention, or dispute, and capable of adapting and responding to these social signals in a polite, unintrusive, or persuasive manner, are likely to be perceived as more natural, efficacious, and trustworthy. For exam-

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

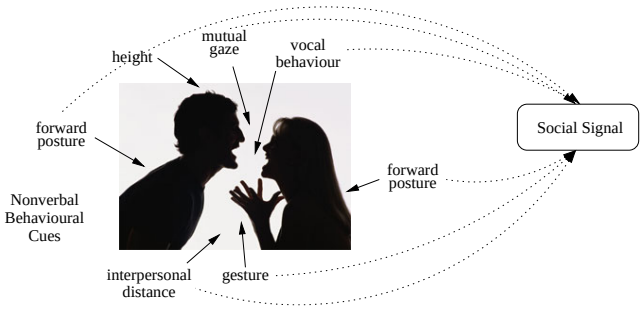

Figure 1: Behavioural cues and social signals. Multiple behavioural cues (vocal behaviour, posture, mutual gaze, interpersonal distance, etc.) combine to produce a social signal (in this case aggressivity or disagreement) that is evident even if the picture shows only the silhouettes of the individuals involved in the interaction.

ple, in education, pupils' social signals inform the teacher of the need to adjust the instructional message. Successful human teachers acknowledge this and work with it; digital conversational embodied agents must begin to do the same by employing tools that can accurately sense and interpret social signals and social context of the pupil, learn successful context-dependent social behaviour, and use a proper socially-adept presentation language (see e.g. [64]) to drive the animation of the agent.

Although the importance of social signals in everyday life situations is evident, and in spite of recent advances in machine analysis and synthesis of relevant behavioural cues like gaze exchange, blinks, smiles, head nods, crossed arms, laughter, and similar, the research efforts in machine analysis and synthesis of human social signals like attention, empathy, politeness, flirting, (dis)agreement, etc., are still tentative and pioneering efforts. The importance of studying social interactions and developing automated assessing of human social behaviour from audiovisual recordings is undisputable. It will result in valuable multimodal tools that could revolutionise basic research in cognitive and social sciences by raising the quality and shortening the time to conduct research that is now lengthy, laborious, and often imprecise. At the same time, and as outlined above, such tools form a large step ahead in realising naturalistic, socially-aware computing and interfaces, built for humans, based on models of human behaviour. Social interaction has commonly been addressed within two different frameworks. One framework comes from cognitive psychology, and focuses on emotion. The key idea is that people perceive others' emotions through stereotyped displays of facial expression, tone of voice, etc. The second framework for understanding social interaction comes from linguistics, and treats social interaction from the viewpoint of dialog understanding. Vocal prosody and gesture are treated as annotations of the basic linguistic information, and used (for instance) to guide attention and signal irony [29].

Social Signal Processing [66][68] is an alternative computational framework, in which speaker attitude or intention is conveyed through the amplitude and frequency of prosodic and gestural activities [65]. This framework is based on the literature of personality and social psychology, and is different from the linguistic framework in that it consists of non-linguistic, largely unconscious, signals about the social situation, and different from the affect framework in that it communicates social relation and not speaker emotion. It is most closely related to the social signaling framework that dominates biology and economics research. It is different in another way as well: it happens over longer time frames than typical linguistic phenomena or emotional displays, treating gestures more like a motion texture than individual actions, and it appears to form a largely independent channel of communication. Social signaling is what you perceive when observing a conversation in an unfamiliar language, and yet find that you can still *see* someone taking charge of a conversation, or establishing a friendly interaction [30].

To the best of our knowledge, this is the first attempt to survey the past work done on SSP. The innovative and multidisciplinary character of the research on SSP is the main reason for this state of affairs. For example, in contrast to the research on human affective behaviour analysis that witnessed tremendous progress in the past decade (for exhaustive surveys in the field see, e.g.,[63]), the research on machine analysis of human social behaviour just started to attract the interest of the research community in computer science. This and the fragmentation of the research over several scientific communities including those in psychology, computer vision, speech and signal processing, make the exercise of surveying the current efforts in machine analysis of human social behaviour difficult.

2. BEHAVIOURAL CUES AND SOCIAL SIGNALS: A TAXONOMY

Is it possible to understand what kind of interactions are having the two individuals portrayed in Figure 1? Are they fighting, laughing, or just having a plain discussion? Are they friends, colleagues, or members of the same family? The picture seems to miss most of the information needed to answer the above questions, but still most of the people watching the image can guess the correct answer: they are hausband and wife and they are fighting. In any case, it is evident to most observers that the two persons have a close relationship and that their affective state is not neutral.

The key elements of such a precise assessment of social interactions, even when limited information is available, are *behavioural cues* and *social signals* (or *social behaviours*). The expression "behavioural cue" is typically used to describe a set of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes), being the main reason for referring to behavioural cues as to *thin slices of behaviour* [3]. Multiple behavioural cues combine to produce social signals (*aggressivity* or *disagreement* in the case of the Figure 1), i.e., an attitude towards others or specific social situations that can last minutes to hours.

Studies performed in the last four decades have shown that social perceptions are shaped mainly by nonverbal behaviour, even if the interactions are typically accompanied by the exchange of verbal messages [3][50]. It is therefore not surprising that SSP focuses on nonverbal communication. Table 1 reports the social signals associated with the behavioural cues that the psychologists consider the most important for conveying social information [44][73].

The cues are grouped into few classes. The first relates to **physical appearance** and includes natural characteristics such as height, body shape, physiognomy, skin and hair color, as well as artificial characteristics such as clothes, ornaments, make up, and other manufacts used to modify/ accentuate the facial/ body aspects. Although common wisdom suggests that the appearance is not important, psychological observations seem to show the contrary. For example, attractiveness elicits desirable social perceptions like high status or good personality even in absence of an objective basis (this phenomenon if referred to as "*what is beautiful is good*" [22]). Tall people are attributed, on average higher social status [30], and the body shapes (round and soft, bony and muscular, or thin and fragile) tend to elicit the attrbution of certain personality traits rather than others [15].

The second class of behavioural cues includes **gestures and postures**. The former are often used consciously, e.g., when waiving hands to greet. However, from an SSP point of view, the most important gestures are those made unconsciously and conveying information about the actual state of people. For example, gestures like self-touching and manipulation of small objects, called *adaptors*, are typically due to boredom or negative attitudes towards others [44]. Postures are typically assumed unconsciously and they are one of the most reliable cues about the rapport between people [73]. Three main criteria define the social meaning of a posture [76]: *inclusion vs. exclusion* (facing in the direction opposite to others shows a negative attitude), *parallel vs. face-to-face* (the choice of face-to-face postures in absence of constraints shows engagement in the interaction), and *congruence vs. non-congruence* (people having satisfying interactions tend to assume the same posture).

Face and eye behaviour are the cues that express social signals with the highest effectiveness. This is evident in psychological experiments where human assessors judge the rapport between people using a single behavioural cue and the results obtained using the facial expressions alone lead to the best accuracy [3]. Facial expressions, typically represented with the *Facial Action Coding System* (FACS) [27], express cognitive states like interest and puzzlement [17], psychological states like suicidal depression [27], social behaviours like accord and rapport [3][17], personality traits like extraversion and temperament [27], and social signals like status, trustworthiness, emblems (i.e., culture-specific interactive signals like wink), regulators (i.e., conversational mediators like nod and gaze exchange), and illustrators (i.e., cues accompanying speech like raised eyebrows) [3].

Vocal nonverbal behaviour includes all spoken cues that surroud the verbal message and influence its actual meaning, namely *voice quality*, *linguistic* and *non-linguistic vocalizations*, *silences*, and *turn-taking patterns*. The *voice quality* corresponds to the prosody and, in perceptual terms, accounts for *how* something is said [35]. It conveys information like emotions [78], and it influences the perception of dominance, extroversion, competence and persuasiveness [77]. *Linguistic vocalizations* include all the non-words that are used as if they were actual words, e.g., "*ehm*","*ah-ah*", "*uhm*", etc. They typically account for embarassment or difficulty with respect to a social interaction [31], but they are also used when someone else speaks (the *back-channel*) to show attention, agreement, wonder or contradiction [82]. The *non-linguistic vocalizations* include nonverbal sounds like laughing, sobbing, crying, whispering, groaning, and similar. These may or may not accompany words, and can be used to reward desirable social behaviour (e.g. through laughter [40]), or to show strong social bonds (e.g. when crying because other people have problems).

	Example Social Signals						
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport
Physical appearance							
height			√	√			
attractiveness		√	√	√	√		√
body shape		√		√			
Gesture and posture							
hand gestures	√				√	√	√
posture	√	√	√	√	√	√	√
walking		√	√	√			
Face and eyes behaviour							
facial expressions	√	√	√	√	√	√	√
gaze behaviour	√	√	√	√	√	√	√
focus of attention	√	√			√	√	√
Vocal behaviour							
prosody	√	√			√		
turn taking			√	√		√	√
vocalizations	√	√		√	√	√	√
silence							√
Space and Environment							
distance		√	√		√		√
seating arrangement				√			√

Table 1: The table shows the behavioural cues associated to some of the most important social behaviours as well as the technologies involved in their automatic detection.

The silence is often interpreted as simple non-speech, but actually plays a major role in vocal behaviour [97]. There are three main kinds of silence [73]: *hesitation silence* (typically due to difficulty and embrassment), *psycholinguistic silence* (typically due to cognitive loads), and *interactive silence* (typically aimed at expressing attitudes like attention or ignoring). The last important aspect of vocal nonverbal behaviour is turn-taking [72]. This includes the regulation of the conversations, and the coordination (or the lack of it) during the speaker transitions. The regulation in conversations includes behaviours (including voice quality and gaze) aimed at maintaining, yielding, denying, or requesting the turn [95]. The second important aspect in turn-taking is the coordination at the speaker transitions [29]. When the interaction is satisfying, the speaker transitions tend to be smooth and no interruptions or long latency times are observed. When the interactions are not positive, interruptions and other behaviours related to aggressivity and dominance appear more frequently [84]. Note, however, that the amount of overlapping speech accounts for up to 10% of the total time even in normal conversations [81].

The last important source of behavioural cues is the use of **space and environment**. Physical distances between individuals often correspond to their social distances. Anthropologists have shown that people tend to split the space

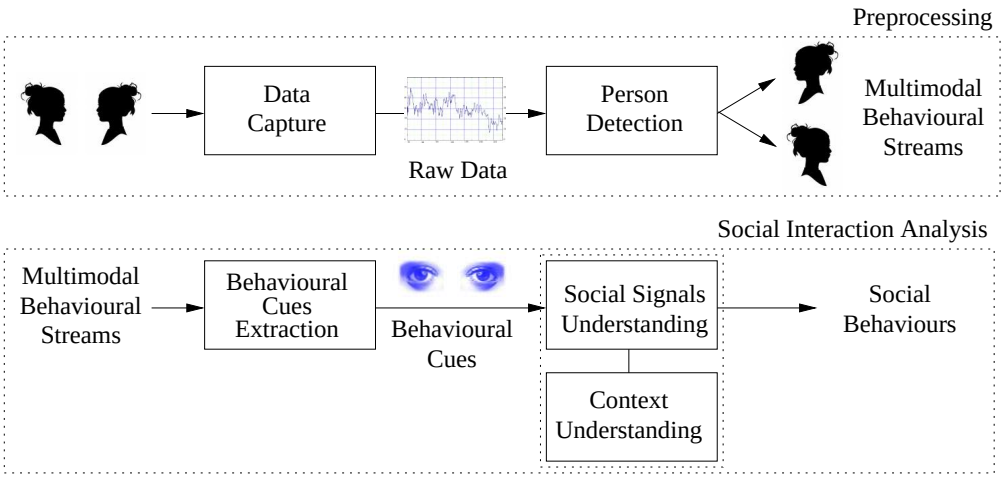

Figure 2: State-of-the-art. The figure shows the general scheme of an SSP approach. The scheme includes two main stages: the *preprocessing*, that takes as input the scene recordings and gives as output the multimodal behavioural streams associated to each person detected in the scene, and the *social interaction analysis*, that maps the multimodal behavioural streams into the social signals.

around them into concetric regions where others are allowed depending on social criteria [34]. The innermost region is called *intimate* and it is open only to the closest family members (typically up to 0.5 m of distance). The second region (between 0.5 and 1.2 m) is called *casual-personal* and it is open to familiar people (e.g., friends and colleagues). The following region (distance between 1.2 and 2.0 m) is called *socio-consultive* and it is used for formal relationships. The rest of the space (beyond 2 m of distance) is called *public* and it is typically beyond the reach of social interactions (in general is used for interaction only in presence of obstacles like long meeting tables or similar).

3. STATE OF THE ART

The problem of machine analysis of human social signals includes two major stages (see Figure 2): *preprocessing* and *social interaction analysis* . The *preprocessing* includes data capture (the recording of the scene with multiple sensors) and detection of the people in the observed scene. In the case of microphones and cameras, the sensors most commonly applied, people detection corresponds to speaker diarization [87], and face [94] or full human figure [51] detection. The *social interaction analysis* includes the extraction of audio and/or visual behavioural cues displayed by people detected in the scene, the interpretation of this information in terms of social signals conveyed by the observed behavioural cues, the sensing of the context in which the scene is recorded, and classification of detected social signals into the target social-behaviour-interpretative categories in a context-sensitive manner.

The preprocessing stage is based on technologies that have been extensively investigated in the recent years and are not specifically oriented to social interactions (e.g., speaker diarization can be performed for other purposes than the analysis of interactions), while the social interaction analysis stage is the actual *core problem* of SSP and it is still largely unexplored, as discussed in the rest of this section.

3.1 Social Signals Detection

This section presents the main approaches applied so far to extract automatically the behavioural cues described in Section 2.

3.1.1 Social Signals from Physical Appearance

To the best of our knowledge, the detection of people appearance has been addressed in relatively few works. These were never aimed at inferring social information and focused rather on biometric and surveillance applications. Several approaches have proposed measures of facial attractiveness based on symmetry and respect of canonical proportions in the geometry of face landmarks (eyes, nose tip, corners of mouth, brows, etc.) [1][25], while others have rather pointed on the adherence to "average" facial models [58]. The modeling of the overall appearance of individuals (color of clothes, skin, hair, etc.) has been investigated for identification purposes in [19].

3.1.2 Social Signals from Gesture and Posture

Gesture recognition is an active research domain, but no attempts have been made, to the best of our knowledge, to interpret gestures in terms of social information, with the exception of few efforts aimed at inferring affective states from gestures (see [62] for more details). The most common approaches for gesture recognition start by detecting the different body parts (arms, legs, trunk, etc.) using features like the orientation of edge histograms, velocity features extracted with stereo cameras, or pixel colors. In the following, they model the temporal dynamics of gestures applying Hidden Markov Models or recurrent neural networks (see [71] for a survey and [9][56] for examples).

Also automatic posture recognition has been addressed in few works, mostly aiming at surveillance [46] (using multiscale morphological method and Kalman motion estimation) and activity recognition [60] (using an eigenspace representation of human silhouettes obtained from Digital Cosine

Figure 3: Basic emotions. Prototypic facial expressions of six basic emotions (disgust, happiness, sadness, anger, fear, and surprise).

Transform coefficients). However, there are few works where the posture is recognized as a social signal, namely to estimate the interest level of children learning to use computers [53], to recognize the affective state of people [20], and the influence of culture on affective postures [43].

3.1.3 Social Signals from Gaze and Face

The face is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression. Personality, attractiveness, age and gender [39] can be also seen from someone's face [3]. Thus the face is a multi-signal sender/receiver capable of tremendous flexibility and specificity. It is therefore not surprising that the experiments (see beginning of Section 2) about the relative weight of the different nonverbal components in shaping social perceptions always show that facial behaviour plays a major role [3][50].

As indicated in [13], most commonly used facial expression descriptors in message judgment approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise; see Fig. 3), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions [39]. In sign judgment approaches [14], a widely used method for manual labeling of facial actions is the Facial Action Coding System (FACS) [26].

FACS associates facial expression changes with actions of the muscles that produce them. It defines 9 different Action Units (AUs) in the upper face, 18 in the lower face, 11 for head position, 9 for eye position, and 14 additional descriptors for miscellaneous actions. AUs are considered to be the smallest visually discernable facial movements. Using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs that produced the expression. As AUs are independent of interpretation, they can be used for any higher order decision making process including recognition of basic emotions (EMFACS; see [26]), cognitive states like interest and puzzlement [17], psychological states like suicidal depression [27] or pain [93], social behaviours like accord and rapport [3][17], personality traits like extraversion and temperament [27], and social signals like status, trustworthiness, emblems (i.e., culture-specific interactive signals like wink), regulators (i.e., conversational mediators like nod and gaze exchange), and illustrators (i.e., cues accompanying speech like raised eyebrows) [3].

Most facial expressions analyzers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger [63][98]. However, several promising prototype systems were reported that can recognize deliberately produced AUs in face images (for overviews, see [61]) and even few attempts towards recognition of spontaneously displayed AUs [48] and towards automatic discrimination between spontaneous and posed facial behaviour such as smiles [89], and pain [47], have been recently reported as well. Although still tentative, few studies have also been recently reported on separating emotional states from non-emotional states and on recognition of non-basic affective states in visual and audiovisual recordings of spontaneous human behaviour [79].However, although messages conveyed by AUs like winks, blinks, frowns, smiles, gaze exchanges, etc., can be interpreted in terms of social signals like turn taking, mirroring, empathy, engagement, etc., no efforts have been reported so far on automatic recognition of social behaviours in recordings of spontaneous facial behaviour. Hence, while the focus of the research in the field started to shift to automatic (non-basic-) emotion and AU recognition in spontaneous facial expressions (produced in a reflex-like manner), efforts towards automatic analysis of human social behaviour from visual and audiovisual recordings of human spontaneous behaviour are still to be made.

3.1.4 Social Signals from Vocal Behaviour

Nonverbal vocal behaviour accounts for roughly 50% of the total time in spontaneous conversations [11], thus it has been extensively investigated in speech processing, but only with the goal of improving speech recognition and synthesis systems (see [35] for an extensive monograph). In other words, no major efforts have been made, to our knowledge, to interpret nonverbal vocal behaviour in social terms.

Section 2 presents the five major components of vocal behaviour, namely voice quality, linguistic and non-linguistic vocalizations, silence and turn-taking patterns. The first corresponds to the prosody and accounts for *how* something is said. The three main prosodic features, called the *Big Three*, are *pitch*, *tempo* and *energy* [16]. The first is the frequency of oscillation of vocal folds during voice emission, the second relates to speaking rate and its variation, and the third is the energy carried by the vocal acoustic waves [35]. The pitch is typically obtained by analyzing the Fourier transform of the speech signal from short intervals (in general 30 ms) [1]. The tempo is measured through the rate of phonetically relevant events like vowels and syllables [70], or through the first spectral moment of the energy [52]. The energy is a property of any digital signal and corresponds to the sum of the square values of the signal samples. In general the energy is extracted from short analysis windows (30 ms like the pitch) [35].

To the best of our knowledge, no efforts have been made to detect non-linguistic vocalizations, with the only exception of laughter [41][88] for its ubiquitous presence in social interactions. The detection is typically performed by classifying vectors of common speech features like *Mel Frequency Cepstral Coefficients* [35] with models like Gaussian Mixture Models and Neural Networks. Recent approaches have shown that the detection performance can be dramatically improved using multimodal approaches based on both audio and visual features [37][69].

[1]Several software packages perform pitch extraction, e.g., Praat [8] and Wavesurfer [83], both publicly available on the web.

On the contrary, linguistic vocalizations have been extensively investigated to detect hesitations in spontaneous speech [80] with the main purpose of improving speech recognition systems. The disfluencies are typically detected by mapping acoustic observations (e.g. pitch and energy) into classes of interest with classifiers like neural networks or Support Vector Machines. The detection of silence is one of the earliest tasks studied in speech analysis and robust algorithms, based on the distribution of the energy, have been developed since the earliest times of digital signal processing [35]. The last important aspect of vocal behaviour, i.e. the turn taking, is typically a side-product of the speaker diarization, i.e. of the segmentation of speech recordings into single speaker segments (see [87] for a survey). This allows to recognize who speaks with whom and to measure the nonverbal behaviour in correspondence of speaker transitions. Turn-taking has been used to model influence and dominance relationships [18].

3.1.5 Social Signals from Use of Space and Environment

Physical proximity information has been used in *reality mining* applications (see Section 4) as a social cue accounting for the simple presence or absence of interaction between people [24][67]. These works use specially equipped cellular phones capable of sensing the presence of similar devices in the vicinity. The automatic detection of seating arrangements has been proposed as a cue for retrieving meeting recordings in [38]. Several approaches developed in computer surveillance to track people across public spaces can potentially be used to address the detection of social signals in the use of the space.

3.2 Context Sensing

No correct interpretation of human behavioural cues in social interactions is possible without taking into account the *context*, namely *where* the interactions take place, *what* is the activity of the individuals involved in the interactions, *when* the interactions take place, and *who* is involved in the interaction. Note, however, that while W4 (*where*, *what*, *when*, *who*) is dealing only with the apparent perceptual aspect of the context in which the observed human behaviour is shown, human behaviour understanding is about W5+ (*where*, *what*, *when*, *who*, *why*, *how*), where the *why* and *how* are directly related to recognizing communicative intention including social behaviours, affective 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 [62]), answering the *why* and *how* questions in a W4-context-sensitive manner when analysing human behaviour is virtually unexplored area of research.

3.3 Social Behaviour Understanding

The two main challenges in human behaviour analysis are the modeling of temporal dynamics and the combination of cues extracted from different modalities and at different time scales.

Temporal dynamics of social behavioural cues (i.e., their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed social behaviour [3][27]. However, relatively few approaches explicitly take into account the temporal evolution of behavioural cues to understand social behaviour. Some of them aim at the analysis of facial expressions involving sequences of Action Units (i.e., atomic facial gestures) [86], as well as coordinated movements of head and shoulders [89]. Others model the evolution of collective actions in meetings using Dynamic Bayesian Networks [21] or hidden Markov models [49].

Social signals are spoken and wordless messages like head nods, bow ties, winks, *uh* and *yeah* utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech. Hence, automated analyzers of human social signals and social behaviours should be multimodal, fusing and analyzing verbal and non-verbal interactive signals coming from different modalities (speech, body gestures, facial and vocal expressions). 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 modelled 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 have been proposed that aim at making use of the correlation between audio and visual data streams, and relax the requirement of synchronization of these streams [28].However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is largely unexplored. A much broader focus on the issues relevant to multimodal temporal fusion is needed including the optimal level of integrating these different streams, the optimal function for the integration, and how estimations of reliability of each stream can be included in the inference process. In addition, how to build context-dependent multimodal fusion is another open and highly relevant issue.

4. MAIN APPLICATIONS OF SOCIAL SIGNAL PROCESSING

The expression *Social Signal Processing* has been used for the first time in [68] to group under a collective definition several pioneering works of Alex Pentland and his group at MIT. These works aimed at two main applications, one one hand, the prediction, with an accuracy of more than 70%, of behavioural outcomes like the result of salary negotiations, hiring interviews, or speed-dating conversations [18]. On the other hand, the analysis of large groups of individuals (around 100 people) through smart cellular phones equipped with proximity detectors and vocal activity analyzers [24][67] (an application called *reality mining*).

In the same years, few other groups worked on the analysis of social interactions in multimedia recordings targeting three main areas: the analysis of interactions in small groups, the recognition of roles, and the sensing of users attitudes towards computer interfaces.

The research on interactions in small groups has focused on the detection of dominant persons and on the recognition of collective actions. The problem of dominance is addressed in [36][74], where multimodal approaches combine

several nonverbal features, mainly speaking energy and body movement, to identify at each moment who is the dominant individual. The same kind of features has been applied in [21][49] to recognize the actions performed in meetings like discussions, presentations, etc. The combination of the information extracted from different modalities is performed with different algorithms including Dynamic Bayesian Networks [54] and layered hidden Markov models [57].

The recognition of roles has been addressed in two main contexts: broadcast material [7][90][92] and small scale meetings [6][23][96]. The works in [90][92] apply Social Network Analysis [91] to detect the role of people in broadcast news and movies, respectively. The approach in [7] recognizes the roles of speakers in broadcast news using vocal behaviour and lexical features. The roles in meetings are recognized using nonverbal behaviour in the case of [6], while a multimodal approach including both audio and visual features is applied in [23][96].

The reaction of users to social signals exhibited by computers has been investigated in several works showing that computers are *social actors*, i.e., they elicit the same reactions and perceptions as humans [55]. This happens, e.g., when children tend to imitate the voice quality of cartoon characters appearing on the interface of didactic applications [59], or when beta testers provide higher appreciation scores for interfaces exhibiting some form of *mimicry*, i.e. of the behaviour imitation typically displayed by humans to mean affiliation and liking [4].

5. CONCLUSIONS AND FUTURE PERSPECTIVES

Social Signal Processing has the ambitious goal of bringing social intelligence [2] in computers. The first results in this research domain have been sufficiently impressive to attract the praise of the technology [32] and businness [10] communities. What is more important is that they have established a viable interface between human sciences and engineering - social interactions and behaviours, although complex and rooted in the deepest aspects of human psychology, can be analyzed automatically with the help of computers. This "cultural" breakthrough is, in our opinion, the most important result of research in SSP so far. In fact, the pioneering contributions in SSP [65][66] have shown that the social signals, typically described as so elusive and subtle that only trained psychologists can recognize them [30], are actually evident and detectable enough to be captured through sensors like microphones and cameras, and interpreted through analysis techniques like machine learning and statistics.

However, this is nothing else than the first step and the survey of the main SSP applications presented in Section 4 clearly shows that current approaches have a number of serious limitations. So far, SSP has been driven by technology researchers with no training in social sciences. Thus important aspects of social interactions are likely to be neglected. For example, most of the current SSP approaches are monomodal, even if social behaviour is multimodal in nature and the integration of multiple cues is a key aspect of social perception in humans. Moreover, most of the SSP approaches presented so far deal with laboratory data created in artificial settings. Hence, it is hard to assess the actual effectiveness of the approaches that might be favored by the specific experimental constraints imposed. For the above reasons, the rest of this section discusses four challenges facing the researchers in the field, for which we believe are the crucial turnover issues that need to be addressed before the research in the field can enter its next phase - the deployment phase.

The first issue relates to *tightening of the collaboration between social scientists and engineers*. The analysis of human behaviour in general, and social behaviour in particular, is an inherently multidisciplinary problem [62]. More specifically, no automatic analysis of social interactions is possible without taking into account the basic mechanisms governing social behaviours that the psychologists have investigated for decades, such as the *chameleon effect* (mutual imitation of people aimed at showing liking or affiliation) [12][45], the interpersonal adaptation (mutual accommodation of behavioural patterns between interacting individuals) [33], the interactional synchrony (degree of coordination during interactions) [42], the presence of roles in groups [5][85], the dynamics of conversations [72][95], etc.

The second issue relates to the need of implementing *multi-cue, multi-modal approaches* to SSP. Nonverbal behaviours cannot be read like words in a book [44][73]; the relationship between behavioural cues and social signals is influenced by a multiplicity of factors that are difficult to model like context and culture. The best way to deal with the resulting inherent ambiguity is to use combinations of multiple cues, if possible extracted from multiple modalities. This corresponds to findings in social psychology, that show that humans use multiple cues to assess effectively social situations [75]. Also, combining multiple classifiers in machine learning has shown to be effective for tackling this problem as long as the single classifiers are *diverse*, i.e., they account for different aspects of the problem of interest.

The third issue relates to *the use of real-world data*. Both psychologists and engineers tend to produce their data in laboratories and artificial settings (see e.g., [18][49]), in order to limit parasitic effects and elicit the specific phenomena they want to observe. However, this is likely to simplify excessively the situation and to improve artificially the performance of the automatic approaches. Moreover, many aspects of the actual social behaviour of people are likely to be missing if the interactions do not have a real impact on the life of the participants. Social interactions are a ubiquitous phenomenon that can be observed and captured in a wide range of real-world scenarios and SSP should focus on these rather than on artificial settings.

The last, but not least, challenging issue relates to the *the identification of applications likely to benefit from SSP*. Applications have the important advantage of linking the effectiveness of detecting social signals to the reality. For example, one of the earliest applications is the prediction of the outcome in transactions recorded at a call center and the results show that the number of successful calls can be increased by around 20% by stopping early the calls that are not promising [10]. This can have not only a positive impact on the marketplace, but also provide *benchmarking procedures* for the SSP research, one of the best means to improve the overall quality of a research domain as extensively shown in fields where international evaluations take place every year.

Acknowledgements. The work of Dr. Vinciarelli and Prof. Bourlard is supported by the Swiss National Science

Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). The work of Dr. Pantic is supported in part by the EU IST Programme project FP6-0027787 (AMIDA) and the EC's 7^{th} Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

6. REFERENCES

[1] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami. The automatic measurement of facial beauty. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 2644–2647, 2001.

[2] K. Albrecht. *Social Intelligence: The new science of success*. John Wiley & Sons Ltd, 2005.

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[4] J. Bailenson, N. Yee, K. Patel, and A. Beall. Detecting digital chameleons. *Computers in Human Behavior*, 24(1):66–87, 2008.

[5] R. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, 1950.

[6] S. Banerjee and A. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, pages 2189–2192, 2004.

[7] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of American Association of Artificial Intelligence Symposium*, pages 679–684, 2000.

[8] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[9] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 405–410, 2002.

[10] M. Buchanan. The science of subtle signals. *Strategy+Business*, 48:68–77, 2007.

[11] N. Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Speech and Language Processing*, 14(4):1171–1178, 2006.

[12] T. Chartrand and J. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[13] J. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 233–238, 2006.

[14] J. Cohn and P. Ekman. Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In J. Harrigan, R. Rosenthal, and K. Scherer, editors, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 9–64. 2005.

[15] J. Cortes and F. Gatti. Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5):432–439, 1965.

[16] D. Crystal. *Prosodic Systems and Intonation in English*. Cambridge University Press, 1969.

[17] D. Cunningham, M. Kleiner, H. Bülthoff, and C. Wallraven. The components of conversational facial expressions. *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*, pages 143–150, 2004.

[18] J. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802–811, 2007.

[19] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.

[20] R. De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Journal of Computational Animation and Virtual World*, 15(3-4):269–276, 2004.

[21] A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25, 2007.

[22] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, 1972.

[23] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 271–278, 2007.

[24] N. Eagle and A. Pentland. Reality mining: sensing complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[25] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2005.

[26] P. Ekman and W. Friesen. Facial action coding system (FACS): Manual. 2002.

[27] P. Ekman and E. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[28] N. Fragopanagos and J. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[29] H. Giles, N. Coupland, and J. Coupland. Accomodation theory: communication, context, consequence. In *Contexts of accomodation: developments in applied sociolinguistics*, pages 1–69. Cambridge University Press, 1991.

[30] M. Gladwell. *Blink: The Power of Thinking without Thinking*. Little Brown & Company, 2005.

[31] C. Glass, T. Merluzzi, J. Biever, and K. Larsen. Cognitive assessment of social anxiety: Development and validation of a self-statement questionnaire. *Cognitive Therapy and Research*, 6(1):37–55, 1982.

[32] K. Greene. 10 emerging technologies 2008. *MIT Technology Review*, February 2008.

[33] S. Gregory, K. Dagan, and S. Webster. Evaluating the relation of vocal accommodation in conversation partners fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43, 1997.

[34] E. Hall. *The silent language*. Doubleday, 1959.

[35] X. Huang, A. Acero, and H. Hon. *Spoken language processing*. Prentice Hall, 2001.

[36] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the ACM International Conference on Multimedia*, pages 835–838, 2007.

[37] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Proceedings of th International Conference on Cyberworlds*, pages 437–444, 2005.

[38] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proceedings of Workshop on Continuous Archival and Retrieval of Personal Experiences*, pages 74–85, 2004.

[39] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J. Haviland-Jones, editors, *Handbook of Emotions*, pages 236–249. 2000.

[40] D. Keltner and J. Haidt. Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5):505–521, 1999.

[41] L. Kennedy and D. Ellis. Laughter detection in meetings. In *Proceedings of the NIST Meeting Recognition Workshop*, 2004.

[42] M. Kimura and I. Daibo. Interactional synchrony in conversations about emotional episodes: A measurement by "the between-participants pseudosynchrony experimental paradigm". *Journal of Nonverbal Behavior*, 30(3):115–126, 2006.

[43] A. Kleinsmith, R. De Silva, and N. Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6):1371–1389, 2006.

[44] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972.

[45] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand. The Chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162, 2003.

[46] Y. Li, S. Ma, and H. Lu. Human posture recognition using multi-scale morphological method and Kalman motion estimation. In *Proceedings of International Conference on Pattern Recognition*, pages 175–177, 1998.

[47] G. Littlewort, M. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 15–21, 2007.

[48] S. Lucey, A. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 275–286. I-Tech Education and Publishing, 2007.

[49] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

[50] A. Mehrabian and S. Ferris. Inference of attitude from nonverbal communication in two channels. *Journal of Counseling Psychology*, 31(3):248–252, 1967.

[51] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[52] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[53] S. Mota and R. Picard. Automated posture analysis for detecting learner's interest level. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 49–56, 2003.

[54] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California Berkeley, 2002.

[55] C. Nass and K. Lee. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001.

[56] A. Oikonomopoulos, M. Pantic, and I. Patras. B-spline polynomial descriptors for human activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[57] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 3–8, 2002.

[58] A. J. O'Toole, T. Price, T. Vetter, J. Bartlett, and V. Blanz. 3D shape and 2D surface textures of human faces: the role of "average" in attractiveness and age. *Image and Vision Computing*, 18(1):9–19, 1999.

[59] S. Oviatt, C. Darves, and R. Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction*, 11(3):300–328, 2004.

[60] L. Ozer and W. Wolf. Real-time posture and activity recognition. In *Proceedings of Workshop on Motion and Video Computing*, pages 133–138, 2002.

[61] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 377–416. I-Tech Education and Publishing, 2007.

[62] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human-centred intelligent human-computer interaction (HCI2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[63] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[64] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 758–765, 2002.

[65] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, 2004.

[66] A. Pentland. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[67] A. Pentland. Automatic mapping and modeling of human networks. *Physica A*, 378:59–67, 2007.

[68] A. Pentland. Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007.

[69] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5117–5121, 2008.

[70] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 945–948, 1998.

[71] R. Pope. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1-2):4–18, 2007.

[72] G. Psathas. *Conversation Analysis - The study of talk-in-interaction*. Sage Publications, 1995.

[73] V. Richmond and J. McCroskey. *Nonverbal Behaviors in interpersonal relations*. Allyn and Bacon, 1995.

[74] R. Rienks, D. Zhang, and D. Gatica-Perez. Detection and application of influence rankings in small group meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 257–264, 2006.

[75] J. Russell, J. Bachorowski, and J. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Reviews in Psychology*, 54(1):329–349, 2003.

[76] A. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27:316–331, 1964.

[77] K. Scherer. *Personality markers in speech*. Cambridge University Press, 1979.

[78] K. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, 2003.

[79] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll. Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 30–37, 2007.

[80] E. Shriberg. Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences*, 1:619–622, 1999.

[81] E. Shriberg, A. Stolcke, and D. Baron. Observations of overlap: findings and implications for automatic processing of multiparty conversation. In *Proceedings of Eurospeech*, pages 1359–1362, 2001.

[82] P. Shrout and D. Fiske. Nonverbal behaviors and social evaluation. *Journal of Personality*, 49(2):115–128, 1981.

[83] K. Sjölander and J. Beskow. Wavesurfer-an open source speech tool. In *Proceedings of International Conference on Spoken Language Processing*, pages 464–467, 2000.

[84] L. Smith-Lovin and C. Brody. Interruptions in group discussions: the effects of gender and group composition. *American Sociological Review*, 54(3):424–435, 1989.

[85] H. Tischler. *Introduction to Sociology*. Harcourt Brace College Publishers, 1990.

[86] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[87] S. Tranter and D. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[88] K. Truong and D. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[89] M. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 38–45, 2007.

[90] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007.

[91] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[92] C. Weng, W. Chu, and J. Wu. Movie analysis based on roles social network. In *proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007.

[93] A. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–455, 2003.

[94] M. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[95] G. Yule. *Pragmatics*. Oxford University Press, 1996.

[96] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006.

[97] B. Zellner. Pauses and the temporal structure of speech. In E. Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. John Wiley & Sons, 1994.

[98] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.