Social Signal Processing: Understanding Social Interactions through Nonverbal Behavior Analysis

A.Vinciarelli and H.Salamin*
Idiap Research Institute
CP592 - 1920 Martigny (Switzerland)
EPFL - 1015 Lausanne (Switzerland)
{vincia,hsalamin}@idiap.ch

M.Pantic†
Imperial College London
108 Queens Gate London
EEMCS - University of Twente
m.pantic@imperial.ac.uk

Abstract

This paper introduces Social Signal Processing (SSP), the domain aimed at automatic understanding of social interactions through analysis of nonverbal behavior. The core idea of SSP is that nonverbal behavior is the machine detectable evidence of social signals, the relational attitudes exchanged between interacting individuals. Social signals include (dis-)agreement, empathy, hostility, and any other attitude towards others that cannot be expressed using just words. Thus, nonverbal behavior analysis is used as a key to automatic understanding of social interactions. This paper presents not only a survey of the related literature and the main concepts underlying SSP, but also an illustrative example of how such concepts are applied, with particular attention to the integration of human sciences (psychology, anthropology, sociology, etc.) findings in technology.

1. Introduction

Imagine to watch the television in a country of which you do not know the language. While you cannot understand what is being said, you can still catch a good deal of information about social interactions taking place on the screen. You can easily spot the most important guest in a talk-show, understand whether the interaction is tense or relaxed, guess the kind of relationships people on the video have in their life (e.g., whether there are couples or members of the same soccer team), etc.

How can we be so effective in interpreting social interactions without the need of understanding what is being said? Psychologists have been studying this phenomenon for decades and they have shown that humans are literally wired for extracting social information from *nonverbal communication*, i.e. from the large variety of nonverbal behavioral cues accompanying human-human interactions [35][54]. Any facial expression, vocal outburst, gesture, posture, etc. elicits the perception, often unconscious, of socially relevant information [3]. Furthermore, this mechanism seems to be so deeply rooted in our brain, that we cannot escape it even when we deal with synthetic faces [10] and voices [44] generated by computers.

If nonverbal communication plays such an important role in our life, why should computers be unable to capture the social meaning of nonverbal cues? This is exactly the problem addressed by Social Signal Processing (SSP), the new, emerging, domain aimed at understanding social interactions through machine analysis of nonverbal behavior [52][68][69][70]. The core SSP idea is that nonverbal behavioral cues that humans so easily sense with their eyes and ears can be detected with microphones, cameras and any other suitable sensor. The cues can then be used as a machine detectable evidence for automatic analysis and understanding of social behavior.

SSP will bring computing closer to *Human-Centred* approaches effectively dealing with psychological and behavioral responses natural for humans, in contrast with computing-centred approaches that require people to operate following technology driven criteria. This will have a major impact on Human-Computer Interaction technologies [48] because interfaces will become more adept to social interactions with users, on multimedia content analysis techniques [22] because content will be analyzed according to the way humans perceive the reality around them, on

*The work of A. Vinciarelli and H.Salamin has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet) and in part by the Swiss National Science Foundation through the National Center of Competence in Re- search on Interactive Multimodal Information Management (IM2).

†The work of M. Pantic has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet), and in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

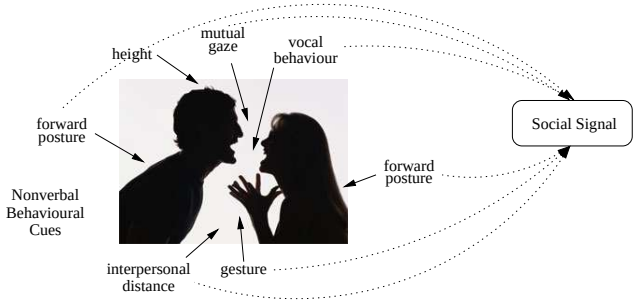

Figure 1. Social signals. A constellation of nonverbal behavioral cues (posture, interpersonal distance, gestures, etc.) is perceived as a social signal (hostility, aggressiveness, disagreement, etc.).

computer mediated communication (e.g., see [25]) because transmission will include the social cues necessary for establishing a natural contact with others, and any other domain where computers must seamlessly integrate the life of people.

This paper starts by introducing the most important aspects of nonverbal communication (Section 2), then it illustrates the main technological components necessary to analyze social behavior (Section 3). After, it continues with an example showing how SSP principles and ideas are applied to a specific case (Section 4), before providing a quick survey of the main SSP applications presented so far in the literature. The final Section 6 proposes final remarks.

2. Nonverbal Behavior and Social Signals

Following one of the most common definitions:

> "Nonverbal communication includes all the messages other than words that people exchange in interactive contexts" [31]

In some cases, the messages are exchanged consciously and nonverbal behaviors have a precise and shared meaning attached to them (e.g., the *thumbs up* gesture). Most frequently, nonverbal behavior *gives off* messages that leak information about the state of people, e.g. about their emotions, self-confidence, status, etc. [26].

SSP focuses on the latter type of communication and, in particular, on *social signals* [2], the *relational attitudes* displayed by people during social interactions. Consider Figure 1: it is not difficult to guess that the two individuals are a couple and they are fighting even if the only information at disposition are their silhouettes. The reason is that the picture shows a sufficient number of nonverbal behavioral cues to correctly understand the kind of interaction taking place. Mouths wide open suggest that the two persons are shouting, the tension of gestures shows that the atmosphere is not relaxed, the distance is too close for persons not sharing an intimate relationship, etc.

For the sake of simplicity, psychologists have grouped all possible nonverbal behavioral cues into five major classes called *codes* [31]. The first is ***physical appearance***, including not only somatic characteristics, but also clothes and ornaments that people use to modify their aspect. While human sciences have extensively investigated the role of appearance in social interactions (e.g., see [18] for the effect of attractiveness and [12] for the influence of body shape on social perceptions), only few works, to the best of our knowledge, have been dedicated to the automatic analysis of the way people look. These are mostly dedicated to the attractiveness of faces (e.g., [28]) and to the recognition of clothes for tracking and surveillance purposes (e.g., [15]).

The second code includes ***gestures and postures***, extensively investigated in human sciences because they are considered the most reliable cue about the actual attitude of people towards others and social situations (see [54] and references therein). Automatic analysis of gestures is a hot topic in technology as well, but the goal is mainly to replace keyboards and mices with hand movements as computer interfaces (see [73] for a survey). Only lately gestures and postures have being analyzed for their affective content (see [29] for a survey).

Face and eye behavior is a crucial code, as face and eyes are our direct and naturally preeminent means of communicating and understanding somebodys affective state and intentions on the basis of the shown facial expression [33]. Not surprisingly facial expressions and gaze behavior have been extensively studied in both human sciences and technology. The first study on facial expressions dates back to Darwin [16], and a comprehensive framework for the description of facial expressions (and messages they convey) has been elaborated in the last decades [21]. Facial expression analysis is a well established domain (see [77] for the most recent and extensive survey), and gaze has been the subject of significant attention in the last years [64].

Vocal behavior is the code that accounts for *how* something is said and includes the following aspects of spoken communication [35][54]: voice quality (prosodic features like pitch, energy and rythm), linguistic (expressions like "*ehm*", "*ah*", etc.) and non-linguistic (laughter, crying, sobbing, etc.) vocalizations, silences (use of pauses), and turn-taking patterns (mechanisms regulating floor exchange) [53][75]. Each one of them relates to social signals that contribute to different aspects of the social perception of a message. Both human sciences and technology have extensively investigated vocal behavior. The former have shown, e.g., that vocal behavior plays a role in expression of emotions [57], is a personality marker [56], and is used to display status and dominance [59]. The speech analysis community has worked on the dection, e.g., of disfluencies [58], non-linguistic vocalizations (e.g., particular laughter [34][62]), or rythm [42], but only with the goal of

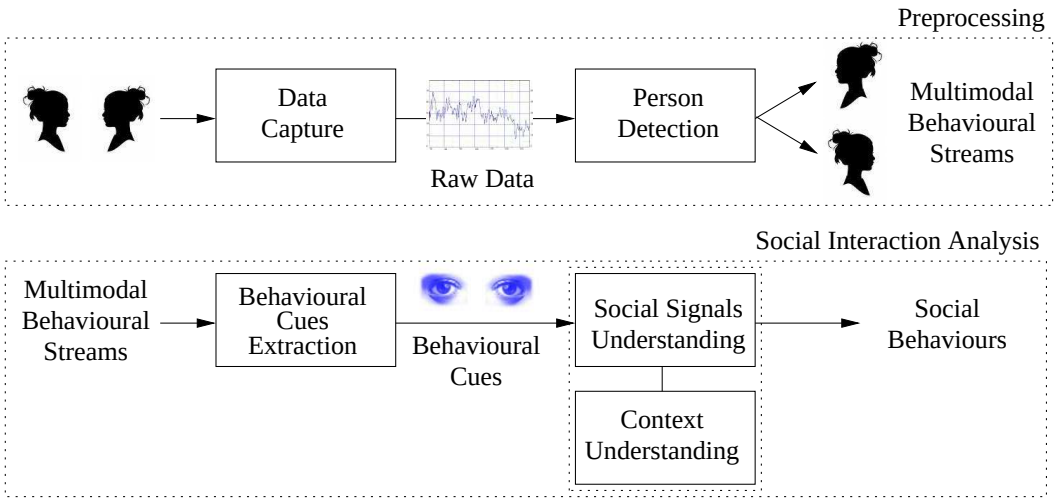

Figure 2. Machine analysis of social signals and behaviours: a general scheme. The process includes two main stages: The preprocessing, takes as input the recordings of social interaction and gives as output the multimodal behavioural streams associated with each person. The social interaction analysis maps the multimodal behavioural streams into social signals and social behaviours.

improving the speech recognition performance. In parallel, the speech synthesis community has investigated the synthesis of vocal behavior as a means to make artificial voices more natural [9].

The last code includes behaviors related to ***space and environment***, i.e. the way people share and organize ambients they have at disposition. Human sciences have investigated this code, showing in particular that people tend to organize the space around them in concentric zones accounting for different relationships they have with others [30]. For example, Figure 1 shows an example of individuals sharing the *intimate zone*, the concentric area closest to each individual. Technology has started only recently to study the use of space, but only for tracking and surveillance purposes.

3. State-of-the-art

Figure 2 shows the main technological components (and their interrelationships) of a general SSP system. The scheme does not correspond to any approach in particular, but most SSP works presented in the literature match, at least partially, the processing chain in the picture (see Section 5).

The first, and crucial, step is the ***data capture***. Depending on how the data is captured, then only certain kinds of processing are possible and not others. The most commonly used capture devices are microphones and cameras (with arrangements that go from a simple laptop webcam to a fully equipped smart meeting room [38][71]), but the literature reports the use of wearable devices [20] and pressure captors [43] (for recognizing posture of sitting people) as well.

In most cases, the raw data involve different persons (e.g., the recording of a conversation where different voices can be heard at different moments in time). Thus, a ***person detection*** step is necessary to know which part of the data corresponds to which person (e.g., who talks when in the recording of a conversation). This is typically performed with speaker diarization [61], face detection [74], tracking [40], or any othe kind of technique that allows one to identify intervals of time or images regions corresponding to specific individuals.

Person detection is the step preliminary to ***Behavioral cues extraction***, i.e. the detection of nonverbal behaviors displayed by each individual. Some approaches for this stage have been cited in Section 2. Furthermore, extensive are available in [68][69][70].

The two main challenges in ***social behavior understanding*** 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 [2][21]. 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) [60], as well as coordinated movements of head and shoulders [63]. Others model the evolution of collective actions in meetings using Dynamic Bayesian Networks [17] or hidden Markov models [39].

To address the second challenge outlined above (combination of cues), a number of model-level fusion methods have been proposed that aim at making use of the corre-

lation between audio and visual data streams, and relax the requirement of synchronization of these streams [23]. However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is largely unexplored.

Context Understanding is desirable because 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 [48]), answering the *why* and *how* questions in a W4-context-sensitive manner when analysing human behaviour is virtually unexplored area of research.

4. An Example: the Analysis of Conflicts

This section aims at providing a concrete example of how principles and ideas outlined in previous sections are applied to a concrete case, i.e. the analysis of conflicts in competitive discussions. Conflicts have been extensively investigated in human sciences. The reason is that they influence significantly the outcome of groups expected to reach predefined targets (e.g., work teams) or to satisfy members needs (e.g., families) [37].

This section focuses on political debates because these are typically built around the conflict between two fronts (including one or more persons each) that defend opposite views or compete for a reward (e.g., the attribution of an important political position) that cannot be shared by the two parts. The corpus used for the experiments includes 45 debates (roughly 30 hours of material) revolving around a *yes/no* question like "*are you favorable to new laws on environment protection?*". Each debate involves one moderator, two guests supporting the *yes* answer, and two guests supporting the *no* answer. The guests state their answer explicitly at the beginning of the debate and this allows one to label them unambiguously in terms of their position.

The goal of the experiments is 1) to identify the moderator, and 2) to reconstruct correctly the two groups (*yes* and *no*) resulting from the structure outlined above. The next sections show how the different steps depicted in Figure 2 are addressed.

4.1. Nonverbal Behavior in Conflicts

Human sciences have studied conversations in depth as these represent one of the most common forms of social interaction [53]. Following [75], conversations can be thought of as markets where people compete for the *floor* (the right of speaking):

> [...] the most widely used analytic approach is based on an analogy with the workings of the market economy. In this market there is a scarce commodity called the ***floor*** which can be defined as the right to speak. Having control of this scarce commodity is called a ***turn***. In any situation where control is not fixed in advance, anyone can attempt to get control. This is called ***turn-taking***.

(boldface as in the original text). This suggests that *turn-taking* is a key to understand conversational dynamics.

In the specific case of conflicts, social psychologists have observed that people tend to react to someone they disagree with rather than to someone they agree with [53][75]. Thus, the social signal conveyed by direct reaction is likely to be *disagreement* and the corresponding nonverbal behavioral cue is adjacency in the speakers sequence. This social psychology finding determines the design of the conflict analysis approach described in the rest of this section.

4.2. Data Capture and Person Detection

The previous section suggests that turn-taking is the key to understand conversational dynamics in conflicts. The data at disposition are television political debates and the turn-taking can be extracted from the audio channel using a speaker diarization approach (see [61] for an extensive survey on diarization). The diarization approach used in this work is not presented here for space limitations (see [1] for a full description). However, what is important for the example presented in this section is that the audio channel of the political debates is converted into a sequence S:

$$S = \{(s_1, t_1, \Delta t_1), \ldots, (s_N, t_N, \Delta t_N)\}, \qquad (1)$$

where each triple accounts for a turn and includes a speaker label $s_i \in A = \{a_1, \ldots, a_G\}$ identifying the person speaking during the turn, the starting time t_i of the turn, and the duration Δt_i of the turn (see Figure 3). Thus, the sequence S contains the whole information the turn-taking corrsponds to, namely *who talks when and how much*. The *purity* (see [66] for a definition of the purity) of the resulting speaker segmentation is 0.92, meaning that the groundtruth speaker segmentation is mostly preserved.

The diarization can be considered a form of person detection because it identifies the parts of the data that correspond to each person. In the case of this work, this allows

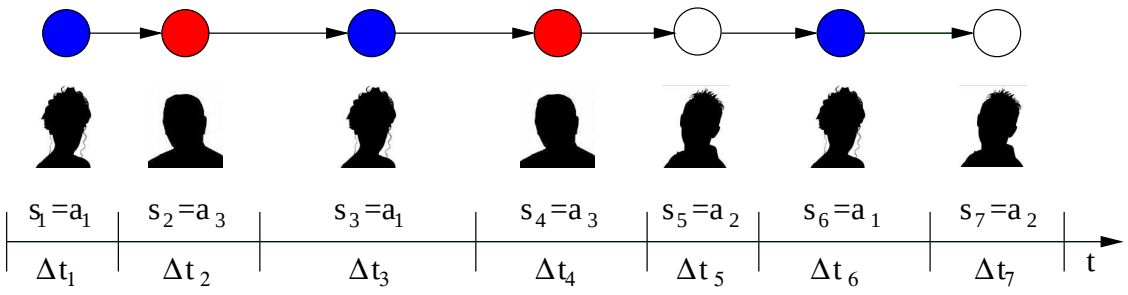

Figure 3. Turn-Taking pattern. The figure shows an example of turn-taking where three persons are assigned to different states.

the identification of speaker adjancencies that are the behavioral cues correlated to the social behaviour of interest, i.e. the expression of a greement and disagreement between debate participants.

4.3. Social Signal Understanding

The suggestion that people tend to react to someone they disagree with rather than to someone they agree with can be expressed, in mathematical terms, by saying that speaker s_i is statistically dependent on speaker s_{i-1} (see Figure 3). Statistical dependence between sequence elements that follow one another can be modeled using a Markov Chain where the set Q of the states contains three elements, namely T_1 (the first group), T_2 (the second group) and M (the moderator).

If $\varphi : A \rightarrow Q$ is a mapping that associates a speaker $s_i \in A$ with a state $q_j \in Q$, then the conflict analysis problem can be thought of as finding the mapping φ^* satisfying the following expression:

$$\varphi^* = \arg \max_{\varphi \in Q^A} p(\varphi(s_1)) \prod_{n=2}^{N} p(\varphi(s_n)|\varphi(s_{n-1})), \quad (2)$$

where N is the number of turns in the turn-taking, $p(\varphi(s_1))$ is the probability of starting with state $q_1 = \varphi(s_1)$, and $p(\varphi(s_n)|\varphi(s_{n-1}))$ is the probability of a transition between state $q_n = \varphi(s_n)$ and state $q_{n-1} = \varphi(s_{n-1})$.

The probability on the left side of Equation (2) has the same value if all the speakers assigned state T_1 are switched to state T_2 and viceversa. In other words, the model is symmetric with respect to an exchange between T_1 and T_2. The reason is that T_1 and T_2 are meant to distinguish between members of different groups and not to account for membership to a specific group.

The Markov Model is trained using a leave-one-out approach: all debates at disposition but one are used as training set, while the left out is used as test set. The experiment is reiterated and each time a different debate is used as test set. The results show that 64.5% of the debates are correctly reconstructed, i.e., the moderator is correctly identified and the two supporters of the same answer are assigned the same state. Such a figure goes up to 75% when using the groundtruth turn-taking (and not the turn-taking as automatically extracted from the audio data). The average performance of an algorithm assigning the states randomly is 6.5% and this means that the above model, even if rather simple, still performs ten times better than chance.

4.4. The SSP Approach

The example presented in this section shows how SSP principles and ideas are applied to a concrete case. First, human sciences provide suggestions about the behavioral cues related to a social phenomenon of interest. In this case, reaction during a debate is identified as a behavioral cue related to disagreement, thus to the composition of groups in competitve discussions. Second, signal processing approaches are applied to the data to extract the behavioral cue of interest. In this case, a speaker diarization technique is used to extract the turn-taking and the sequence of speakers in a competitve discussion. Third, a Machine Learning approach is used to model the behavioral cue and understand the social phenomenon of interest. In this case, a simple Markov Model captures the structure of a political debate in terms of opposite groups and person who plays the moderator role.

5. Main SSP Applications

The first extensive surveys of SSP applications have been proposed in [68][69][70], after that the expression *Social Signal Processing* has been introduced for the first time in [52] to group under a collective definition several pioneering works published by Alex Pentland and his group at MIT.

The earliest SSP works focused on vocal behavior with the goal of predicting (with an accuracy higher than 70%) the outcome of dyadic interactions such as salary negotiations, hiring interviews, and speed dating conversations [14]. One of the most important contributions of these works is the definition of a coherent framework for the analysis of vocal behavior [49][50], where a set of cues accounts for *activity* (the total amount of energy in the speech sig-

nals), *influence* (the statistical influence of one person on the speaking patterns of the others), *consistency* (stability of the speaking patterns of each person), and *mimicry* (the imitation between people involved in the interactions). Recent approaches for the analysis of dyadic interactions include the visual analysis of movements for the detection of interactional synchrony [41][41].

Other approaches, developed in the same period as the above works, have aimed at the analysis of small group interactions [37], with particular emphasis on meetings and broadcast data (talk-shows, news, etc.). Most of the works have focused on recognition of collective actions [17][39], dominance detections [32][55], and role recognition [6][19][24][36][76]. The approaches proposed in these works are often multimodal [17][19][32][39][55][76], and the behavioral cues most commonly extracted correspond to speaking energy and amount of movement. In many cases, the approaches are based only on audio, with features that account for turn-taking patterns (when and how much each person talks) [6][36], or for combinations of social networks and lexical features [24].

Social network analysis has been applied as well in [65][67][72] to recognize the roles played by people in production environment data (movies, radio and television programs, etc.), and in an application domain known as *reality mining*, where large groups of individuals equipped with smart badges or special cellular phones are recorded in terms of proximity and vocal interactions and then represented in a social network [20][51].

The reaction of users to social signals exhibited by computers has been investigated in several works showing that people tend to behave with machines as they behave with other humans. The effectiveness of computers as *social actors*, i.e., entities involved in the same kind of interactions as the humans, has been explored in [44][45][46], where computers have been shown to be attributed a personality and to elicit the same reactions as those elicited by persons. Similar effects have been shown in [13][47], where children interacting with computers have modified their voice to match the speaking characteristics of the animated personas of the computer interface, showing adaptation patterns typical of human-human interactions [8]. Further evidence of the same phenomenon is available in [4][5], where the interaction between humans and computers is shown to include the *Chameleon effect* [11], i.e. the mutual imitation of individuals due to reciprocal appreciation or to the influence of one individual on the other.

6. Conclusion

This paper has introduced the core SSP principles and an example of how these are applied to a concrete case, namely the analysis of conflicts in political debates. While still in its pioneering phase, SSP has produced results sufficiently convincing to attract the praise of both technology [27] and businness [7] communities. However, the most important result is that a viable interface between human sciences and technology has been established with the purpose of analyzing social phenomena rooted in the deepest aspects of human psychology.

References

[1] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication*, 40(3):351–363, 2003. 4

[2] N. Ambady, F. Bernieri, and J. Richeson. Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In M. Zanna, editor, *Advances in Experimental Social Psychology*, pages 201–272. 2000. 2, 3

[3] M. Argyle. *The Psychology of Interpersonal Behaviour*. Penguin, 1967. 1

[4] J. Bailenson and N. Yee. Virtual interpersonal touch and digital chameleons. *Journal of Nonverbal Behavior*, 31(4):225–242, 2007. 6

[5] J. Bailenson, N. Yee, K. Patel, and A. Beall. Detecting digital chameleons. *Computers in Human Behavior*, 24(1):66–87, 2008. 6

[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. 6

[7] M. Buchanan. The science of subtle signals. *Strategy+Business*, 48:68–77, 2007. 6

[8] J. Burgoon, L. Stern, and L. Dillman. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995. 6

[9] N. Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Speech and Language Processing*, 14(4):1171–1178, 2006. 3

[10] J. Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000. 1

[11] 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. 6

[12] J. Cortes and F. Gatti. Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5):432–439, 1965. 2

[13] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *International Conference on Spoken Language Processing*, pages 2689–2692, 2002. 6

[14] 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. 5

[15] 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. 2

[16] C. Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, 1872. 2

[17] A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25, 2007. 3, 6

[18] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, 1972. 2

[19] 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. 6

[20] N. Eagle and A. Pentland. Reality mining: sensing complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006. 3, 6

[21] 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. 2, 3

[22] A. Elgammal. Human-Centered Multimedia: representations and challenges. In *Proc. ACM Intl. Workshop on Human-Centered Multimedia*, pages 11–18, 2006. 1

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

[24] 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. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008. 6

[25] J. Gemmell, K. Toyama, C. Zitnick, T. Kang, and S. Seitz. Gaze awareness for video-conferencing: A software approach. *IEEE Multimedia*, 7(4):26–35, 2000. 2

[26] E. Goffman. *The presentation of self in everyday life*. Anchor Books, 1959. 2

[27] K. Greene. 10 emerging technologies 2008. *MIT Technology Review*, february 2008. 6

[28] H. Gunes and M. Piccardi. Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64(12):1184–1199, 2006. 2

[29] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In J. Or, editor, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218. 2008. 2

[30] E. Hall. *The silent language*. Doubleday, 1959. 3

[31] M. Hecht, J. De Vito, and L. Guerrero. Perspectives on nonverbal communication. codes, functions and contexts. In L. Guerrero, J. De Vito, and M. Hecht, editors, *The nonverbal communication reader*, pages 201–272. 2000. 2

[32] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language: Special Issue on Multimedia, to appear*, 2009. 6

[33] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J. Haviland-Jones, editors, *Handbook of Emotions*, pages 236–249. 2000. 2

[34] L. Kennedy and D. Ellis. Laughter detection in meetings. In *Proceedings of the NIST Meeting Recognition Workshop*, 2004. 2

[35] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972. 1, 2

[36] K. Laskowski, M. Ostendorf, and T. Schultz. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 148–155, 2008. 6

[37] J. Levine and R. Moreland. Small groups. In D. Gilbert and G. Lindzey, editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press, 1998. 4, 6

[38] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 748–751, 2003. 3

[39] 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. 3, 6

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

[41] L. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 181–188, 2008. 6

[42] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proceedings of Eurospeech*, pages 2079–2082, 1997. 2

[43] S. Mota and R. Picard. Automated posture analysis for detecting learners interest level. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 49–56, 2003. 3

[44] C. Nass and S. Brave. *Wired for speech: How voice activates and advances the Human-Computer relationship*. The MIT Press, 2005. 1, 6

[45] 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. 6

[46] C. Nass and J. Steuer. Computers and social actors. *Human Communication Research*, 19(4):504–527, 1993. 6

[47] 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. 6

[48] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human-centred intelligent human-computer interaction (HCI^2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008. 1, 4

[49] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, 2004. 5

[50] A. Pentland. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005. 5

[51] A. Pentland. Automatic mapping and modeling of human networks. *Physica A*, 378:59–67, 2007. 6

[52] A. Pentland. Social Signal Processing. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007. 1, 5

[53] G. Psathas. *Conversation Analysis - The study of talk-in-interaction*. Sage Publications, 1995. 2, 4

[54] V. Richmond and J. McCroskey. *Nonverbal Behaviors in interpersonal relations*. Allyn and Bacon, 1995. 1, 2

[55] 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. 6

[56] K. Scherer. *Personality markers in speech*. Cambridge University Press, 1979. 2

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

[58] E. Shriberg. Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences*, 1:619–622, 1999. 2

[59] 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. 2

[60] 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. 3

[61] 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. 3, 4

[62] K. Truong and D. Leeuwen. Automatic detection of laughter. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 485–488, 2005. 2

[63] 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. 3

[64] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 301–308, 2001. 2

[65] 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. 6

[66] A. Vinciarelli and S. Favre. Broadcast news story segmentation using social network analysis and hidden markov models. In *Proceedings of the ACM International Conference on Multimedia*, pages 261–264, 2007. 4

[67] A. Vinciarelli and J.-M. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006. 6

[68] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: survey of an emerging domain. *Image and Vision Computing, to appear*, 2009. 1, 3, 5

[69] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM International Conference on Multimedia*, pages 1061–1070, 2008. 1, 3, 5

[70] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: A survey. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 61–68, 2008. 1, 3, 5

[71] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, and R. Stiefelhagen. SMaRT: the Smart Meeting Room task at ISL. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 752–755, 2003. 3

[72] C. Weng, W. Chu, and J. Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009. 6

[73] Y. Wu and T. Huang. Vision-based gesture recognition: A review. In *Proceedings of the International Gesture Workshop*, pages 103–109, 1999. 2

[74] 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. 3

[75] G. Yule. *Pragmatics*. Oxford University Press, 1996. 2, 4

[76] 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. 6

[77] 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*, 31(1):39–58, 2009. 2