Capturing Order in Social Interactions

Alessandro Vinciarelli

Following Aristotle, "*Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human*" (Politika, ca. 328 BC). This is more than an abstract philosophical statement if, twenty five centuries after the great Greek philosopher, domains as diverse as psychology, physiology and neurology, just to mention a few, still investigate how humans are the perfect machines for social interaction: the muscles of our faces are aimed at expressing our subtlest feelings and emotions to others [1], our ears are tuned to human voices more than to any other sound [2], specific brain structures (the mirror neurons) are aimed at imitating and learning from others [3], and the list could continue.

As humans appear to be literally wired for social interaction, it is not surprising to observe that social aspects of human behavior and psychology attract interest in the computing community as well [4][5]. The gap between social animal and unsocial machine was tolerable when computers were nothing else than improved versions of old tools (e.g., word processors replacing typewriters), but nowadays computers go far beyond that simple role. Today, computers are the natural means for a wide spectrum of new, inherently social, activities like remote communication, distance learning, online gaming, social networking, information seeking and sharing, training in virtual worlds, etc. In this new context, computers must integrate human-human interaction as seamlessly as possible and deal effectively with spontaneous social behaviors of their users. In concise terms, computers need to become *socially intelligent* [6].

Such an ambitious plan of filling the social intelligence gap between humans and machines starts from a fundamental problem, namely how to make social phenomena accessible to computers when the only evidence these have at disposition about the world are signals captured with devices like microphones and cameras. The consequent question is: "*Do social phenomena leave physical, machine detectable, traces in signals captured with sensors?*"

One possible answer comes from the findings of human sciences (sociology, anthropology, social psychology, etc.) showing that *social phenomena, while appearing unconstrained and spontaneous, are governed by principles and laws and give rise to ordered and predictable behavioral patterns* [7]. For example, during social interactions, people tend to mirror postures and facial expressions of individuals they like, play with pencils and other little objects when they are uncomfortable, avoid exchanging mutual gaze with people they consider of a superior social level, interrupt others to show disagreement, and give

off many other behavioral cues that have no other function than conveying socially relevant information (see [8] for an extensive monography).

These ordered and predictable patterns allow people to make sense, often unconsciously, of social interactions they both observe and participate in [2]. Patterns that are accessible to eyes and ears are typically detectable through microphones and cameras (or any other suitable sensor) and, once detected, they can be automatically understood in terms of social information they convey. Since one of the most important facets of social intelligence is exactly about understanding of socially relevant behavioral patterns, an automatic approach including both detection and understanding of these patterns can be considered as a form of *artificial social intelligence*.

The rest of this article shows a few examples of how above ideas can be applied to the analysis of social phenomena taking place in conversations. In particular, the examples show how turn-taking patterns, one of the most salient behavioral cues in any conversation, can be analyzed and interpreted in terms of roles that people play, social groups that form around different subjects, and conflict dynamics in competitive discussions. After the examples, the article outlines some of the most promising research directions aimed at artificial social intelligence in computing and signal processing communities.

CAPTURING ORDER IN CONVERSATIONS

Conversation is the most common form of social interaction, one of the most important situations where social intelligence operates to understand, beyond the verbal content of messages being exchanged, the social phenomena at work. Human sciences have extensively investigated conversations and suggest turn-taking as a key evidence of social interaction processes:

> [...] 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*** [9].

In technical terms, the turn-taking is a sequence of pairs S encoding *who talks when and how much*:

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

where N is the number of turns, Δt_i is the length of turn i, and s_i is a participant identifier, with $s_i \in A = \{a_1, \ldots, a_G\}$ (G is the number of conversation participants).

From a machine analysis point of view, turn-taking is appealing because it can be effectively extracted with a large variety of speaker diarization approaches, i.e. techniques aimed at segmenting audio recordings into single speaker intervals. Futhermore, human sciences provide insights about the way social phenomena

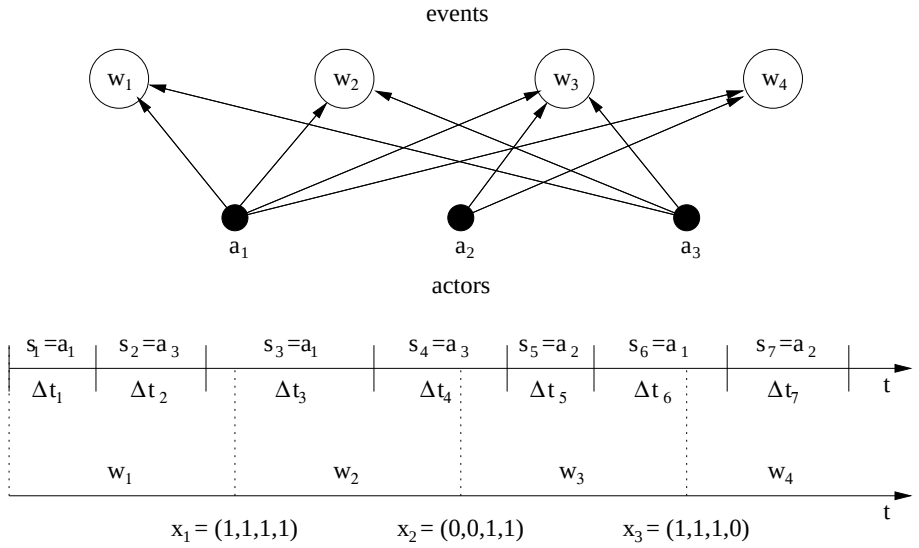

Fig. 1. Extraction of a Social Affiliation Network from the turn-taking. Actors correspond to participants and events to uniform nonoverlapping segments spanning the whole length of the conversation.

shape turn-taking. However, two major questions remain open: does such a simple object as S actually convey enough information about social interactions? Are order and predictability induced by social phenomena robust to speaker diarization errors? The rest of this section shows a few examples where the answer to the above questions is positive.

Role recognition

As they are ubiquitous in everyday life, social interactions take the most diverse forms in terms of settings, goals, contexts, etc. However, there is one aspect that they all have in common, their participants play roles: "*People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their interactions involve behaviors associated with defined statuses and particular roles.*" [7]. This section addresses the problem of automatically recognizing roles in formal settings like news and talk-shows (where roles correspond to functions like *anchorman*, *guest*, *headline person*, etc.), or meetings (where roles correspond to company positions like *project manager*, *industrial designer*, etc.).

Do roles leave traces in turn-taking? Social psychology suggests that conversations involving more than two persons can be thought of as sequences of one-to-one interactions between pairs of participants. Thus, for two individuals, proximity in time of respective turns is likely to account for direct interaction. Such a simple information allows one to build a Social Affiliation Network (SAN) capturing the overall interaction structure of a conversation under exam [10]. If roles actually leave a trace, they are likely to do it in such a structure because a person playing a given role tends to interact only with people playing

TABLE I

ROLE RECOGNITION RESULTS.

setting	size	α	π	α^*	avg G	$\|\mathcal{R}\|$
news	18h 56m	81.2%	0.82	95.3%	12	6
talk-shows	27h 00m	83.9%	0.78	96.5%	30	5
meetings	45h 38m	43.6%	0.99	49.5%	4	4

certain roles and not with others.

A SAN [10] is a graph with two kinds of nodes, *actors* and *events* (see Figure 1). In conversations, actors correspond to participants and events correspond to, as a simple approximation, uniform non-overlapping segments spanning the whole length of the conversation. Actors are linked to an event when they participate in it (in this case when they talk during the corresponding segment). Each actor a_i is represented with a n-tuple $\mathbf{x}_i$, where component x_{ij} accounts for participation of a_i in event w_j. In the simplest case, x_{ij} is set to 1 when a_i participates in event w_j and to 0 otherwise (see lower part of Figure 1).

Such a simple representation has been applied in extensive experiments performed over roughly 90 hours of material including news, talk-shows, and meetings (see all details in [11]). The overall approach includes three different steps, automatic extraction of of turn-taking with an unsupervised diarization approach, extraction of SAN and representation of actors as described above, and mapping of n-tuples $\mathbf{x}_i$ into roles belonging to a predefined set $\mathcal{R}$. If $\mathbf{r}$ is a G-tuple such that r_i is the role of a_i, then the role recognition step can be thought of as finding the G-tuple $\mathbf{r}^*$ satisfying the following equation:

$$\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}^G} p(X,T|\mathbf{r})p(\mathbf{r}), \tag{2}$$

where $\mathcal{R}$ is the set of predefined roles, $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_G\}$ contains the n-tuples representing the participants, and $T = \{\tau_1,\ldots,\tau_G\}$ contains the fractions τ_i of time each actor talks for (see above for the meaning of other symbols). After assuming that $\mathbf{x}_i$ and τ_i are statistically independent given the role and that roles are independent, the above expression boils down to:

$$\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}^G} \prod_{i=1}^{G} p(\mathbf{x}_i|r_i)p(\tau_i|r_i)p(r_i). \tag{3}$$

The term $p(\mathbf{x}_i|r_i)$ is estimated with Bernoulli distributions, $p(\tau_i|r_i)$ with Gaussians, and *a-priori* role probabilities $p(r)$ with the fraction of training set each role accounts for.

Table I reports the results and shows interaction setting, size of the corresponding dataset, overall accuracy α (percentage of time correctly labeled in terms of role), purity π of the speaker diarization

(the closer to 1 the better), accuracy α^* achieved over the groundtruth turn-taking, average number of participants, and cardinality of predefined role set $\mathcal{R}$.

The performances seem to suggest that *roles actually bring order and predictability in turn-taking*. The effect is machine detectable and an automatic approach, based on a simple representation of turn-taking behavior, recognizes roles with a performance significantly higher than chance even in highly spontaneous settings like meetings. The difference between α and α^* shows that, at least in the case of news and talk-shows, errors are mostly due to speaker diarization. However, role related turn-taking patterns are still evident enough to achieve satisfactory performances.

Roles are played individually by each person involved in a given setting. However, other social phenomena can be understood only in terms of *social groups*, subsets of interaction participants that develop mutual bonds tighter than those they have with others. The next example shows how social groups form around the different subjects discussed during a conversation.

Groups and stories

In general, conversations are sequences of *stories*, semantically coherent segments during which participants discuss about a single and specific subject. Whether the sequence is dictated by an agenda or follows a spontaneous evolution, social psychologists have observed that each story involves only a fraction of participants. In other words, each story corresponds not only to a specific subject, but also to a *social group*, a subset of participants characterized by a high degree of mutual interaction (see Figure 2b). This applies in particular when conversations involve a large number of individuals and simultaneous participation of all of them is impractical.

Does the presence of social groups induce order in turn-taking? This question has been addressed through experments performed over 27 hours of talk-shows where people interact spontaneously, but still follow a plan expected to pass through some major predefined topics (see [12] for a full description). The applied approach includes three main steps, the extraction of the turn-taking with an unsupervised diarization approach, the building of a Social Affiliation Network like the one described in the previous section, and the automatic alignment of the sequence of turns (see below for their representation) with a sequence of stories.

The n-tuples $\mathbf{x}$ used for role recognition (see previous section) capture information about groups as well. When people belong to the same social group, they tend to participate in the same events (in this case to talk during the same time intervals), thus to be represented with similar n-tuples. The turn-taking S includes the speaker sequence $\{s_1, \ldots, s_N\}$. This can be converted into a sequence of observations $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $\mathbf{y}_i$ is obtained by applying Principal Components Analysis (PCA) to $\mathbf{x}_i$, the

TABLE II

STORY SEGMENTATION PERFORMANCE IN TERMS OF PURITY.

	variance fraction			
speak. segm.	70%	80%	90%	100%
manual	0.80	0.80	0.80	0.82
automatic	0.74	0.76	0.76	0.77

n-tuple representing the speaker talking at turn i.

If a conversation is actually a story sequence, then Y is the observable evidence of an underlying, hidden, sequence of stories $H = \{h_1, \ldots, h_N\}$ as depicted in Figure 2b. The problem of reconstructing the story sequence, and identifying the corresponding social groups, can be thought of as finding the sequence H^* satisfying the following equation:

$$H^* = \arg \max_{H \in \mathcal{H}_N} p(Y|H)p(H), \tag{4}$$

where $\mathcal{H}_N$ is the set of all possible story sequences of length N. The term $p(Y|H)$ is estimated with a fully connected, ergodic, Hidden Markov Model, and the term $p(H)$ is estimated with a trigram language model:

$$p(H) = \prod_{i=1}^{N} p(h_i|h_{i-1}, h_{i-2}). \tag{5}$$

The goal of $p(H)$ is to ensure that the order of the story is respected, i.e. that story k always follows story $k-1$ and precedes story $k+1$.

Table II reports the results in terms of *purity*, a measure of the coherence between groundtruth and automatic story segmentation (the closer to 1 the better). The results are reported, for both automatically extracted and groundtruth turn-taking, for several amounts of variance retained after applying PCA to n-tuples $\mathbf{x}$. The main stories, those who are sufficiently long to allow the formation of a group, are correctly captured, while others, those that are too short to let a social group to form, are typically missed. However, the performance is satisfactory for browsing applications aimed at bringing a user in correspondence of the main talk-shows stories.

Like in the case of roles, a social phenomenon like group forming results into order and predictability in the turn-taking. Once again, the effect is machine detectable and the story segmentation performance shows that the approach can detect at least the most evident social groups, those that correspond to the stories that have been discussed for more time and thus are likely to be more important. Furthermore, the effect is robust with respect to the errors of the speaker diarization process used to extract the turn-taking from the original data.

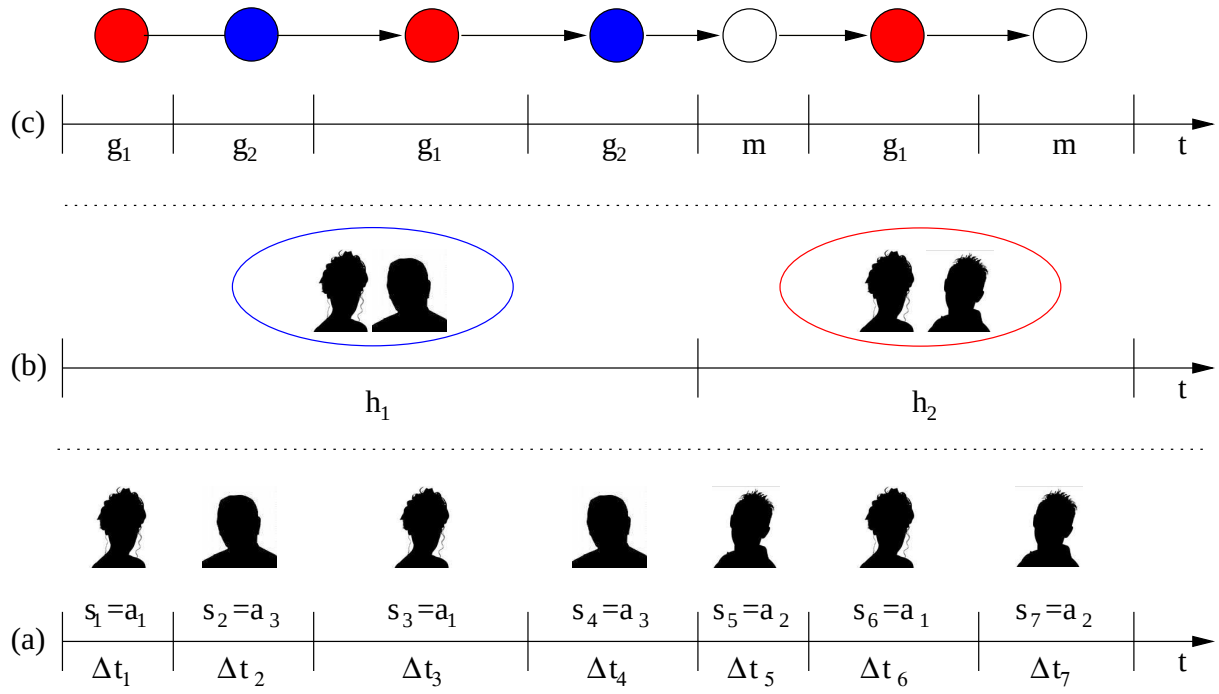

Fig. 2. The turns of figure *a* can be grouped into stories that correspond to social groups (figure *b*), or used to model conflict dynamics with Markov chains (figure *c*).

Order in conflicts

Conflicts are among the most investigated social phenomena as they have a major impact on the life of any group of individuals expected to interact with one another. Whether the group is a professional team working towards a common task, a family addressing the basic needs of its members, or simply a circle of friends sharing their saturday evening, a conflict can jeopardize the welfare of individual members as well as of the group as a whole.

Do conflicts leave machine detectable traces in turn-taking? Whoever has been involved in a heated discussion knows that this is definitely the case. During conflicts, people are prone to break the rules of a normal conversation and do not hesitate to shout, interrupt, speak when others are speaking, etc. What is less evident is that there is an order underlying these behaviors, even if they seem to introduce noise and desorder in the normal flow of non-conflictual interactions. Furthermore, the same ordered and predictable patterns emerge not only when conflicts are hot, but also when they are cold, i.e. when people express their disagreement while still applying the norms of non-conflictual conversations.

Social psychologists have observed that, in the presence of a conflict (hot or cold), people tend to react to someone they disagree with rather than to someone they agree with. This means that the participant talking at turn k is statistically dependent on the participant talking at turn $k-1$ (see Figure 2c). This information can be easily captured with a Markov chain, a probability density function defined over the

space of state sequences $Q = \{q_1, \ldots, q_N\}$, where each q_i belongs to a predefined set $\mathcal{Q}$ of states:

$$p(Q) = p(q_1) \prod_{k=2}^{N} p(q_k|q_{k-1}), \tag{6}$$

$p(q_1)$ is the probability of starting with state q_1, $p(q_k|q_{k-1})$ is the probability of a transition from q_{k-1} to q_k, and N is the number of states in Q.

This simple model has been applied in experiments performed over a dataset of 45 political debates (27 hours and 56 minutes of material in total) built around the conflict between two fronts opposing one another on the issue of the day. Each debate revolves around a central *yes/no* question (e.g., "*are you favorable to new education laws?*") and involves five persons: one moderator, two participants on the *yes* side and two others on the *no* one. The goal of the experiments is to automatically identify the moderator and to reconstruct correctly the two fronts. The applied approach starts with an unsupervised speaker diarization that extracts automatically the turn-taking, then uses a Markov chain to map the resulting sequence of turns into a sequence of states corresponding to the two fronts and to the moderator, i.e. $\mathcal{Q} = \{g_1, g_2, m\}$, like depicted in Figure 2c.

More formally, if $\varphi : A \rightarrow \mathcal{Q}$ is a mapping that associates a participant $s_i \in A$ with a state $q_j \in \mathcal{Q}$, then the problem can be thought of as finding the mapping φ^* satisfying the following expression:

$$\varphi^* = \arg\max_{\varphi \in \mathcal{Q}^A} p(\varphi(s_1)) \prod_{n=2}^{N} p(\varphi(s_k)|\varphi(s_{k-1})). \tag{7}$$

By construction, the probability on the right hand side of Equation (7) has the same value if states g_1 and g_2 are switched. The reason is that g_1 and g_2 are simply meant to distinguish between members of different fronts and not to account for a specific front.

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 actually assigned the same front. This figure goes up to 75% when using the groundtruth speaker segmentation (and not the speaker segmentation automatically extracted from the data). The average performance of an algorithm assigning the states randomly is 6.5% and this means that the simple above model performs ten times better than chance. Thus, conflicts, that seem to be a moment where any social norm is broken, turn out to be a source of order as the other social phenomena described so far.

SOCIAL COMPUTERS FOR THE SOCIAL ANIMAL

So far we have shown how several social phenomena (roles, group forming, and conflicts) leave physical, machine detectable, traces in terms of predictable behavioral patterns. These have been detected in turn-taking (*who talks when and how much*), a phenomenon shaped by social processes in the settings considered for the experiments (talk-shows, news, debates and meetings). The integration of social psychology

into automatic approaches has been shown to be effective and to lead to a form of artificial social intelligence. The works described in the previous section are just examples, but their core idea, to capture order induced by social interactions through integration of human sciences findings, lies at the hart of both Social Computing (SC) [4] and Social Signal Processing (SSP) [5][13], the main domains aimed at bringing social intelligence in computers. The two domains are partially overlapping, but they are complementary under two fundamental respects: the behavioral patterns they investigate, and the scale of the interactions they consider. The rest of this section outlines the main aspects of the two domains and delineates some future research perspectives.

Social Computing

Social Computing focuses on *electronic* or *computer mediated* behaviors [4]. These include actions like credit card payments, cellular phone calls, e-mail exchanges, use of instant messaging, posting of data to social media like Flickr or Youtube, social networking activities through sites like Facebook or Linkedin, e-shopping via web based services like Amazon or eBay, writing blogs, and any other action that can be detected through a large-scale computing infrastructure [14].

Analysis of these behaviors involves hundreds to millions of participants (depending on the cases) that contribute to large-scale collective behavioral patterns. Order emerges through a large number of individual actions and interactions and leads to phenomena like *online communities* that group thousands of people around a subject or a common interest even if none of the members states it explicitly, applications like *recommendation systems* that provide suggestions inferred from the choices of thousands of other individuals showing similar behavioral patterns, technology approaches like *tagging* that learn to describe the data content from the millions of descriptions people spontaneously share on social media, devices like *smart badges for reality mining* that constantly monitor the acivities of their holders and those of the neighboring people to devise common behavioral and interaction patterns, etc. [4][14].

Social Signal Processing

Social Signal Processing is the new, emerging, domain aimed at automatic understanding of social interactions through analysis of nonverbal behavioral patterns [5][15]. Several decades of research in human sciences have shown thet people display *social signals*, i.e. relational attitudes corresponding to their feeling about ongoing interactions and social contexts, in terms of aggregates of nonverbal behavioral cues. Social signals include phenomena like politeness, attention, interest, disagreement, ostracism, hostility, etc. Socially relevant nonverbal patterns include face and eyes behavior (facial expressions, gaze exchanges, etc.), vocal behavior (vocal outbursts, turn-taking, silences and pauses, etc.), gestures and postures (head

movements, body orientation with respect to others, etc.), physical appearance (somatotype, clothes, etc.), and use of space and environment (seating arrangements, interpersonal distances, etc.).

SSP considers small (2 to 4 participants) to medium (5 to 25 participants) scale interactions like those analyzed in the examples of previous sections. The typical social phenomena investigated so far in the SSP community include dominance, social and functional roles, conflicts, group dynamics, interest, engagement, agreement and disagreement, personality, etc. This has led to technologies that predict the outcome of dyadic interactions (salary negotiations, job interviews, customer-operator transactions, etc.), to approaches aimed at detecting symptoms of mental and psychological problems (depression, alzheimer disease, autism, etc.), to systems that automatically extract the content of multimedia material on the basis of the portrayed social interactions, etc. (see [5] for an extensive survey).

Furthermore, since people tend to interact with computers in the same way as they do with other humans, SSP investigates how dynamics of human-human interaction can be applied to Human-Machine interaction as well. This has led to synthetic voices and faces that convey relational attitudes and allow a natural interaction with computers and robots, to data retrieval approaches adapting their results to the attitude of users, etc. (see [2] for a monography on this aspect).

SSP is an inherently multidisciplinary domain as it requires not only a tight collaboration between technology and human sciences, but also the integration of different technological disciplines (e.g., computer vision and speech processing). On one hand, the examples of the previous section clearly show how automatic approaches would not be capable of correctly understanding social phenomena without integrating the findings of human sciences. On the other hand, one of the clearest indications emerging from current SSP state-of-the-art is that, in most cases, social interactions analysis is reliable only if several behavioral cues are analyzed jointly (e.g. facial expressions and accompanying vocalizations) and this typically requires multimodal approaches. The reason is that, individually, nonverbal behavioral patterns are ambiguous and using multiple cues is the only way to improve robustness of understanding approaches.

Future perspectives

In their complementarity, SSP and SC aim at transforming computers into social actors following the same mechanics as humans in natural and spontaneous interactions, whether these take place face-to-face or through computing infrastructures. Both SSP and SC have shown that integration between human sciences and technology is a key towards success and they are ready to continue in this directions despite all the difficulties in establishing a multidisciplinary field [5][14]. Furthermore, both domains have clearly identified order and predictablity as a viable evidence for analysis, synthesis and understanding of social interactions. It is a promising starting point towards the creation of *social computers for the social animal*,

the common long term goal of all the efforts described in this article.

Acknowledgments. This work 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 author wishes to thank Sarah Favre, Fernando Fernandez Martinez and Hugues Salamin for their precious help.

The author. Alessandro Vinciarelli (vincia@idiap.ch) is Senior researcher at Idiap Research Institute (Switzerland). He is IEEE Member since 2006.

REFERENCES

[1] B. Waller, J. Cray, and A. Burrows, "Selection for universal facial emotion.," *Emotion*, vol. 8, no. 3, pp. 435, 2008.

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

[3] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annual Reviews of Neuroscience*, vol. 27, pp. 169–192, 2004.

[4] F.Y. Wang, K.M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 79–83, 2007.

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

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

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

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

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

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

[11] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliation networks and discrete distributions," in *Proceedings of International Conference on Mutlimodal Interfaces*, 2008.

[12] A. Vinciarelli and S. Favre, "Broadcast news story segmentation using Social Network Analysis and Hidden Markov Models," in *Proceedings of ACM International Conference on Multimedia*, 2007, pp. 261–264.

[13] A. Pentland, "Social Signal Processing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007.

[14] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, pp. 721–723, 2009.

[15] A. Pentland, *Honest signals: how they shape our world*, MIT Press, 2008.