Social Signal Processing: Understanding Nonverbal Communication in Social Interactions

Alessandro Vinciarelli
University of Glasgow / Idiap Research Institute
Sir A.Williams Bldg – G12 8QQ Glasgow (UK)
CP592 – 1920 Martigny (Switzerland)
vincia@dcs.gla.ac.uk

Fabio Valente
Idiap Research Institute
CP592 – 1920 Martigny (Switzerland)
fvalente@idiap.ch

ABSTRACT
This paper provides a short overview of Social Signal Processing, the domain aimed at bridging the social intelligence gap between people and machines. The focus of Social Signal Processing is on nonverbal behavioral cues that human sciences (psychology, anthropology, sociology, etc.) have identified as conveying social signals, i.e. relational attitudes towards others and social situations. The rationale is that such cues are the physical, machine detectable and synthesizable evidence of phenomena non-otherwise accessible to computers such as empathy, roles, dominance, personality, (dis-)agreement, interest, etc. After providing a brief state-of-the-art of the domain, the paper outlines its future perspectives and some of its most promising applications.

Author Keywords
Social Signal Processing, human-human communication, nonverbal behavior, social interactions.

ACM Classification Keywords
A.1 INTRODUCTORY AND SURVEY. H.1.2 User/ Machine Systems. I.2 ARTIFICIAL INTELLIGENCE. I.2.10 Vision and Scene Understanding. J.4 SOCIAL AND BEHAVIORAL SCIENCES.

INTRODUCTION
Several decades of research in human sciences have shown that nonverbal communication is the main channel through which we express *social signals* [3], i.e. our relational attitudes (e.g., sympathy, interest, hostility, agreement, etc.) towards others and social situations. Nonverbal communication is the wide spectrum of nonverbal behavioral cues (e.g., facial expressions, vocalizations, postures, gestures, etc.) that we display when we interact with others [10], with machines [8] and with media [11]. From a computing point of view, this is important for two reasons: The first is that nonverbal behavioral cues play the role of a physical, hence machine detectable evidence of social signals. The second is that nonverbal cues synthesized through some form of embodiment (conversational agents, robots, etc.) express the same relational attitudes as when they are displayed by humans [4], thus are likely to synthesize social signals [8].

Social Signal Processing (SSP) relies on the above to bridge the social intelligence gap between humans and machines [13,15]. Social intelligence [2] is the facet of our cognitive abilities that aims at dealing effectively with social interactions and, at its core, it includes two main aspects: The first is the correct interpretation, in terms of social signals, of nonverbal behavioral cues displayed by others. The second is the generation of nonverbal cues expressing social signals appropriate in a given situation. In other words, SSP brings social intelligence in machines via *modeling*, *analysis* and *synthesis* of nonverbal behavior in social interactions [13,15].

Modeling means the investigation of principles and laws underlying human-human interactions, along the same line of research that has identified the importance of nonverbal behavior in social interactions. Analysis means the development of automatic approaches for understanding social signals and social phenomena, mainly based on signal processing and machine intelligence techniques. Synthesis means the automatic generation of social signals under the form of embodied conversational agents, artificial faces, avatars, robots or any other device capable of displaying understandable nonverbal behavioral cues.

Correspondingly, SSP addresses three main research questions:

- Is it possible to detect automatically nonverbal behavioral cues in data captured with suitable sensors (e.g., microphones and cameras)?
- Is it possible to automatically infer social signals from nonverbal behavioral cues detected through sensors?
- Is it possible to synthesize social signals for embodiment of social behaviors in artificial agents,

robots or other devices?

While still in its early and pioneering stages, SSP has attracted significant attention in both scientific and business communities. The SSP state-of-the-art is rich and constantly expands towards new research directions, but the domain is still characterized by high entry barriers, in particular the need of large annotated corpora and software tools covering a wide spectrum of functionalities (e.g., facial expression analysis, prosody extraction, data annotation, etc.). In this respect, a European collaboration called SSPNet (Social Signal Processing Network) is building an extensive online repository (www.sspnet.eu) of articles, data and software tools. The goal is to smooth the entry barriers and allow any potentially interested researcher to start working on SSP [12].

The rest of this paper provides a short survey of the SSP state-of-the-art (in particular when it comes to social interactions understanding), outlines some future perspectives from both scientific and application points of view, and draws some conclusions.

STATE-OF-THE-ART

Extensive surveys of SSP, at least for the analysis component, are available in [13,15]. This section provides an overview of the main themes addressed in modeling and analysis of nonverbal behavior, with a particular attention to turn-taking, the cue that, so far, has led to the most satisfactory results in social behavior understanding.

On the modeling side, current efforts aim at a systematic and rigorous definition of social signals as well as at the identification of behavioral variables to be taken into account in automatic analysis and synthesis of social signals [3]. Furthermore, several works explore the possibility of using computational approaches to validate psychological findings like, e.g., the impact of facial features on the perception of personality traits, or the effect of depression on nonverbal cues. This kind of works is particularly interesting because it closes the loop between human and computing sciences: on one hand, computational approaches integrate human sciences findings to automatically analyze nonverbal behavior, on the other hand, human sciences apply computational approaches to confirm and assess their findings.

On the analysis side, the state-of-the-art concentrates on interactions in small groups, the most common and primordial forms of social exchange [7]. The most extensively addressed problem is the recognition of roles people play in different situations, including radio and television programs, where the setting is highly formal and roles correspond to specific tasks (e.g., *anchorman* or *guest*) [14], and spontaneous meetings, where roles correspond to social functions (e.g., *attacker* or *supporter*) [16].

Automatic role recognition is mostly based on the analysis of turn-taking patterns, i.e. *on who talks when, to whom and how much*. Thus, the first step of the process is typically the application of a speaker clustering approach that segments the audio channel of the interaction recordings into time intervals expected to correspond to an individual voice. In other words, speaker clustering techniques identify *turns*, i.e. time segments during which one person talks and the others listen to her. In the meantime, each turn is automatically assigned a label corresponding to a speaker so that the process not only identifies the points where the speaker changes, but also what are the turns during which each speaker talks.

Speaker clustering techniques are typically based on agglomerative clustering approaches that group vectors of acoustic observations, extracted at regular time steps (typically once every 10 milliseconds), based on their similarity, i.e. on how likely they are to belong to the same voice, and on their temporal proximity, i.e. on how likely they are to belong to the same turn. Agglomerative clustering techniques are iterative approaches where, at each step, the two most similar clusters are merged. The process is continued until a model selection criterion (typically the Bayesian Information Criterion) is met. Two clusters are considered similar when, after having been merged, the fitness of the clustering to the data (typically measured in terms of log likelihood) improves.

Once the speaker clustering has been performed, the actual role recognition step can take place using two main approaches. The first is to represent the turn-taking pattern of each person with a feature vector and then to map this last into one of the roles using a classifier (Support Vector Machines, Neural Networks, Bayesian Networks, etc.). The second is to extract a feature vector from each turn and then to align the resulting sequence of observations with a sequence of roles using probabilistic sequential models (Hidden Markov Models, Dynamic Bayesian Networks, Conditional Random Fields, etc.).

In the first case, the most common features are the number of times a given speaker talks, the fraction of total conversation time a speaker talks for, how many adjacent turns each speaker has with all of the others, what is the centrality of each speaker (i.e. how many times the speaker talks between each pair of two other speakers), etc. In the second case, the most common features aim at capturing sequential aspects such as the number of times a given sequence of speakers is observed, how many turns there are between two consecutive interventions of the same speaker, etc.

There is no evidence about what approach is better, but sequential approaches seem to be more promising because they allow one to assign different roles to the same person in the course of the same interaction. This is an important requirement when considering roles inspired by social

theories (like those proposed by Bales in his works on small groups), more general than those related to specific scenarios considered so far like broadcast data and meetings.

Furthermore, some variants of the most common probabilistic sequential approaches, e.g. Factorial Hidden Markov Models, Influence Models, Layered Hidden Markov Models, Latent Conditional Random Fields, Dynamic Bayesian Networks, etc., allow one to model multiple streams of observations which might correspond to several persons (particularly suitable to investigate how people react to one another) as well as to behavioral cues extracted from multiple modalities (particularly suitable to study how multiple cues concur to convey the same social signal).

Besides roles, other phenomena often investigated include conflict and disagreement [2], given the disruptive impact they can have on the life of a group, and recognition of dominant individuals, given the impact these have on the group outcome [6]. In these cases as well, the turn-taking component plays a major role (see above), but more behavioral cues are taken into account such as facial expressions, gaze behavior and movement. This requires to jointly analyze multiple modalities, a major problem when the behavioral cues expressed in each of the modalities take place at different time-scales like, for example, facial expressions (half a second to a second) and turns (few seconds to some minutes).

Two main approaches have been followed in these cases. The first is called early fusion and simply consists of using a single probabilistic sequential model (see above) fed with observations extracted from multiple modalities. This is based on the assumption, often unrealistic, that the processes taking places in different modalities are lockstep, i.e. that observations resulting from different modalities are always determined by the same hidden, underlying state. The second approach, called late fusion, analyzes processes taking place in different modalities separately and then fuses the output of the models with classifier combination approaches. This approach is based on the assumption, once again unrealistic, that behaviors captured through different modalities are independent.

While being based on patently wrong assumptions, both early and late fusion approaches lead to satisfactory results even if, in general, one modality alone is responsible for most of the performance, the others simply bringing small relative improvements.

FUTURE PERSPECTIVES

The state-of-the-art addresses a long list of social phenomena, but new directions are still emerging in SSP research.

On the short term, research efforts explore aspects of nonverbal behavior that, while having been extensively investigated in human sciences, have been neglected in the computing community. One example is the use of space and environment to express social relational messages [5]. This includes analysis and synthesis of socially relevant information from the spatial configurations (called F-formations) people assume during interactions, the inference of social distance from physical distance, analysis of territoriality, social behavior in surveillance and monitoring scenarios, etc. Other examples are the multimodal generation of spontaneous behavior, or the simulation of subtle behavioral phenomena like mirroring and phonetic convergence.

In both cases, signal processing and machine learning approaches applied so far are not necessarily suitable and major challenges must be faced. In particular, the state-of-the-art has concentrated on small groups (four to six participants), but many important scenarios involve larger numbers of individuals. This is likely to require approaches focusing less on the detailed behavior of each individual and more on collective phenomena. This might lead towards Social Network Analysis like techniques where no prediction is made at the individual level, and analysis is possible only in terms of presence of social groups, detection of prominent individuals (in terms of the number of network paths passing through them), overall connectedness of the network, etc.

The perspectives are rich in terms of potential applications as well. Multimedia indexing is likely to profit from SSP because social interactions are one of the main channels through we access reality and to index the data in terms of social interaction means to make retrieval approaches closer to our perception of data content. Healthcare applications, especially when it comes to mental problems or cognitive deterioration due to ageing and related problems, can apply SSP to identify subtle symptoms in the first stages of illnesses. Human Computer Interaction can be improved thanks to the introduction of socially adept technologies capable of dealing with users like humans deal with other people. Computer mediated communication can profit from SSP by allowing the detection and transmission of those nonverbal cues like gaze that most contribute to the naturalness of face-to-face interactions.

CONCLUSIONS

This paper has provided a short introduction to Social Signal Processing, including an overview of its main principles and goals, a survey of the most important results so far, and some of the most promising research perspectives currently emerging in the community.

Social aspects of human behavior attract attention in many different areas because they seem to provide an explanation for many experimental observations not otherwise understandable. For example, neurosciences have identified

social interaction and learning through imitation as the main goal of mirror neurons, physiology has shown that our ears are tuned to human voices more than to any other sound to maximize the chance of social contacts, some psychology theories explain the existence of stable personality traits as a means to ensure predictability in social exchanges, ethology recognizes social interaction as one of the main reasons behind observable behaviors, and the list could continue.

Computing sciences could not be immune to such a wave of interest. Nowadays, computers are much more than an improved version of old tools (like word processors used to be with respect to typewriters), they are the platform through which we communicate, we entertain ourselves, we shop, we join and form large communities of interest, etc. Furthermore, computers are at the core of technologies expected to seamlessly integrate our everyday life such as smart ambients, robots, smart interfaces and, more in general, human centered technologies [9], i.e. technologies dealing with their users following the natural interacting modes of humans.

As social interactions, and their non-verbal component in particular, are such a natural aspect of our behavior, SSP technologies, centered around non-verbal communication, are likely to bring a major improvement in all of the scenarios outlined above. Last, but not least, computational approaches for analysis and synthesis appear to be an instrument helping human sciences to better understand human behavior. This in turn might further improve SSP and open a cycle where human and computing sciences are not only integrated, but also mutually supporting.

ACKNOWLEDGMENTS

This work has been funded by the EC (FP7/2007-2013) under grant agreement No. 231287 (SSPNet), and by the Swiss National Science Foundation through the National Center of Competence in Research IM2.

REFERENCES

1. Albrecht, K. *Social Intelligence: The new science of success*. John Wiley & Sons Ltd, 2005.
2. Bousmalis, K., Mehu, M. and Pantic, M. Spotting Agreement and Disagreement: A Survey of Nonverbal Audiovisual Cues and Tools. Proceedings of IEEE Workshop on Social Signal Processing, 2009.
3. Brunet, P.M., Donnan, H., McKeown, G., Douglas-Cowie, E. and Cowie, R. Social signal processing: What are the relevant variables? And in what ways do they relate? Proceedings of the *IEEE International Workshop on Social Signal Processing* (2009).
4. Cassell, J. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67-83.
5. Cristani, M., Murino, V. and Vinciarelli, A. Socially Intelligent Surveillance and Monitoring: Analysing Social Dimensions of Physical Space. In *Proceedings of International Workshop on Socially Intelligent Surveillance and Monitoring* (2010), to appear.
6. Hung, H. and Jayagopi, D. and Yeo, C. and Friedland, G. and Ba, S. and Odobez, J.M. and Ramchandran, K. and Mirghafori, N. and Gatica-Perez, D. Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the ACM International Conference on Multimedia* (2007), 838-841.
7. Levine, J. and Moreland, R. Small groups. In *The handbook of social psychology*, D. Gilbert and G. Lindzey (eds.). Oxford University Press (2008), 415-469.
8. Nass, C. and Brave, S. *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, USA, 2005.
9. Pantic, M. and Nijholt, A. and Pentland, A. and Huang, T.S. Human-Centred Intelligent Human Computer Interaction (HCI2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems* 1, 2 (2008), 168-207.
10. Poggi, I. *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler Buchverlag Berlin, 2007.
11. Reeves, B. and Nass, C. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press New York, NY, USA, 1996.
12. Vinciarelli, A., Pantic, M. *sspnet.eu, a web portal for Social Signal Processing*. IEEE Signal Processing Magazine, to appear, (2010).
13. Vinciarelli, A., Pantic, M. and Bourlard, H. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing Journal* 27, 12 (2009), 1743-1759.
14. Vinciarelli, A. Capturing Order in Social Interactions. IEEE Signal Processing Magazine 26, 12 (2009), 133-137.
15. Vinciarelli, A., Pantic, M., Bourlard, H. and Pentland, A. Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM International Conference on Multimedia* (2008), 1061-1070
16. Zancanaro, M., Lepri, B. and Pianesi, F. Automatic detection of group functional roles in face-to-face interactions. *In Proceedings of the International Conference on Multimodal Interfaces*, (2006), 28–31.