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)
{vancia,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 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 is expressed not only by words but by nonverbal behaviors such as facial expression and body posture as well. 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 to the analysis of conflicts in competitive discussions.

1. Introduction

Imagine watching 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 have (e.g., whether they 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 extracting social information from nonverbal communication is hard wired in the human brain [33][54]. Any facial expression, vocal outburst, gesture or posture triggers often unconscious analysis of socially relevant information [4]. 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 [42] generated by computers.

If nonverbal communication plays such an important role in our life, shouldn’t we enable computers to sense and interpret social meaning of human user’s 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 [51][68][69]. The core idea of SSP is that nonverbal behavioral cues can be detected with microphones, cameras, and any other suitable sensors. The cues can then be used as a machine detectable evidence for automatic analysis and understanding of social behavior shown by the human user.

SSP enables Human-Centred computing paradigm [46], effectively dealing with psychological and behavioral responses natural to humans, in contrast to computing-centred paradigm that requires people to operate computers following technology-driven criteria. This will have a major impact on various domains of computing technology such as Human-Computer Interaction which will become more adept to social interactions with users [46], multimedia content analysis which will be analyzed according to the way humans perceive the reality around them [22], computer mediated communication (e.g., see [24]) because transmission will include social cues necessary for establishing a
natural contact with others, and other domains where computers must seamlessly integrate into the life of people.

The paper starts by introducing the most important aspects of nonverbal communication (Section 2). It illustrates the main technological components necessary to analyze social behavior (Section 3) and provides an example showing how SSP principles and ideas are applied to the analysis of conflicts in competitive discussions (Section 4). It also provides a brief survey of the main SSP applications presented so far in the literature. Section 6 concludes the paper.

2. Nonverbal Behavior and Social Signals

Nonverbal communication includes all the messages other than words that people exchange in interactive contexts. In some cases, messages are exchanged consciously, and nonverbal behaviors have a precise meaning attached to them (e.g., the thumbs up gesture). More frequently, nonverbal behavior gives away messages, leaking information about the state of people, e.g. about their emotions, self-confidence, status, etc. [25].

SSP focuses on human nonverbal communication and, in particular, on social signals [3], 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 that 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 occurring in social interactions into five major classes called codes [30]. The first is physical appearance, including not only somatic characteristics, but also clothes and ornaments that people use to modify their appearance. 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., [27]) and to the recognition of clothes for tracking and surveillance purposes (e.g., [15]).

The second code relates to gestures and postures, extensively investigated in human sciences because they are considered the most reliable cue revealing actual attitude of people towards others (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 mouses with hand movements as computer interfaces (see [72] for recent technologies). Gestures and postures have been also analyzed for their affective content (see [28] for a survey). However, there are only a few works presented so far addressing the problem of interpreting gestures and postures in terms of social signals (see [68] 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 somebody’s affective state and intentions on the basis of the shown facial expression [32]. 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 [76] 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 [33][54]: voice quality (prosodic features like pitch, energy and rhythm), linguistic vocalizations (expressions like “ehm”, “ah”, etc.) and non-linguistic vocalizations (laughter, crying, sobbing, etc.), silence (use of pauses), and turn-taking patterns (mechanisms regulating floor exchange) [53][74]. 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 detection, e.g., of disfluencies [58], non-linguistic vocalizations (e.g., particular laughter [52][62]), or rhythm [40], but with the goal of improving the speech recognition performance rather than
analysing social behavior.

The last code relates to space and environment, i.e. the way people share and organize the space 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 [29]. 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 follow, at least partially, the processing chain in the picture (see Section 5).

The first, and crucial, step is the data capture. 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 [36][70]), but the literature reports the use of wearable devices [20] and pressure captors [41] (for recognizing posture of sitting people) as well.

In most cases, the raw data involve recordings of 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 [73], or any other kind of technique that allows one to identify intervals of time or scene regions corresponding to specific individuals.

Person detection is the step preliminary to behavioral cues extraction, i.e. the detection of nonverbal signals displayed by each individual. Some approaches for this stage have been mentioned in Section 2. Extensive overviews are available in [68][69].

The two main challenges in social behavior understanding are the modeling of temporal dynamics and fusing the data extracted from different modalities at different time scales.

Temporal dynamics of social behavioral cues (i.e., their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed social behavior [3][21]. However, relatively few approaches explicitly take into account the temporal evolution of behavioral cues to understand social behavior. 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 [37].

To address the second challenge outlined above (temporal, multimodal data fusion), 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 (see [76] for a survey). 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 behavioral 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 behavior is shown, human behavior understanding is about W5+ (where, what, when, why, how), where the why and how are directly related to recognizing communicative intention including social behaviors, affective and cognitive states of the observed person [47]. 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 [47]), answering the why and how questions in a W4-context-sensitive manner when analyzing human behavior 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., deadlines) or to satisfy members needs (e.g., in families) [35].

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 two parties. The corpus used for the experiments includes 45 debates (roughly 30 hours of material) revolving around yes/no questions 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 [74], 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.

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][74]. Thus, the social signal conveyed as a direct reaction is likely to be disagreement. Hence, the corresponding nonverbal behavioral cue is adjacency in speakers turns. 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 that proposed in [1]. The audio channel of the political debates is converted into a sequence $S$:

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

(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 $\Delta t_i$ of the turn (see Figure 3). Thus, the sequence $S$ contains the entire information about the turn-taking, namely who talks when and how much. The purity (see [67] 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 for the identification of speaker adjacencies representing the target cue based on which agreement and disagreement between debate participants will be detected.
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 $\varphi^*$ satisfying the following expression:

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

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 expression 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 vice versa. 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 simply meant to distinguish between members of different groups.

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 one is used as the test set. The experiment is reiterated and each time a different debate is used as the 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. This figure goes up to 75% when using the truth speaker segmentation and not the speaker segmentation 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.

5. Main SSP Applications

The first extensive surveys of SSP applications have been proposed in [68][69], after the expression Social Signal Processing has been introduced for the first time in [51] to denote 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 [48][49], where a set of cues accounts for activity (the total amount of energy in the speech signals), influence (the statistical influence of one person on the speaking patterns of the others), consistency (the 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 [38][39].

Other approaches, developed in the same period as the above works, have aimed at the analysis of small group interactions [35], with particular emphasis on meetings and broadcast data (talk-shows, news, etc.). Most of the works have focused on recognition of collective actions [17][37], dominance detections [31][55], and role recognition [7][19][23][34][75]. The approaches proposed in these works are often multimodal [17][19][31][37][55][75], 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) [7][34], or for combinations of social networks and lexical features [23].

Social network analysis has been applied as well...
in [65][66][71] to recognize the roles played by people in broadcast 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][50].

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 [42][43][44], 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][45], 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 [9]. Further evidence of the same phenomenon is available in [5][6], 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

The long term goal of SSP is to give computers social intelligence [2]. This is one of the multiple facets of human intelligence, maybe the most important because it helps to deal successfully with the complex web of interactions we are constantly immersed within, whether this means to be recognized as a leader on the workplace, to be a good parent, or to be a person friends like to spend time with. The first successes obtained by SSP are impressive and have attracted the praise of both technology [26] and business [8] communities. However, there is still a long way to go before artificial social intelligence and socially-aware computing become a reality.

Several major issues need to be addressed in this direction. The first is to establish an effective collaboration between human sciences and technology. SSP is inherently multidisciplinary, no effective analysis of social behavior is possible without taking into account the basic laws of human-human interaction that psychologists have been studying for decades. Thus, technology should take into account findings of human sciences, and these should formulate their knowledge in terms suitable for automatic approaches. The second issue is the development of approaches dealing with multiple behavioral cues (typically extracted from several modalities), often evolving at different time scales while still forming a coherent social signal. This is necessary because single cues are intrinsically ambiguous, sometimes they actually convey social meaning, but sometimes they simply respond to contingent factors (e.g., postures can communicate a relational attitude, but also be determined by the search for comfort). Finally, an important issue is the use of real-world data in the experiments. This will lead to more realistic assessments of technology effectiveness and will link research to potential application scenarios.

The strategic importance of the domain is confirmed by several large projects funded at both national and international level around the world. In particular, the European Network of Excellence SSPNet (2009-2014) aims not only at addressing the issues outlined above, but also at fostering research in SSP through the diffusion of knowledge, data and automatic tools via its web portal (www.sspnet.eu). In this sense, the portal is expected to be not only a site delivering information, but also an instrument allowing any interested researcher to enter the domain with an initial effort as limited as possible.

References


