Towards a Technology of Nonverbal Communication: Vocal Behavior in Social and Affective Phenomena

Alessandro Vinciarelli
University of Glasgow - Department of Computing Science
Sir Alwyn Williams Building, Glasgow G12 8QQ, UK
Idiap Research Institute - CP592, 1920 Martigny, Switzerland
Gelareh Mohammadi
Idiap Research Institute - CP592, 1920 Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne - EPFL, 1015 Lausanne, Switzerland

ABSTRACT

Nonverbal communication is the main channel through which we experience inner life of others, including their emotions, feelings, moods, social attitudes, etc. This attracts the interest of the computing community because nonverbal communication is based on cues like facial expressions, vocalizations, gestures, postures, etc. that we can perceive with our senses and can be (and often are) detected, analyzed and synthesized with automatic approaches. In other words, nonverbal communication can be used as a viable interface between computers and some of the most important aspects of human psychology such as emotions and social attitudes. As a result, a new computing domain seems to emerge that we can define "technology of nonverbal communication". This chapter outlines some of the most salient aspects of such a potentially new domain and outlines some of its most important perspectives for the future.

INTRODUCTION

Nonverbal communication is one of the most pervasive phenomena of our everyday life. On one hand, just because we have a body and we are alive, we constantly display a large number of nonverbal behavioral cues like facial expressions, vocalizations, postures, gestures, appearance, etc. (Knapp & Hall, 1972; Richmond & McCroskey, 1995). On the other hand, just because we sense and perceive the cues others display, we cannot avoid interpreting and understanding them (often outside conscious awareness) in terms of feelings, emotions, attitudes, intentions, etc. (Kunda, 1999; Poggi, 2007). Thus, "*We cannot not communicate"* (Watzlawick et al., 1967) even when we sleep and still display our feelings (of which we are unaware) through movements, facial expressions, etc., or when we make it clear that we do not want to communicate:

> If two humans come together it is virtually inevitable that they will communicate something to each other [...] even if they do not speak, messages will pass between them. By their looks, expressions and body movement each will tell the other something, even if it is only, "I don't wish to know you: keep your distance"; "I assure you the feeling is mutual. I'll keep clear if you do".

(Argyle, 1979).

As nonverbal communication is such a salient and ubiquitous aspect of our life, it is not surprising to observe that computing technology, expected to integrate our daily life seamlessly and naturally like no one else, identifies automatic understanding and synthesis of nonverbal communication as a key step towards *human-centered* computers, i.e. computers adept to our natural modes of operating and communicating (Pantic et al., 2007; Pantic et al., 2008). This is the case of *Affective Computing*, where the aim is automatic understanding and synthesis of emotional states (Picard, 2000), of certain trends in *Human-Computer Interaction*, where the goal is to interface machines with the psychology of users (Reeves & Nass, 1996; Nass & Brave, 2005), of research in *Embodied Conversational Agents*, where the goal is to simulate credible human behavior with synthetic characters or robots (Bickmore & Cassell, 2005), and of the emerging field of *Social Signal Processing*, where the target is to understand mutual relational attitudes (*social signals*) of people involved in social interactions (Vinciarelli et al., 2008; Vinciarelli et al., 2009).

This list of domains is by no means complete, but it is sufficient to show how a *nonverbal communication technology* is actually developing in the computing community. Its main strength is an intense cross-fertilization between machine intelligence (e.g., speech processing, computer vision and machine learning) and human sciences (e.g., psychology, anthropology and sociology) and its main targets are artificial forms of social, emotional and affective intelligence (Albrecht, 2005; Goleman, 2005). Furthermore, social and psychological research increasingly relies on technologies related to nonverbal communication to develop insights about human-human interactions, like in the case of large scale social networks (Lazer et al., 2009), organizational behavior (Olguin et al., 2009), and communication in mobile spaces (Raento et al., 2009).

This chapter aims at highlighting the most important aspects of this research trend and includes two main parts. The first introduces the main aspects of nonverbal communication technology and the second shows how this last is applied to the analysis of social and affective phenomena. The first part introduces a general model of human-human communication, proposes a taxonomy of nonverbal behavioral cues that can be used as perceivable stimuli in communication, and outlines the general process that nonverbal communication technology implements. The second part illustrates the most important phenomena taking place during social interactions, provides a survey of works showing how technology deals with them, and proposes the recognition of emotions in speech as a methodological example of the inference of social and affective phenomena from vocal (nonverbal) behavior. The chapter ends with a description of the emerging domain of Social Signal Processing (the most recent research avenue centered on nonverbal communication) and a list of application domains likely to benefit from the technologies described in this chapter.

PSYCHOLOGY AND TECHNOLOGY OF NONVERBAL COMMUNICATION

Nonverbal communication is a particular case of human-human communication where the means used to exchange information consists of nonverbal behavioral cues (Knapp & Hall, 1972; Richmond & McCroskey, 1995). This is appealing from a technological point of view because nonverbal cues must necessarily be accessible to our senses (in particular sight and hearing) and this makes them detectable through microphones, cameras or other suitable sensors, a *conditio sine qua non* for computing technology. Furthermore, many nonverbal behavioral cues are displayed outside conscious awareness and this makes them *honest*, i.e. sincere and reliable indices of different facets of affect (Pentland, 2008). In other words, nonverbal behavioral cues are the physical, machine detectable evidence of affective phenomena not otherwise accessible to experience, an ideal point for technology and human sciences to meet.

The rest of this section outlines the most important aspects of nonverbal communication from both psychological and technological points of view.

Psychology of Nonverbal Communication

In very general terms (Poggi, 2007), communication takes place whenever an *Emitter E* produces a *signal* under the form of a *Perceivable Stimulus PS* and this reaches a *Receiver R* who interprets the signal and extracts an *Information I* from it, not necessarily the one that *E* actually wanted to convey. The emitter, and the same applies to the receiver, is not necessarily an individual person, it can be a group of individuals, a machine, an animal or any other entity capable of generating perceivable stimuli. These include whatever can be perceived by a receiver like sounds, signs, words, chemical traces, handwritten messages, images, etc.

Signals can be classed as either *communicative* or *informative* on one hand, and as either *direct* or *indirect* on the other hand. A signal is said to be communicative when it is produced by an emitter with the intention of conveying a specific meaning, e.g. the "thumb up" to mean "OK", while it is informative when it is emitted unconsciously or without the intention of conveying a specific meaning, e.g. crossing arms during a conversation. In parallel, a signal is said to be direct when its meaning is context independent, e.g. the "thumb up" that means "OK" in any interaction context, and indirect in the opposite case, e.g. crossing arms when used by workers in strike to mean that they refuse to work.

In this framework, the communication is said to be *nonverbal* whenever the perceivable stimuli used as signals are *nonverbal behavioral cues*, i.e. the miriad of observable behaviors that accompany any human-human (and human-machine) interaction and do not involve language and words: facial expressions, blinks, laughter, speech pauses, gestures (conscious and unconscious), postures, body movements, head nods, etc. (Knapp & Hall, 1972; Richmond & McCroskey, 1995). In general, nonverbal communication is particularly interesting when it involves informative behavioral cues. The reason is that these are typically produced outside conscious awareness and can be considered *honest signals* (Pentland, 2008), i.e. signals that leak reliable information about the actual inner state and feelings of people, whether these correspond to emotional states like anger, fear and surprise, general conditions like arousal, calm, and tiredness, or attitudes towards others like empathy, interest, dominance and disappointment (Ambady et al., 2000; Ambady & Rosenthal, 1992).

Social psychology proposes to group all nonverbal behavioral cues into five classes called *codes* (Hecht et al., 1999): *physical appearance*, *gestures and postures*, *face and eyes behavior*, *vocal behavior*, and *space and environment*. Table 1 reports some of the most common nonverbal behavioral cues of each code and shows the social and affective phenomena most closely related to them. By "related" it is meant that the cue accounts for the phenomenon taking place and/or influences the perception of the same phenomenon. The cues listed in this section are the most important, but the list is by no means exhaustive. The interested reader can refer to specialized monographs (Knapp & Hall, 1972; Richmond & McCroskey, 1995) for an extensive survey. In the following, codes and some of their most important cues are described in more detail.

Physical appearance: Aspect, and in particular attractiveness, is a signal that cannot be hidden and has a major impact on the perception of others. After the first pioneering investigations (Dion et al., 1972), a large body of empirical evidence supports the "*Halo effect*", also known as "*What is beautiful is good*", i.e. the tendence to attribute socially desirable characteristics to physically attractive people. This has been measured through the higher success rate of politicians judged attractive by their electors (Surawski & Osso, 2006), through the higher percentage of individuals significantly taller than the average among

CEOs of large companies (Gladwell, 2005), or through the higher likelihood of starting new relationships that attractive people have (Richmond & McCroskey, 1995). Furthermore, there is a clear influence of people somatotype (the overall body shape) on personality traits attribution, e.g., thin people tend to be considered lower in emotional stability, while round persons tend to be considered higher in openness (Cortes & Gatti, 1965).

Gestures and Postures: Gestures are often performed consciously to convey some specific meaning (e.g., the thumb up gesture that means "OK") or to perform a specific action (e.g., to point at something with the index finger), but in many cases they are the result of some affective process and they are displayed outside conscious awareness (Poggi, 2007). This is the case of *adaptors* (self-touching, manipulation of small objects, rhythmic movements of legs, etc.) that typically account for boredom, uncomfort, and other negative feelings, and self-protection gestures like folding arms and crossing legs (Knapp & Hall, 1972; Richmond & McCroskey, 1995). Furthermore, recent studies have shown that gestures express emotions (Coulson, 2004; Stock et al., 2007) and accompany social affective states like shame and embarrassment (Costa et al., 2001; Ekman & Rosenberg, 2005).

Postures are considered among the most reliable and honest nonverbal cues as they are typically assumed unconsciously (Richmond & McCroskey, 1995). Following the seminal work in (Scheflen, 1964), postures convey three main kinds of social messages: *inclusion and exclusion* (we exclude others by orienting our body in the opposite direction with respect to them), *engagement* (we are more involved in an interaction when we are in front of others), and *rapport* (we tend to imitate others posture when they dominate us or when we like them).

	Affective Behaviors							**Tech.**		
Nonverbal cues	emotion	Personality	status	dominance	persuasion	regulation	rapport	Speech analysis	Computer vision	biometry
Physical appearance										
Height			✓	✓					✓	✓
Attractiveness		✓	✓	✓	✓		✓		✓	✓
Body shape		✓		✓					✓	✓
Gesture and posture										
Hand gestures	✓	✓			✓	✓	✓		✓	✓
Posture	✓	✓	✓	✓	✓	✓	✓			
Walking		✓	✓	✓						
Face and eye behavior										
Facial expressions	✓	✓	✓	✓	✓	✓	✓		✓	✓
Gaze behavior	✓	✓	✓	✓	✓	✓	✓		✓	
Focus of attention	✓	✓	✓	✓	✓	✓	✓		✓	
Vocal behavior										
Prosody	✓	✓		✓	✓		✓	✓		
Turn taking	✓	✓	✓	✓		✓	✓	✓		
Vocal outbursts	✓	✓		✓	✓	✓	✓	✓		
Silence	✓		✓				✓	✓		
Space and environment										
Distance	✓	✓	✓		✓		✓		✓	
Seating arrangement				✓	✓		✓		✓	

Table 1. This table shows the most common nonverbal behavioral cues for each code and the affective aspects most commonly related to them. The table has been published in Vinciarelli et al. 2009 and it is courtesy of A.Vinciarelli, M.Pantic and H.Bourlard.

Face and gaze behavior: Not all nonverbal behavioral cues have the same impact on our perception of other's affect and, depending on the context, different cues have different impact (Richmond & McCroskey, 1995). However, facial expressions and, in more general terms, face behaviors are typically the cues that influence most our perception (Grahe & Bernieri, 1999). Nonverbal facial cues account for cognitive states like interest (Cunningham et al., 2004), emotions (Cohn, 2006), psychological states like suicidal depression (Ekman & Rosenberg, 2005) or pain (Williams, 2003), social behaviors like accord and rapport (Ambady & Rosenthal, 1992; Cunningham et al., 2004), personality traits like extraversion and temperament (Ekman & Rosenberg, 2005), and social signals like status, trustworthiness (Ambady & Rosenthal, 1992). Gaze behavior (*who looks at whom and how much*) plays a major role in exchanging the floor during conversations, and in displaying dominance, power and status.

Vocal Behavior: Vocal behavior accounts for all those phenomena that do not include language or verbal content in speech. The vocal nonverbal behavior includes five major components: *prosody*, *linguistic* and *non-linguistic vocalizations*, *silences*, and *turn-taking patterns* (Richmond & McCroskey, 1995). Prosody accounts for *how* something is said and it influences the perception of several personality traits, like competence and persuasiveness (Scherer, 1979). Linguistic vocalizations correspond to sounds like "ehm", "ah-ah", etc. that are used as words even if they are something different. They typically communicate hesitation (Glass et al., 1982) or support towards others speaking. Non-linguistic vocalizations include cry, laughter, shouts, yawns, sobbing, etc. and are typically related to strong emotional states (we cry when we are very happy or particularly sad) or tight social bonds (we laugh to show pleasure of being with someone). Silences and pauses typically express hesitation, cognitive effort (we think about what we are going to say), or the choice of not talking even when asked to do so. Last, but not list, turn-taking, the mechanism through which people exchange the floor in conversations, has been shown to account for roles, preference structures, dominance and status, etc.

Space and Environment: Social and physical space are tightly intertwined and, typically, the distance between two individuals corresponds to the kind of relationship they have, e.g. *intimate* (less than 0.5 meters in western cultures), *casual-personal* (between 0.5 and 1.2 meters) or *socio-formal* (between 1 and 2 meters) following the terminology in (Hall, 1959). Furthermore, the kind of relationship between people sitting around a table influences the seating positions, e.g. people collaborating tend to sit close to one another, while people discussing tend to sit in front of one another (Lott & Sommer, 1967).

Technology of Nonverbal Communication

Is it possible to make technological value out of social psychology findings about nonverbal communication? This is a core question for domains like affective computing (Picard, 2000) and Social Signal Processing (Vinciarell et al., 2009; Vinciarelli, 2009), where nonverbal behavioral cues are used as a physical, *machine detectable* evidence of emotions and social relational attitudes, respectively. Both domains start from the simple consideration that we sense nonverbal behavioral cues (most of the times unconsciously) through our eyes and ears. Thus, it must be possible to sense the same nonverbal cues with cameras, microphones and any other suitable sensor. Furthermore, both domains consider that there is an inference process (in general unconscious) between the behavior we observe and the perceptions we develop in terms of emotional and social phenomena. Thus, automatic inference approaches, mostly based on machine learning, could be used to automatically understand emotional and social phenomena.

Figure 1 shows the main technological components involved in approaches for automatic understanding of nonverbal communication. The scheme does not correspond to any approach in particular, but any work in the literature matches, at least partially, the process depicted in the picture. Furthermore, the scheme illustrated in Figure 1 is not supposed to describe how humans work, but only how machines can understand social and affective phenomena. Overall, the process includes four major steps described in more detail in the rest of this section.

Capture: Human behavior can be sensed with a large variety of devices, including cheap webcams installed on a laptop, fully equipped smart meeting rooms where several tens of microphones and cameras record everything happens (McCowan et al., 2003; Waibel et al., 2003), mobile devices equipped with haptic and proximity sensors (Raento et al., 2009; Murray-Smith, 2009), pressure captors that detect posture and movements (Kapoor et al., 2004), eyefish cameras capturing spontaneous interactions, etc. Capture is a fundamental step because, depending on the sensors, certain kinds of analysis will be possible and others not. However, what is common to all possible sensing devices is that they give as output *signals* and these must be analyzed automatically to complete the whole process.

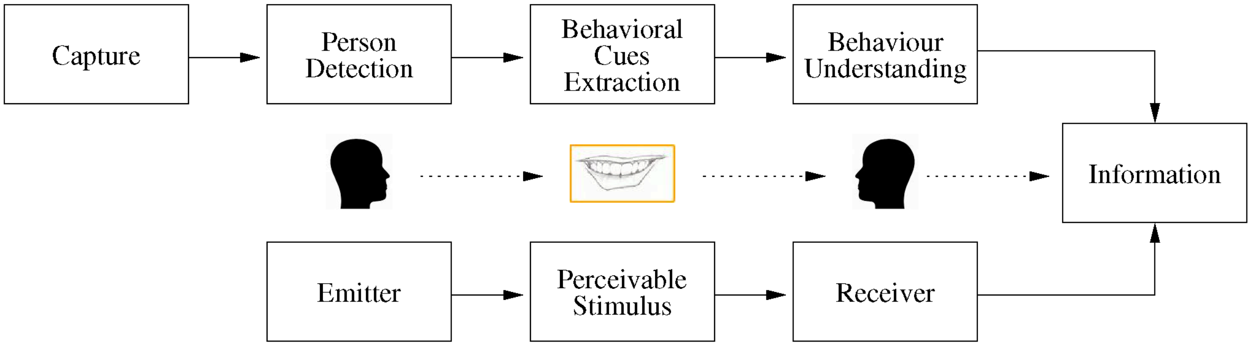

Figure 1. This picture draws a parallel between the communication process as it takes place between humans and as it is typically implemented in a machine. The correspondence does not mean that the process implemented in the machine actually explains and or described a human-human communication process, but simply helps to understand how technology deals with nonverbal communication.

Person detection: In general, signals obtained through capture devices portray more than one person. This is the case, for example, of audio recordings where more than one person talk, of video recodings where different persons interact with one another, etc. This requires a person detection step aimed at identifying what parts of the data correspond to which person. The reason is that nonverbal behavioral cues can be extracted reliably only when it is clear what individual corresponds to a signal under analysis. Person detection includes technologies like speaker diarization, detecting who talks when in audio data (Tranter & Reynolds, 2006), face detection, detecting what part of an image corresponds to the face of one person (Yang et al., 2002), tracking, following one or more persons moving in a video (Forsyth et al., 2006), etc. The application of one person detection technology rather than another one depends on the capture device, but the result is always the same: the signals to be analyzed are segmented into parts corresponding to single individuals.

Behavioral cues detection: The technological components described so far can be considered as a preprocessing phase that gives the raw data a form suitable for actual analysis and understanding of nonverbal communication. Behavioral cues are the perceivable stimuli that, in the communication process, are used by the emitter to convey information and by the receiver to draw information, possibly the same that the emitter wants to communicate. Detection of nonverbal behavioral cues is the first step of the process that actually deals with nonverbal behavior and it includes well developed domains like facial expression recognition (Zeng et al., 2009), prosody extraction (Crystal, 1969), gesture and posture recognition (Mitra & Acharya, 2007), head pose estimation (Murphy-Chutorian & Trivedi, 2009), laughter detection (Truong & Van Leeuwen, 2007), etc. (see Vinciarelli et al., 2009) for an extensive survey on techniques applied at all processing steps). These are the perceivable stimuli that we both produce and sense in our everyday interactions to communicate with others.

Nonverbal behavior understanding: In the communication process, receivers draw information from perceivable stimuli. The information corresponds, in general, to what the emitter actually wants to convey, but this is not necessarily the case. Nonverbal behavior understanding corresponds to this step of the communication process and aims at inferring information like the emotional state or the relational attitude of the receiver from the nonverbal behavioral cues detected at the previous stage of the process. This step of the process relies in general on machine learning and pattern recognition approaches and it is the point where human sciences findings are integrated in technological approaches. Most of the efforts have been dedicated at the recognition of emotions (Picard, 2000) and social signals, i.e. relational attitudes exchanged by people in social interactions (Vinciarelli et al., 2009).

PSYCHOLOGY AND TECHNOLOGY OF FACE-TO-FACE INTERACTIONS

The most natural setting for nonverbal communication is face-to-face interaction, in particular conversations that are considered the "*primordial site of social interaction*" (Schegloff, 1987). As such, conversations are the natural context for a wide spectrum of social phenomena that have a high impact on our life as well as on the life of the groups we belong to (Levine & Moreland, 1998), whether these are work teams expected to accomplish some complex collaborative tasks, circles of friends trying to organize an entertaining Saturday evening, or families aimed at supporting the well being of their members.

This section focuses in particular on those social phenomena that have been not only investigated from a psychological point of view, but that have been the subject of technological research as well (Vinciarelli et al., 2008, 2009).

Psychology of Face-to-Face Interactions

Three main social phenomena recognized as fundamental by psychologists have been addressed by computer scientists as well, namely *roles*, *dominance* and *conflict* (or *disagreement*). This section provides a description of each one of them.

Roles are a universal aspect of human-human interaction (Tischler, 1990), whenever people convene to interact, they play roles with the (unconscious) goal of fulfilling others expectations (if you are the head of a group you are expected to provide guidance towards the fulfillment of group goals), give meaning to their behaviors (helping a patient as a doctor is a professional duty while helping the same patient as a family member is a form of love and attachment), and provide predictability to other interactants (when teachers enter their classroom it is likely they will give a lecture and this helps students to behave accordingly). Some roles correspond to explicit functions (like the examples given above) and can be easily identified and formalized, while others are more implicit and embody deeper aspects of human-human interaction like the *attacker*, the *defender* or the *gate-keeper* in theories of human interactions (Bales, 1950). From a behavioral point of view, roles corresponding to explicit functions tend to induce more regular and detectable behavioral patterns than others and are thus easier to be analyzed automatically (Salamin et al., 2009).

Conflict and disagreement are among the most investigated social phenomena because their impact on the life of a group is significant and potentially disruptive. In some cases, conflicts foster innovation and enhance group performance, but in most cases they have a contrary effect and can lead to the dissolution of the group (Levine & Moreland, 1998). From a social point of view, the most salient aspects of conflicts and disagreement are activities of some of the members that have negative effects on others, attempts of increasing power shares at the expense of others, bargaining between members and formation of coalitions (Levine & Moreland, 1990). In terms of nonverbal behavior, conflicts are typically associated with interruptions, higher fidgeting and voice loudness typical of anger, pragmatic preference structures such that people tend to react to those they disagree with rather than to those they agree with (Bilmes, 1988; Vinciarelli, 2009), longer periods of overlapping speech, etc.

Dominance accounts for ability to influence others, control available resources, and have higher impact on the life of a group, whatever its goal is. Dominance can be interpreted as a personality trait (the predisposition to dominate others), or as a description of relationships between group members (Rienks & Heylen, 2006). While being a hypothetical construct (it cannot be observed directly), dominance gives rise to a number of nonverbal behavioral cues that allow observers to agree on who is (or are) the dominant individuals in a given group. These include seating in positions allowing direct observation of others like the shortest side of a rectangular table (Lott & Sommer, 1967), being looked at by others more than looking at others (Dovidio & Ellyson, 1982), talking longer than others (Mast, 2002), etc.

Technology of Face-to-Face Interactions

Given the centrality of small group interactions in psychology research, it is not surprising to observe that computing technology efforts aimed at the analysis of social and affective phenomena have focused on face-to-face interaction scenarios like meetings, talk-shows, job interviews, etc. (Vinciarelli et al., 2008, 2009). This section proposes a brief survey of the most important approaches dedicated to this problem in the literature, with particular attention to those dealing with role recognition, conflict and disagreement analysis, and dominance detection, i.e. those dealing with the social phenomena identified above as among the most important ones from a social psychology point of view. Table 2 reports results and some of the experimental characteristics of the works surveyed in this section.

Role recognition is typically based on automatic analysis of speaking activity, the physical, machine detectable aspect of behavior that seems to be more correlated with the roles people play in a conversation. By speaking activity is meant here the simple act of speaking or remaining silent, the use of certain words rather than others, the tendency to speak while others are speaking, the number and length of turns during a conversation, etc. Temporal proximity of different speakers interventions is used in (Vinciarelli, 2007; Salamin et al., 2009) to build social networks and represent each person with a feature vector. This is then fed to Bayesian classifiers mapping individuals into roles belonging to a predefined set. A similar approach is used in several other works (Barzilay et al., 2000; Liu, 2006; Garg et al., 2008; Favre et al., 2009) in combination with approaches for the modeling of lexical choices like the BoosTexter (Barzilay et al., 2000) or the Support Vector Machines (Garg et al., 2008). Probabilistic sequential approaches are applied to sequences of feature vectors extracted from individual conversation turns in (Liu, 2006; Favre et al., 2009), namely Maximum Entropy Classifiers and Hidden Markov Models, respectively. An approach based on C4.5 decision trees and empirical features (number of speaker changes, number of speakers talking in a given time interval, number of overlapping speech intervals, etc.) is proposed in (Banerjee & Rudnicky, 2004). A similar approach is proposed in (Laskowski et al., 2008), where the features are probability of starting speaking when everybody is silent or when someone else is speaking, and role recognition is performed with a Bayesian classifier based on Gaussian distributions. The only approaches including features non related to speaking activity are presented in (Zancanaro et al., 2006; Dong et al., 2007), where fidgeting is used as an evidence of role. However, the results seem to confirm that speaking activity features are more effective.

Conflict and disagreement analysis is a domain attracting increasingly significant interest in the last years (Bousmalis et al., 2009). Like in the case of roles, behavior evidences based on speaking activity seem to account reliably for conflict, agreement and disagreement, though psychology insists on the importance of facial expressions, head nods and bodily movements (Poggi, 2007). The coalitions forming during television debates are reconstructed in (Vinciarelli, 2009) through a Markov model keeping into account that people tend to react to someone they disagree with more than to someone they agree with. Similarly, pairs of talk spurts (short turns) are first modeled in terms of lexical (which words are uttered), durational (length, overlapping, etc.), and structural (spurts per speaker, spurts between two speakers, etc.) features and then classified as expressions of agreement or disagreement with a Maximum Entropy Model in (Hillard et al., 2003; Galley et al., 2004).

Dominance detection is one of the most extensively investigated problems in machine analysis of human behavior (Vinciarelli et al., 2009). In contrast with the other two problems considered above, speaking activity features are here accompanied by other nonverbal behavioral cues as well. This happens in (Otsuka et al., 2005), where Dynamic Bayesian Networks are used to model speaking based features (see description of role recognition) and gaze behavior (who looks at whom). Another multimodal approach (Jayagopi et al., 2009) combines speaking activity features with movement based cues (e.g., time during which a person moves, number of time intervals during which a person moves, etc.). In both approaches, movement and gaze help, but speaking features still seem to be the most effective. This seems to be confirmed by other works that achieve god results by using only speaking activity features (Rienks et al., 2006; Rienks & Heylen, 2006).

Article	Data	Performance
Role Recognition		
Salamin et al. (2009)	Broadcast+AMI (90h)	80% frame accuracy
Laskowski et al. (2008)	AMI (45h)	53% frame accuracy
Garg et al.(2008)	AMI (45h)	67.9% frame accuracy
Dong et al. (2007)	MSC (4h.30m)	75% role assignment accuracy
Liu (2006)	Broadcast (17h)	77.0% story accuracy
Banerjee & Rudnicky (2004)	Meetings (45m)	53.0% analysis segments accuracy
Barzilay et al. (2000)	Broadcast (17h)	80.0% story accuracy
Dominance Detection		
Jayagopi et al. (2009)	AMI subset (5h)	80% dominant person recognition rate
Rienks & Heylen (2006)	AMI and M4 subset (95m)	75% dominance level recognition rate
Rienks et al. (2006)	AMI-40	70% dominance level recognition rate
Otsuka et al. (2005)	Broadcast (17h)	N/A
Analysis of (Dis-) Agreement		
Vinciarelli (2009)	Canal9 (43h)	66% (dis-)agreement recognition rate
Hillard et al. (2003)	ICSI subset (8094 talk spurts)	78% (dis-)agreement recognition rate
Galley et al. (2004)	ICSI subset	86.9% (dis-)agreement recognition rate

Table 2. This table presents the main works where nonverbal behavioral cues have been used, in different contexts, to interpret automatically social interactions. Whenever possible, the table reports the amount of data used in the experiments and a performance measure.

A METHODOLOGICAL EXAMPLE: EMOTION RECOGNITION IN SPEECH

The above survey has shown that, from an automatic behavior analysis point of view, cues extracted from speech tend to be more effective than cues extracted from other communication modalities (gestures, movement, facial expressions, etc.). This is in contrast with the results of psychology experiments showing that vocal cues, while having a major impact on the perception of social and emotional phenomena, still have less influence than other behavioral cues, in particular facial expressions and gaze. The most likely reason is that speech analysis techniques are more robust than other technologies (e.g. facial expression analysis and gesture recognition) with respect to conditions in naturalistic interaction settings (e.g., people assuming unconstrained positions, moving, occluding one another, etc.). In other words, machines represent a bottleneck through which some cues, effective when perceived by humans, become difficult to detect and interpret.

For this reason, this section proposes the recognition of emotions in speech as a methodological example of technology of nonverbal communication. The domain has been investigated for long time in both psychology and computing science and the results, if not conclusive, still have a high degree of maturity.

	Joy	Boredom	Neutral	Sadness	Anger	Fear	Surprise	Stress	Depression	Happiness	Disgust	Annoyance	Frustration	Anxiety	Dislike
Pitch	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intensity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rhythm	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Formants		✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓
Cross sectional Areas								✓							
MFCC		✓	✓	✓	✓	✓	✓	✓		✓					✓
LFPC				✓	✓	✓	✓			✓					✓
LPC	✓		✓	✓	✓	✓	✓				✓				
Spectral-band Intensity			✓	✓	✓	✓		✓	✓	✓	✓				
Cepstral Coefficients												✓	✓		
Voice Quality Parameters		✓	✓	✓	✓					✓				✓	

Table 3. Paralinguistic features, used in recognition of different emotional states

Emotion and Vocal Behavior

Generally, the term "emotion" describes subjective feelings lasting for short periods of time, as well as mental and physiological states associated with a wide variety of feelings. No definitive taxonomy of emotions exists, though numerous taxonomies have been proposed; the two commonly used models are called *discrete* and *dimensional*. Most studies in the field of vocal effects of emotion have used the discrete model that groups all emotional states into few discrete categories: happiness, sadness, anger, fear, disgust and surprise. In the dimensional model, different emotional states are mapped in a two-or three-dimensional space. The two main dimensions are *valence* and *activity*; the third dimension is often power or control.

Darwin believed that the voice is the primary channel for expressing emotion in both humans and animals (Knapp & Hall, 1997). Studies of content free speech have shown that emotion perception depends on changes in pitch, speaking rate, volume and other paralinguistic characteristics of voice (Scherer, 2003). Davitn, as cited in (Knapp & Hall, 1997), said in 1964 that "Regardless of the technique used, all studies of adults thus far reported in the literature agree that emotional meanings can be communicated accurately by vocal expression". Significant efforts have been made to identify the vocal cues actually carrying emotional information (Scherer, 2003) and the result is that there is no "dictionary" of emotions in terms of paralinguistic cues, i.e. there is no one-to-one correspondence between observed cues and emotions being expressed. Furthermore, there are several factors interfering with vocal cues including verbal aspects of communication, culture dependency and the rest of nonverbal behavior; however, it has been possible to show which vocal features tend to be associated with which emotions (Knapp & Hall, 1997; Polzin & Waibel, 2000; Scherer, 2003). This is evident in Table 3, where the cues most commonly used in automatic emotion recognition are reported with the respective emotions they tend to be associated with.

Emotion Recognition in Speech

There is a long list of paralinguistic features employed for the recognition of emotions in speech (Knapp & Hall, 1997; Morrison et al., 2007; Scherer, 2003; Ververidis & Kotropoulos, 2006) and they can be categorized into four main groups:

1. **Prosodic Features:** these are features which are reflecting rhythm, stress and intonation of speech. In acoustic terms, the three main classes of prosodic features are pitch, energy, and rhythm (Morrison et al., 2007). Rhythm is represented by various features like, number of pauses, ZCR (Zero-cross ratio which represents the number of times that the speech signal touches the level zero), speech rate (SR), voiced-segment length and unvoiced-segment length.

2. **Spectral-based features:** These are features accounting for the speech signal behavior in the frequency domain. MFCC (Mel-frequency Cepsteral Coefficients), LPC (Linear Prediction Coefficients), LPCC (Linear Prediction Cepstral Coefficients), PLP (Perceptual Linear Prediction), LFPC (Log Frequency Power Coefficients) and Energy in spectral bands are the most commonly applied spectral-based features, typicallly used in speech and emotion recognition (Oudeyer, 2003; Nwe et al., 2001; Nwe et al., 2003; Ververidis & Kotropoulos, 2006;Womack & Hansen, 1996).

3. **Articulatory-based features:** these are features measuring the changes in shape and structure of vocal tract during articulation, e.g., formants, which are a representation of the vocal tract resonance, and cross-section areas when the vocal tract is modeled as a series of concatenated lossless tubes (Ververidis & Kotropoulos, 2006; Womack & Hansen, 1996).

4. **Voice quality features:** Voice quality is the outcome of the human voice excitation; voice quality is the result of glottal pulse shape, its rate and time variations. In particular, voice quality features describe the properties of the glottal source. Jitter, shimmer (Li et al., 2007) and VQP (voice quality parameters) (Lugger & Yang, 2006) are used in emotion recognition as a representation of voice quality.

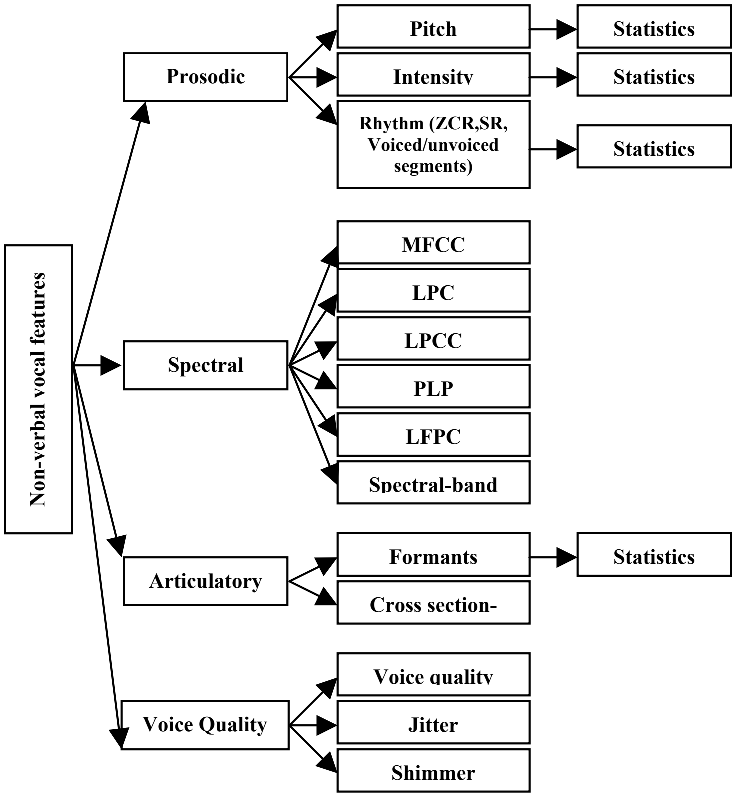

Figure 2. Feature groups

The features in these four main groups are called "*primary*", they are estimated on a frame basis and they reflect short-term characteristics of vocal behavior. Other features, called "*secondary*", are obtained from

Study Groups	Emotional States	Feature Groups	Recognition Method	Accuracy (%)
Womack, & Hansen (1996)	Different stress conditions	Pitch, Rhythm, Cross Sectional Areas, Mel-based	Back-propagation Neural Network	91
Nicholson et al. (1999)	Joy, Neutral, Sadness, Anger, Fear, Surprise, Disgust, Teasing	Pitch, Intensity, LPC, Delta-LPC	OCO Neural Network ACO Neural Network Learning Vector Quantization	50 55 33
Amir et al., (2000)	Joy, Sadness, Anger, Fear, Disgust	Pitch, Intensity, Rhythm	Fuzzy Classifier	43.4
France et al. (2000)	Control, Depression, Suicidal Risk	Pitch, Intensity, Formants, spectral-band Intensity	Quadratic Classifier	70 (F), 77(M)
Polzin, & Waibel (2000)	Neutral, Sadness, Anger	Pitch, Intensity, Jitter, MFCC, + Verbal features	GMM Classifier	62
McGilloway et al. (2000)	Happiness, Neutral, Sadness, Anger, Fear	Pitch, Intensity, Spectral-band Intensity	Gaussian SVM Linear Discriminant	52 55
Lee et al. (2001)	Negative from Non-negative Emotions	Pitch, Intensity	KNN	80 (F), 76 (M)
Zhou et al. (2001)	Different stress conditions	Spectral-ban Intensity	HMM Classifier	91
New et al. (2001)	Happiness, Sadness, Anger, Fear, Surprise, Dislike	LFPC, Mel-based	HMM Classifier	78
Park et al. (2002)	Neutral, Anger, Surprise, Laugh	Pitch	Recurrent Neural Network	Not Reported
Ang et al. (2002)	Annoyance+Frustration vs. Else	Pitch, Intensity, Rhythm, Cepstral Coef.	Decision Tree	83

Oudeyer (2003)	**Calm, Sadness, Anger**	**Pitch, Intensity, Spectral-band Intensity, MFCC**	**KNN Classifier Decision Tree (C4.5) Decision Rules/PART Kernel Density Kstar Linear Regression LWR Voted Perceptrons SVM VFI M5Prime Naive Bayes AdaBoost**	**90 93 94 90 86 85 89 75 94 84 92 91 95**
Kwon et al. (2003)	**Stress, Neutral / Happiness, Boredom, Neutral, Sadness, Anger**	**Pitch, Intensity, Rhythm, Formants, MFCC, Spectral-band Energy**	**Gaussian SVM**	**42.3**
Ververidis (2004)	**Happiness, Neutral, Sadness, Anger, Surprise**	**Pitch, Intensity, Formants**	**Bayes Classifier**	**51.6**
Bhatti et al., (2004)	**Happiness, Sadness, Anger, Fear, Dislike, Surprise**	**Pitch, Intensity, Rhythm**	**A standard Neural Network KNN Classifier Modular Neural Network**	**80.69 79.31 83.31**
Schuller et al. (2005)	**Joy, Neutral, Sadness, Anger, Fear, Disgust, Surprise**	**Pitch, Intensity, Rhythm, Spectral-band Intensity**	**StackingC**	**71.6**
Hyun et al. (2005)	**Joy, Neutral, Sadness, Anger**	**Pitch, Intensity, Rhythm**	**Bayes Classifier**	**71.1**
Shami, & Kamel (2005)	**Approval, Attention, Prohibition Weak, Soothing, Neutral**	**Pitch, Intensity, Rhythm, MFCC**	**KNN SVM**	**87 83**
Morrison et al. (2007)	**Happiness, Sadness, Anger, Fear, Dislike, Surprise**	**Pitch, Intensity, Rhythm, Formants**	**StackingC Unweighted Vote**	**72.18 70.54**
Lugger & Yang (2008)	**Happiness, Boredom, Neutral, Sadness, Anger, Anxiety**	**Pitch, Intensity, Rhythm, Formants, MFCC, VQ Parameters**	**Bayes Classifier**	**88.6**
Yang, & Lugger (2009)	**Happiness, Boredom, Neutral, Sadness, Anger, Anxiety**	**Pitch, Intensity, Rhythm, Formants, Length, Harmony(derived from pitch), Voice Quality**	**Bayes Classifier**	**73.5**
Rong et al. (2009)	**Happiness, Sadness, Anger, Fear**	**Pitch, Intensity, Rhythm, MFCC**	**Decision Tree + Random Forest**	**72.25**
Busso et al. (2009)	**Neutral, Emotional**	**Pitch**	**Distance from Neutral speech model (GMM)**	**70**

Table 4. Previous studies in emotional speech recognition

primary features by, e.g., estimating their average, median, minimum, maximum, range, variance, mean, contour, and, in more general terms, by extracting statistical information about the variation of primary features over a time interval.

Identifying exactly what features account for what emotion is not evident, however, Table 3 shows the paralinguistic features most commonly used for recognizing each emotion. Furthermore, Table 4 provides the most important results in terms of emotion recognition obtained so far in the literature.

Pitch, intensity and rhythm appear to be the most effective paralinguistic features in emotion recognition (Morrison et al., 2007). Pitch is the perceived fundamental frequency of voice and it is the rate of vibration of vocal folds (see Figure 3).

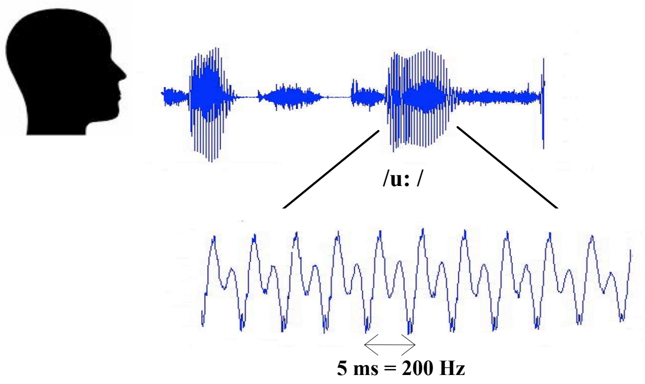

Figure 3. Pitch Estimation

The pitch contour has been shown to vary depending on emotional state (Ang et al., 2002; Lee et al., 2001; Mozziconacci & Hermes, 1995; Park et al., 2002; Polzin & Waibel, 2000). For example, neutral speech has narrower pitch range than emotional speech, i.e. people tend to change more their pitch while emotionally affected. This is particularly evident in angry speech that shows higher pitch range (Scherer, 1996), as well as increased mean pitch and intensity with respect to neutral speech (Scherer, 2003). The same can be observed for happiness while fear was discovered to have a high pitch median, wide pitch range, medium inflection range, and a moderate rate variation. Emotions characterized by lower arousal levels like sadness and disgust typically have lower physiological activation levels and this is evident in speech as well, e.g., sadness results into lower pitch mean and narrower pitch range. Disgust generally has a low pitch median, wide pitch range, lower inflectional range, and lower rate of pitch change during inflection (Morrison et al., 2007; Rong et al., 2009; Scherer, 2003; Ververidis & Kotropoulos, 2006). Table 5 shows a summary of how different emotions affect pitch.

	Pitch		
	Mean	**Range**	**Variance**
Anger	>>	>	>>
Boredom		<	<
Disgust	<	$>_M, <_F$	
Fear	>>	>	>
Joy	>	>	>
Sadness	<	<	<
Stress	>		
Surprise	>=	>	

Table 5. Emotional states and pitch. The symbols show whether the pitch increases (> and >>), decreases (< and <<) or remain unchanged for a given emotion (=). The same symbols are used in the following tables 6 and 7.

Voice intensity is another important paralinguistic feature that can be used as an emotional marker (McGilloway et al., 2000; Polzin & Waibel, 2000; Scherer, 2003). Voice intensity contour provides useful information to discriminate sets of emotions; For instance angry speech is shown to have a significant increase in energy envelope in contrast to sadness and disgust that typically lead to a decrease in intensity. Happiness is shown to have on intensity roughly the same affects as anger. Scherer (2003) notes that anger, fear and joy determine an increase in high frequency intensity while sadness determines, over the same parameter, a decrease. In summary, emotions with high excitation levels such as happiness, anger and surprise have typically higher intensity whereas sadness, fear and disgust have lower intensity (Rong et al., 2009; Ververidis & Kotropoulos, 2006). Table 6 is an abstract of intensity changes affected by different emotional states.

	Intensity		
	Mean	Range	High-freq. mean
Anger	$>>_{M},>_{F}$	>	>
Disgust	<		
Fear	>=		>
Joy	>	>	>=
Sadness	<	<	<
Stress	>		

Table 6. Emotional states and Intensity (see caption of Table 5 for an explanation of symbols).

Rhythm-based characteristics of speech can be used as another index in motion recognition (Ang et al., 2002; Kwon et al., 2003; Schuller et al., 2005; Shami & Verhelst, 2007; Womack & Hansen, 1996; Yang & Lugger, 2009). Rhythm-based characteristics include length of voice/unvoiced segments, pauses between them, Zero Cross Ratio and speech rate.

In several studies it has been shown that speaking rate is higher in anger while on the other side it is lower in sadness; it is also noted that in sadness speech contains “irregular pauses” (Morrison et al., 2007). Table 7 reviews speech rate variation for different emotions.

	Rhythm
	Speech Rate
Anger	><
Boredom	<
Disgust	>
Fear	><
Joy	>
Sadness	<
Surprise	=

Table 7. Emotional states and changes in speech rate (see caption of Table 5 for an explanation of symbols).

Formants are among effective features in emotion recognition (France et al., 2000; Morrison et al., 2007; Kwon et al., 2003; Yang & Lugger, 2009). Formants are resonances of the human vocal tract. The frequency of resonance depends upon the shape and physical dimensions of vocal tract. Under different emotional states the length and width of vocal tract changes. It has been shown in previous studies that during anger, vowels are produced “with a more open vocal tract” and from that they concluded that the first formant frequency is higher in mean than natural speech. Neutral speech usually has a “uniform formant structure and glottal vibration pattern” which is in contrast with formant contours of sadness, fear and anger. Scherer (2003) found that in happiness first formant (F1) mean is decreased but the bandwidth of F1 is increased while in sadness, fear and anger it is opposite (Scherer, 2003), Table 8.

	Formants			
	F1		F2	
	Mean	Bandwidth	Mean	Bandwidth
Anger	>		><	
Disgust	>	<	<	
Fear	>	<	<	
Joy	<	>		
Sadness	>	<	<	

Table 8. Emotional states and Formants (see caption of Table 5 for an explanation of symbols).

CONCLUSIONS AND FUTURE PERSPECTIVES

The main reason why computing science is interested in Nonverbal Communication is that this represents an ideal interface between machines and some of the most important aspects of human psychology, in particular emotions and social attitudes. In this respect, automatic analysis, synthesis and modeling of nonverbal behavior concur towards human-computer confluence and have the potential for bridging the emotional and social intelligence gap between humans and machines. As a result, there are at least two major computer communities working towards technology of nonverbal communication, i.e. affective computing (Picard 2000), dealing with emotions, and Social Signal Processing (Vinciarelli et al., 2008, 2009), dealing with social relational attitudes. Affective Computing is a well established domain and it has been extensively investigated for at least one decade. In contrast, Social Signal Processing is an emerging domain that aims at automatically understanding and synthesizing the social relational attitudes that people exchange during social interactions.

At its core, Social Signal Processing aims at answering three main questions:

- Is it possible to detect nonverbal behavioral cues in recordings of social interactions captured with microphones, cameras and other kinds of sensors?
- Is it possible to infer social signals from nonverbal behavioral cues as detected from data captured with different kinds of sensors?
- Is it possible to synthesize nonverbal behavioral cues eliciting desired social perceptions?

The first two questions pertain to the problem of analyzing social behavior and the involved technological components are those depicted in Figure 1 and described in the section about psychology and technology of nonverbal behavior. The third question concerns the problem of embodiment of social behavior in Artificial Agents, Robots, Avatars, Artificial Characters and any other manufact supposed to simulate human behavior in human-machine interactions. While being in its early stages, Social Signal Processing has attracted significant attention in the business community for its potential impact on organizational sciences (Buchanan, 2007). Furthermore, major efforts are being made towards the creation of a publicly available web-based repository hosting the most important resources necessary to work on SSP, i.e. publications, benchmarks and tools (www.sspnet.eu).

Some issues still remain open and should be addressed in the next years. The first is the effect of context. It has been observed in a previous section that signals can be context dependent (or indirect). In current technology, the context is typically never modeled and the meaning of a given nonverbal cue can be misunderstood. On the other hand, the context is difficult to define and it is unclear how to address its modeling. Furthermore, social and affective aspects of behavior tend to be considered separately (SSP and Affective Computing are typically presented as separate domains) while they overlap to a certain extent and common aspects could be used to reinforce domain technologies.

Several application domains are likely to take significant profit from the development of technology of nonverbal communication (the list is not exhaustive): *multimedia content analysis* can rely on techniques for automatic understanding of social and affective phenomena to enrich the description of multimedia data content. This is particularly important because people and their interactions are among the most important cues we use to access reality and indexing the data in terms of social and affective phenomena is expected to bring retrieval systems closer to human needs (Dumais et al., 2003). *Computer mediated communication* (e.g., videoconferencing) will benefit significantly from the transmission of nonverbal cues (which includes both automatic understanding and synthesis of nonverbal behavior) as their lack seems to be one of the main sources of unnaturalness in mediated communication like, e.g., the lack of gaze contact in videoconferences (Crowley, 2006). In a similar vein, *communication in mobile spaces* can benefit from the use of devices like gyroscopes and haptic sensors capable of stimulating natural nonverbal phenomena like mimicry and coordination (Murray-Smith, 2009). Early detection of cognitive and mental problems can be performed by identifying problems in nonverbal communication (e.g., lack of gestures accompanying speech or unnatural delays in reacting to others in conversation), thus automated systems for analysis of nonverbal communication can help in *healthcare*, particularly for aging related diseases like Alzheimer and Parkinson. *Videogames* have significantly increased their degree of interactivity in the last years and a better understanding of users via their nonverbal behaviors, as well as characters more convincing in the naturalness of their behaviors is likely to further improve gaming experience. *Marketing* is likely to benefit from the automatic analysis of customers behavior in retail spaces as well as from the identification of nonverbal cues capable of establishing a trust relationships between customers and sellers (Ambady et al., 2006). Furthermore, new application domains are likely to emerge like the development of tools for supporting and enhancing human-human communication, the creation of technologies for helping workers using communication (e.g., teachers) in their job. Last, but not least, the use of automatic approaches is likely to help psychologists to make their observations way more extensive and objective with respect to current standards mostly based on observation in the laboratory. These are just few examples, but many more can be identified by looking at how pervasive and ubiquitous computers are becoming in our everyday life.

ACKNOWLEDGEMENTS

This work is supported by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet). Furthermore, it is supported by the Swiss National Science Foundation through the Indo-Swiss Joint Research Project ``Cross-Cultural Personality Perception" (Grant 122936), and through the National Centre of Competence in Research on Interactive Multimodal Information Management (IM2).

REFERENCES:

Albrecht, K. (2005). *Social intelligence: The new science of success*. John Wiley & Sons Ltd.

Ambady, N., Krabbenhoft, M., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluatesales effectiveness. *Journal of Consumer Psychology,* 16(1):4-13.

Ambady, N., Bernieri, F., & Richeson, J. (2000). Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In M. Zanna (Ed.), *Advances in experimental social psychology* (p. 201-272).

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

Amir, N., Ron, S., & Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. In *Proceedings of the ISCA Workshop on Speech & Emotion* (pp. 29-33).

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of ICSLP'02* (pp.2037-2040).

Argyle, M., & Trower, P. (1979). *Person to person: ways of communicating*. HarperCollins Publishers.

Bales, R. (1950). *Interaction process analysis: A method for the study of small groups*. Addison-Wesley.

Banerjee, S., & Rudnicky, A. (2004). 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* (pp. 221-231).

Barzilay, R., Collins, M., Hirschberg, J., & Whittaker, S. (2000). The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of the 17th national conference on artificial intelligence* (pp. 679-684).

Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *Proceedings of the International Symposium on Circuits & Systems, ISCAS'04,* (Vol. 2, pp. II-181-4).

Bickmore, T., & Cassell, J. (2005). Social dialogue with embodied conversational agents. In J. van Kuppevelt, L. Dybkjaer, & N. Bernsen (Eds.), *Advances in natural, multimodal, dialogue systems* (p. 23-54). New York: Kluwer.

Bilmes, J. (1988). The concept of preference in conversation analysis. *Language in Society, 17* (2), 161-181.

Bousmalis, K., Mehu, M., & Pantic, M. (2009). Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *Proceedings of the international conference on affective computing and intelligent interaction* (Vol. II, pp. 121-129).

Buchanan, M. (2007). The science of subtle signals. Strategy+Business, 48, 68–77.

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transaction on Audio, Speech and Language Processing, 17*(4), 582-596.

Cohn, J. (2006). Foundations of human computing: facial expression and emotion. In *Proceedings of the ACM international conference on multimodal interfaces* (pp. 233-238).

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

Costa, M., Dinsbach, W., Manstead, A., & Bitti, P. (2001). Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior, 25* (4), 225-240.

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior, 28* (2), 117-139.

Crowley, J.L. (2006). Social Perception. *ACM Queue*, 4(6):43-48.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Cunningham, D., Kleiner, M., Bültho, H., & Wallraven, C. (2004). The components of conversational facial expressions. *Proceedings of the Symposium on Applied Perception in Graphics and Visualization* (pp. 143-150).

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24* (3), 285-290.

Dong, W., Lepri, B., Cappelletti, A., Pentland, A., Pianesi, F., & Zancanaro, M. (2007). Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on multimodal interfaces* (pp. 271-278).

Dovidio, J., & Ellyson, S. (1982). Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly, 45* (2), 106-113.

Dumais, S., Cutrell, E., Cadiz, J.J., Jancke, G., Sarin, R. & Robbins, D.C. (2003). Stuff I've seen: a system for personal information retrieval and re-use. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 72-79).

Ekman, P., & Rosenberg, E. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs)*. Oxford University Press.

Favre, S., Dielmann, A., & Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *Proceedings of ACM international conference on multimedia*.

France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., & Wilkes, D.M. (2000). Acoustical Properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering, 47*(7), 829-837.

Forsyth, D., Arikan, O., Ikemoto, L., O'Brien, J., & Ramanan, D. (2006). Computational studies of human motion part 1: Tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision, 1* (2), 77-254.

Galley, M., McKeown, K., Hirschberg, J., & Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: use of Bayesian Networks to model pragmatic dependencies. In *Proceedings of meeting of the association for computational linguistics* (pp. 669-676).

Garg, N., Favre, S., Salamin, H., Hakkani-Tur, D., & Vinciarelli, A. (2008). Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In *Proceedings of the acm international conference on multimedia* (pp. 693-696).

Gladwell, M. (2005). *Blink: The power of thinking without thinking*. Little Brown & Company.

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

Goleman, D. (2005). *Emotional intelligence*. Random House Publishing Group.

Grahe, J., & Bernieri, F. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23* (4), 253-269.

Hall, E. (1959). *The silent language*. Doubleday.

Hecht, M., De Vito, J., & Guerrero, L. (1999). Perspectives on nonverbal communication-codes, functions, and contexts. In L. Guerrero, J. De Vito, & M. Hecht (Eds.),
The nonverbal communication reader - classic and contemporary readings (p. 3-18). Waveland Press.

Hillard, D., Ostendorf, M., & Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the north American chapter of the association for computational linguistics - human language technologies conference*.

Hyun, H.K., Kim, E.H., & Kwak, Y.K. (2005). Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept. *IEEE International Workshop on Robots and Human Interactive Communication*, ROMAN 2005, (pp. 312-316).

Jayagopi, D., Hung, H., Yeo, C., & Gatica-Perez, D. (2009). Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing, 17* (3), 501-513.

Kapoor, A., Picard, R., & Ivanov, Y. (2004). Probabilistic combination of multiple modalities to detect interest. In *Proceedings of the international conference on pattern recognition* (pp. 969-972).

Knapp, M., & Hall, J. (1972). *Nonverbal communication in human interaction*. Harcourt Brace College Publishers.

Kwon, O.W., Chan, K., Hao, J., & Lee, T.W. (2003). Emotion Recognition by Speech Signals. In *Proceedings of International Conference EUROSPEECH* (pp. 125-128).

Kunda, Z. (1999). *Social cognition*. MIT Press.

Laskowski, K., Ostendorf, M., & Schultz, T. (2008). Modeling vocal interaction for text independent participant characterization in multi-party conversation. In *Proceedings of the 9th isca/acl sigdial workshop on discourse and dialogue* (pp. 148-155).

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science, 323*, 721-723.

Lee, C.M., Narayanan, S., & Pieraccini, R. (2001). Recognition of negative emotions from the speech signals. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 240-243).

Levine, J., & Moreland, R. (1990). Progress in small roup research. *Annual Reviews Psychology, 41*, 585-634.

Levine, J., & Moreland, R. (1998). Small groups. In D. Gilbert & G. Lindzey (Eds.), *The handbook of social psychology,* (Vol. 2, pp. 415-469). Oxford University Press.

Li, X., Tao, J., Johanson, M.T., Soltis, J., Savage, A., Leong, K.M., & Newman, J.D. (2007). Stress and emotion classification using jitter and shimmer features. In *Proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing, ICASSP2007* (Vol. 4, pp. IV-1081-4).

Liu, Y. (2006). Initial study on automatic identication of speaker role in broadcast news speech. In *Proceedings of the human language technology conference of the naacl, companion volume: Short papers* (pp. 81-84).

Lott, D., & Sommer, R. (1967). Seating arrangements and status. *Journal of Personality and Social Psychology, 7* (1), 90-95.

Lugger, M., & Yang, B. (2006). Classification of different speaking groups by means of voice quality parameters*. ITG-Sprach-Kommunikation*.

Lugger, M., & Yang, B. (2008). Cascaded emotion classification via psychological emotions using a large set of voice quality parameters. In *Proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing, ICASSP2008* (pp. 4945-4948.

Mast, M. (2002). Dominance as expressed and inferred through speaking time: A metaanalysis. *Human Communication Research, 28* (3), 420-450.

McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G., Monay, F., Moore, D., Wellner, P., & Bourlard, H. (2003). Modeling human interaction in meetings. In *Proceedings of IEEE international conference on acoustics, speech and signal processing* (pp. 748-751).

McGilloway, S., Cowie, R., & Douglas-Cowie, E. (2000). Approaching automatic recognition of emotion from voice: a rough benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 207-212).

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37* (3), 311-324.

Morrison, D., Wang, R., & De Silva, L.C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication, 49*, 98-112.

Mozziconacci, S.J.L., & Hermes, D.J. (1995). A study of intonation patterns in speech expressing emotion or attitude: production and perception. In *Proceedings of 13th International Congress of Phonetic Sciences (ICPh'95)* (Vol. 3, pp. 178-181).

Murphy-Chutorian, E., & Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31* (4), 607-626.

Murray-Smith, R. (2009). Empowering people rather than connecting them. *International Journal of Mobile HCI,* to appear.

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

Nicholson, J., Takahashi, K., & Nakatsu, R. (1999). Emotion recognition in speech using neural networks. In *Proceedings of ICONIP'99, 6th International Conference on Neural Information Processing* (Vol. 2, pp. 495-501).

Nwe, T.L., Foo, S.W., & De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*, 603-623.

Nwe, T.L., Wei, F.S., & De Silva, L.C. (2001). Speech based emotion classification. In *Proceedings of IEEE Region 10 International Conference on Electrical & Electronic Technology* (Vol. 1, pp. 291-301).

Olguin Olguin, D., Waber, B., Kim, T., Mohan, A., Koji, A., & Pentland, A. (2009). Sensible organizations: technology and methodology for automatically measuring organizational behavior*. IEEE Transactions on Systems, Man abd Cybernetics – Part B, 39* (1), 43-55.

Otsuka, K., Takemae, Y., & Yamato, J. (2005). A probabilistic inference of multiparty conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of ACM international conference on multimodal interfaces* (pp. 191-198).

Oudeyer, P.Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, *59*, 157-183.

Pantic, M., Nijholt, A., Pentland, A., & Huang, T. (2008). Human-Centred Intelligent Human-Computer Interaction (HCI2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems, 1* (2), 168-187.

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2007). Human computing and machine understanding of human behavior: A survey. In *Lecture notes in articial intelligence* (Vol. 4451, p. 47-71). Springer Verlag.

Park, C.H., Lee, D.W., & Sim, K.B. (2002). Emotion recognition of speech based on RNN. In *Proceedings of International Conference on Machine Learning & Cybernetics(ICMLC'02)* (Vol. 4, pp.2210-2213).

Pentland, A. (2008). *Honest signals: how they shape our world*. MIT Press.

Picard, R. (2000). *Affective computing.* Cambridge (MA), USA: The MIT Press.

Poggi, I. (2007). *Mind, hands, face and body: A goal and belief view of multimodal communication*. Weidler Buchverlag Berlin.

Polzin, T.S., & Waibel, A. (2000). Emotion-sensetive human-computer interfaces. In *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 201-206).

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: an emerging tool for social scientists. *Sociological Methods & Research, 37* (3), 426.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York (USA): Cambridge University Press New York, NY, USA.

Richmond, V., & McCroskey, J. (1995). *Nonverbal behaviors in interpersonal relations*. Allyn and Bacon.

Rienks, R., & Heylen, D. (2006). Dominance Detection in Meetings Using Easily Obtainable Features. In *Lecture notes in computer science* (Vol. 3869, p. 76-86). Springer.

Rienks, R., Zhang, D., & Gatica-Perez, D. (2006). Detection and application of in fluence rankings in small group meetings. In *Proceedings of the international conference on multimodal interfaces* (pp. 257-264).

Rong, J., Li, G., & Chen, Y.P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Management, 45*, 315-328.

Salamin, H., Favre, S., & Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: Using social aliation networks for feature extraction. *IEEE Transactions on Multimedia, 11* (7), 1373-1380.

Scheflen, A. (1964). The significance of posture in communication systems. *Psychiatry, 27*, 316-331.

Schegloff, E. (1987). Single episodes of interaction: an exercise in conversation analysis. *Social Psychology Quarterly, 50* (2), 101-114.

Scherer, K. (1979). *Personality markers in speech*. Cambridge University Press.

Scherer, K. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40*, 227-256.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of IEEE International Conference on Multimedia & Expo (ICME'05).*

Shami, M.T., & Kamel, M.S. (2005). Segment-based approach to the recognition of emotions in speech. *IEEE International Conference on Multimedia & Expo (ICME'05)*, Amsterdam, The Netherlands.

Shami, M.T., & Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication, 49*, 201-212.

Stock, J. Van den, Righart, R., & Gelder, B. de. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion, 7* (3), 487-494.

Surawski, M., & Osso, E. (2006). The eects of physical and vocal attractiveness on impression formation of politicians. *Current Psychology - Developmental - Learning - Personality - Social, 25* (1), 15-27.

Tischler, H. (1990). *Introduction to sociology*. Harcourt Brace College Publishers.

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

Truong, K., & Van Leeuwen, D. (2007). Automatic discrimination between laughter and speech. *Speech Communication, 49* (2), 144-158.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: resources, features, and methods. *Speech Communication, 48*, 1162-1181.

Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing* (Vol. 1, pp. I-593-6).

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

Vinciarelli, A. (2009). Capturing order in social interactions. *IEEE Signal Processing Magazine, 26* (5), 133-137.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal, 27* (12), 1743-1759.

Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008). Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM international conference on multimedia* (pp. 1061-1070).

Waibel, A., Schultz, T., Bett, M., Denecke, M., Malkin, R., Rogina, I., & Stiefelhagen, R. (2003). SMaRT: the Smart Meeting Room task at ISL. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing* (pp. 752-755).

Watzlawick, P., Beavin, J., & Jackson, D. (1967). *The pragmatics of human communication*. New York: Norton.

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

Womack, B., & Hansen, J.L.H. (1996). Classification of speech under stress using target driven features. *Speech Communication, 20*, 131-150.

Yang, B., & Lugger, M. (2009). Emotion recognition from speech signals using new harmony features. Article in Press, *Signal Processing* (pp. 1-9).

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

Zancanaro, M., Lepri, B., & Pianesi, F. (2006). Automatic detection of group functional roles in face to face interactions. In *Proceedings of international conference on multimodal interfaces* (pp. 47-54).

Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31* (1), 39-58.

Zhou, G., Hansen, J.H.L., & Kaiser, J.F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transaction on Speech and Audio Processing, 9* (3), 201-216.

KEY TERMS & DEFINITIONS

Social Signal Processing. Domain aimed at modeling, analysis and synthesis of nonverbal behavior in social interactions.

Affective Computing. Domain aimed at modeling, analysis and synthesis of human emotions.

Nonverbal Communication. Form of communication based on nonverbal behavioral cues (facial expressions, vocalizations, gestures, postures, etc.)

Vocal behavior. Ensemble of speech phenomena that do not include words or language (pauses, laughter, fillers, prosody, rhythm, intensity, etc.).

Social interactions. Every form of interaction including at least two persons modifying their behavior accordingly to others.

Emotion. Subjective feeling lasting for short periods of time, as well as mental and physiological state associated with a wide variety of feelings.

Paralinguistic features. Features extracted from speech data that account for nonverbal vocal behavior.