Cognitive Computation manuscript No.
(will be inserted by the editor)

Open Challenges in Modeling, Analysis and Synthesis of Human Behaviour in Human-Human and Human-Machine Interactions

Alessandro Vinciarelli[1] · **Anna Esposito**[2] · **Elisabeth André**[3] · **Francesca Bonin**[4] · **Mohamed Chetouani**[5] · **Jeffrey F. Cohn**[6] · **Marco Cristani**[7] · **Ferdinand Fuhrmann**[8] · **Elmer Gilmartin**[4] · **Zakia Hammal**[9] · **Dirk Heylen**[10] · **Rene Kaiser**[8] · **Maria Koutsombogera**[11] · **Alexandros Potamianos**[12] · **Steve Renals**[13] · **Giuseppe Riccardi**[14] · **Albert Ali Salah**[15]

Received: date / Accepted: date

Abstract Modeling, analysis and synthesis of behaviour are the subject of major efforts in computing science, especially when it comes to technologies that make sense of human-human and human-machine interactions. This article outlines some of the most important issues that still need to be addressed to ensure substantial progress in the field, namely 1) development and adoption of virtuous data collection and sharing practices, 2) shift of the focus of interest from individuals to dyads and groups, 3) endowment of artificial agents with internal representations of users and context, 4) modeling of cognitive and semantic processes underlying social behaviour, and 5) identification of application domains and strategies for moving from laboratory to the real-world products.

Keywords Human Behaviour; Social Interactions; Virtuous Data Practices; Multimodal Embodiment; Cognitive Modeling; Semantic Processing; Roadmap to Application.

1 Introduction

Modeling, analysis and synthesis of human behaviour are the subject of major efforts in computing science [128, 130]. In principle, the problem can be addressed in purely technological terms, i.e. by applying the same methodologies and approaches that can be used for any other type of data accessible to machines. For example, speech has been analyzed using methodologies that can be applied to any other signal and, similarly, computer vision has addressed the problem of tracking people using the same methodologies that can be used to track any other moving object. Furthermore, robotics and computer graphics addressed the synthesis of human motion by simply reproducing its observable aspects.

However, human behaviour is governed by cognitive, social and psychological phenomena that, while not being observable, must be taken into account to build technologies more robust, effective and human-centered. The first attempts in this direction were done in the early nineties, when automatic analysis and synthesis of facial expressions were addressed for the first time not only in terms of observable facial muscles activity, but also in terms of emotion expression [41]. Interdisciplinary collaboration between computing on one side and, on the other side, psychology and cognitive sciences proved to be a crucial and fruitful milestone.

Nowadays, domains like Affective Computing [93], Social Signal Processing (SSP) [128], Social Robotics [23], Intelligent Virtual Agents [26], Human Communication Dynamics [80], etc. are well established and have a well delimited and recognized scope in the computing community. Recent technological achievements include, e.g., social robots that deal with autistic children [109], computers that make sense of human personality in various contexts [127], artificial agents that sustain emotionally rich conversations with their users [113], approaches that detect phenomena as subtle as mimicry [37], and the list could continue.

However, modeling, analysis and synthesis of human behaviour are far from being solved problems. This article outlines a few major issues that need to be addressed to substantially improve the current state-of-the-art:

– *Virtuous practices for design, collection and distribution of data*. Without data it is difficult, if not impossible, to develop technologies revolving around human behaviour. However, widely shared practices for making

[1]University of Glasgow (United Kingdom), [2]Second University of Naples (Italy), [3]University of Augsburg (Germany), [4]Trinity College Dublin (Ireland), [5]University Pierre et Marie Curie Paris (France), [6]University of Pittsburgh (USA), [7]University of Verona (Italy), [8]Joanneum Research (Austria), [9]Robotics Institute, Carnegie Mellon University, [10]University of Twente (Netherlands), [11]Institute for Language and Speech Processing (Greece), [12]National Technical University of Athens (Greece), [13]University of Edinburgh (United Kingdom), [14]University of Trento (Italy), [15]Bogazici University (Turkey).

of data an asset for the entire community are still missing (see Section 2) [19, 65].
- *From individuals to interaction*. A group of interacting people is more than the mere sum of its members. However, most current analysis approaches still focus on individuals. Furthermore, methodologies addressing groups as a whole, especially when it comes to mutual influence processes, are still at an early stage of development (see Section 3) [27, 54].
- *From shallow to deep interactions*. Human-human interactions take place in highly specific contexts where people typically have a long history of previous relations. However, current artificial agents typically miss an internal representation of both context and others, resulting into shallow interactions with their users. Attempts to go beyond such a state-of-affairs are still limited (see Section 4) [7, 108].
- *Integration of semantic and cognitive aspects*. Social life is determined to a large extent by unconscious, cognitive processes. However, most current approaches for analysis and synthesis of human behaviour do not try to model how people make sense of others and give meaning to their experiences (see Section 5) [32, 96].
- *Applications*. Real-world applications are the ultimate testbed for any technology expected to interact with humans. However, only a relatively few domains are seriously planning the adoption of technologies dealing with human behaviour (see Section 6) [61, 99, 100, 124].

The rest of the article describes each of the above issues in details.

2 The Data

Corpora and data collections are a necessary prerequisite for modeling, analysis and synthesis of human behaviour. In fact, analysis and synthesis are not possible without learning from data showing contexts and phenomena of interest. Furthermore, efficiency in experiments and replicability of results are difficult without a framework for comprehensive and easily interoperable data annotation and analysis. In other words, the multimodal research community cannot progress without virtuous data collection, annotation and sharing practices that make high-quality data accessible and easy to process. This section outlines the challenges arising at various stages of corpus design, collection, annotation, curation and distribution, and proposes strategies that should underpin the best practices.

2.1 Data Design and Collection

Collections of data portraying multimodal interaction behaviours cover a wide spectrum of verbal, nonverbal, social and communicative phenomena. However, most current resources do not address all aspects of social interactions, but focus on the investigation of specific contexts and settings. The probable reason is that the range of spoken interactions, or "*speech-exchange systems*" [106], humans engage in is enormous. It is an open question whether basic mechanisms such as turn-taking or the temporal distribution of cues such as back-channel, gestures or disfluencies vary with the type of interaction. In other words, it is not sure whether observations made over certain data generalize to other data as well and, if yes, to what extent. This requires one to carefully consider what is the genre of the corpus at the design stage.

Corpora may consist of audiovisual material gathered from conventional media (radio and television) and the web, or recordings made during laboratory experiments, possibly using advanced sensors (e.g., High Definition cameras, gaze trackers, microphone arrays, RGB Depth cameras like the Kinect, physiological sensors, etc.). Overall, the large number of settings and data acquisition approaches reflects the wide variety of design and research goals that data collections are functional to [128].

When interactions are recorded in a laboratory setting, the most common way to ensure that people actually engage in social exchanges is to use tasks aimed at eliciting conversation. Typical cases include the description of routes on a map (e.g., the HCRC Map Task Corpus [5]), spotting differences between similar pictures (e.g., the London UCL Clear Speech in Interaction Database [12] and the Wildcat Corpus [125]), participation in real or simulated professional meetings (e.g., the ICSI Meeting Corpus [60] and the AMI Corpus [77]), etc. This way of collecting data produces useful corpora of non-scripted dialogues. However, it is unclear whether the motivation of subjects involved in an artificial task can be considered genuine. Therefore, it is not sure that the resulting corpora can be used to make reliable generalizations about natural conversations [68, 85].

Attempts to collect data in more naturalistic settings have focused initially on real-world phone calls like, e.g., the suicide hotline and emergency line conversations described in [106]. A broadest domain of topics is available in the corpora of real phone conversations (e.g., Switchboard [47] and ESP-C [25]). The main drawbacks of these resources are that they are unimodal and, furthermore, it is not clear whether phone mediated and face-to-face conversations can be considered equivalent. The effort of capturing face-to-face, real-life spoken interactions has led to collection of corpora like Santa Barbara [39], ICE [49], BNC [16] and Gothenburg Corpus [1]. However the effectiveness of these collections is still limited by unimodality (the only exception is the Gothenburg Corpus) and relatively low quality of the recordings.

What emerges from the above is that the collection of data suitable for multimodal research entails a trade-off be-

tween the pursuit of real-life, naturalistic resources and the need of high quality material suitable for automatic processing. This typically leads to the choice of laboratory settings where the sensing apparatus is as unobtrusive as possible and the scenario is carefully designed to avoid biases. This led to hybrid multimodal corpora showing encounters recorded in the laboratory, but without pre-scribed task or subject of discussion imposed on participants. These include collections of free-talk meetings, or first encounters between strangers (e.g., the Swedish Spontal [40], the NOMCO and MOMCO Danish and Maltese corpora [88], casual conversations between acquaintances and strangers [86], etc.). Some of the latest corpora include physiological signals, motion capture information (e.g., DANS and Spontal [40]), and breathing data.

The availability of new sensors, capable of capturing information non accessible in previous corpora, make old data less useful due to sparsity of the type of signals collected (many are audio only) and the impossibility of investigating the range of interconnected signals and cues of interest to current researchers. This issue could serve as a caution to current data collectors. It would be very useful if researchers future-proofed corpora by gathering a range of signals as wide as possible at the data collection stage, hopefully slowing the onset of data obsolescence.

2.2 Data Annotation, Curation and Distribution

Creating recordings is becoming increasingly cheaper and easier, but annotating them in view of modeling, analysis and synthesis of social behaviour remains a time-consuming and labour-intensive task. In fact, enriching data with descriptive and semantic information is usually done manually. Recent advances in sensing technologies have introduced flexibility in automatically collecting features of interest enabling the creation of datasets rich with information on multimodal behaviour that can be further augmented with manual encodings. However, analysis and modelling of multimodal interaction is hampered by the lack of a comprehensive annotation scheme or taxonomy incorporating speech, gestures, and other multimodal interaction features.

Many spoken dialogue annotation schemes are based on speech/dialogue acts and their function in updating dialogue state [24][31][67]. The ISO 24617-2 standard for functional dialogue annotation [57] comprehensively covers information transfer and dialogue control/interaction management functions of utterances, but coverage of social or interactional functions is restricted to "*social obligation management*" (salutations, self-introduction, apologizing, thanking, and valedictions). There is also a need to include annotation of multimodal cues. The MUMIN scheme [2] allows coding of multimodal aspects of dialogue, particularly in terms of their contribution to interaction management and turn-taking, but has not yet been integrated into larger dialogue taxonomies. An important advantage of the ISO standard and indeed of the information state update paradigm [20] is its multi-dimensionality, whereby a "*markable*" or "*area of interest*" can be tagged in several orthogonal ways. This scheme may thus be extensible to account for many interactional and multimodal aspects of interaction. A more extensive taxonomy of communicative acts encompassing various modalities is highly desirable.

While many databases are publicly available, many others are still not shared. The shortage of desirable annotated data is also due to lack of standardisation, Intellectual Property Rights (IPR) restrictions, and privacy issues arising from research ethics. Datasets involved in tasks related to human behaviour analysis come with strict terms of use. Data providers should thus ensure that data reuse is permitted through a set of appropriate licensing conditions. More importantly, datasets should be indexed so that all interested parties are able to identify different types of resources they wish to access and/or acquire. The multidisciplinarity of the field also calls for true and continuous cooperation among disciplines to make the most of complementary expertise in resource development and processing [116].

2.3 Open Issues and Challenges

Methodologies aimed at data creation and dissemination should be fostered by both users and providers to maximize availability and usability. The goal should be the creation of data ecosystems that support the whole multimodal value chain - from design to distribution - through definition of best practices (e.g., like those available in Natural Language Processing) and setup of infrastructures for resource use and sharing [95]. These infrastructures will address the following needs:

- a framework for managing and sharing data collections;
- legal and technical solutions for privacy protection in a number of use scenarios;
- data visibility and encouragement to data sharing, reuse and repurposing for new research questions;
- identification of gaps and missing resources.

Establishing such an ecosystem in the area of multimodal interaction is necessary to support the increasingly demanding requirements of real-world applications. In particular, the creation of an effective data ecosystem promises to have the following advantages:

- integration of social and multimodal annotation into standard dialogue annotation schemes;
- building of knowledge-bases informing the design of real-world and impact-oriented applications;

– coverage of a wide, possibly exhaustive spectrum of contexts and situations;
– better analysis of context and genre in social interactions.

Overall, a solid shared data ecosystem would greatly streamline the acquisition of relevant scientific understanding of multimodal interaction, and thus expedite the use of this knowledge in the research and development of a range of novel real-world applications (see Section 6). The challenge remains at defining, labeling and annotating the high-level behaviours associated with human-human interaction. For this purpose, experts in multimodal signal processing and machine learning work hand-in-hand with psychologists, clinicians and other domain experts to transfer knowledge gained over years of labeling human behaviours to a machine readable code that is amenable to computational manipulation.

3 Behaviour Analysis

Previous research on social behaviour analysis has focused on individuals, whether it comes to the detection of specific actions and cues (e.g., facial expressions, gestures, and prosody) or the measurement of social and psychological phenomena (e.g., valence and arousal and personality traits). With advances in methodology, there is increasing interest in advancing beyond action detection in individuals to detection and understanding of interpersonal influence. Recent work includes comparing patterns of interpersonal influence under different conditions (e.g., with or without visual feedback, during high- versus low conflict, and during negative and positive affect [52, 54, 78, 117, 126]) and the relation between interpersonal influence and outcome variables (e.g., friendship or relationship quality [3, 97]). Key issues are feature extraction and representation, time-series methodologies, and outcomes. Unless otherwise noted, the rest of this section focuses on dyads (i.e., two interacting individuals) rather than larger social groups.

3.1 Detection of Behavioural Cues

The first step in computing interpersonal influence is to extract and represent relevant behavioural features from one or more modalities. Methodologies include motion-tracking [9] for body motion, computer vision [36, 54, 78, 126] for facial expression, head motion, and other visual displays, signal processing [59, 126] for voice quality, timing, and speech, and manual measurements by human observers [30, 71, 74]. Because of their objectivity, quantitative measurement, efficiency, and reproducibility, automatic measures are desirable. We address their limitations and challenges in Section 3.3.

3.2 Modeling Interpersonal Influence

Independent of specific methods of feature extraction, two main approaches have been used to analyze interpersonal influence. The first includes analytic and descriptive models that seek to quantify the extent to which behaviour of an individual account for the behaviour of another. The second includes prediction and classification models that seek to measure behavioural matching between interactive partners.

3.2.1 Analytic/Descriptive Models

Windowed cross-correlation is one of the most commonly used measures of similarity between two time-series [3, 97]. It uses a temporally defined window to measure successive lead-lag relationships over relatively brief time-scales [17, 53, 54, 78]. By using small window sizes, assumptions of stationarity are less likely violated. When time series are highly correlated at zero lag, they are said to be synchronous. When they are highly correlated at negative or positive lags, reciprocity is indicated. Patterns of cross-correlation may change across multiple windows, consistently with descriptions of mismatch and repair processes (e.g., in mother-infant dyads [29]).

Other approaches are recurrence analysis, accommodation, and spectral analysis. Recurrence analysis [104] seeks to detect similar patterns of change or movement in time series, which are referred to as "*recurrence points*". Accommodation [112], also referred to as convergence, entrainment, or mimicry [91, 92], refers to the tendency of dyadic partners to adapt their communicative behaviour to each other. Accommodation is based on a time-aligned moving average between time series. Spectral methods are particularly suitable for rhythmic processes. Spectral analysis measures phase shifts [87, 104] and coherence [36, 103, 104] or power spectrum overlap [87]. The methods cited above may suggest that one of the interaction participants influences the other (e.g., infants smile in response to their mother's smile), but it is more rigorous to say that they detect co-occurrence patterns that do not necessarily correspond to causal or influence relationships. Correlation or co-occurrence across multiple time series might be due to chance. A critical issue when attempting to detect dependence between time series is to rule out random cross-correlation or random cross-phase coherence. Two types of approaches may be considered. One of the most common is to apply surrogate statistical tests [9, 36, 71, 103]. For instance, the time series may be randomized. Statistics that summarize the relation between time series (e.g., correlations) then can be compared between the original and randomized series. If the statistics differ between the original and randomized series, that suggests a non-random explanation.

To investigate directionality requires one to consider alternative approaches. For instance, is synchrony achieved by one or both partners modifying their own behaviour in response to the other [29]? In behavioural science, both time- and frequency-domain approaches have been proposed to address this problem [48, 71]. These approaches involve statistical analysis of observational measures. Yet another approach is to introduce experimental perturbations into naturally occurring behaviour. In a video-conference, the output of one person's behaviour may be processed using an active appearance model and modified in real time without their knowledge. Using this approach, it has been found that attenuated head nods in an avatar resulted in increased head nods and lateral head turns in the other person [18]. Recent advances in image processing make possible real time experimental paradigms to investigate the direction of effects in interpersonal influence.

3.2.2 Prediction/Classification Models

In many applications, it is of interest to detect moments of similar behaviour between partners. For example, smiles in interactions between mothers and infants could be learned, and then their joint occurrence detected automatically. Mutual or synchronous head nodding, as in back-channeling, would be another example. A method to detect joint states using semi-supervised learning was proposed in [131]. Similarly, one could use supervised or unsupervised methods to learn phase relations between partners. This would include coordinated increasing or decreasing intensity of positive affect or mimicry. In [107], Hidden Markov Models (bi-grams) are employed to learn parent-infant interaction dynamics. This modeling is coupled with Non-Negative Matrix Factorization for a the extraction of a social signature of typical and autistic children. In [38], a set of One-Class SVM-based models are used to recognize the gestures of task partners during EEG hyper-scanning. A measure of "*imitation*" is then derived from the likelihood ratio between the models.

To reveal causal relations between time series, parametric approaches such as Actor-Partner analysis have been proposed [64]. These approaches assume that each person is potentially both cause and effect of others' behaviour. In [54], Actor-Partner analysis was used to measure the reciprocal relationship between head movements of intimate partners in conflict and non-conflict interaction. Each participant's head movements were used as both predictor and an outcome variable in the analyses. In behavioural science, it often is of interest to discover predictors of interaction outcomes from patterns of social interaction over time. For example, it was discovered that when husbands signaled contempt during arguments with their respective wives, divorce was in the offing [69]. Over 90% of divorces could be predicted in this way.

3.3 Open Issues and Challenges

Critical challenges are access to well-annotated data from dyads or other social groups (see Section 2), further advances in automated measurement, and improved analysis methodologies (see Section 5). Because distribution of spontaneous social interaction data has been constrained by confidentiality restrictions, investigators have been unable to train on and analyze each other's data. That limits advances in our methods. Often, however, participants would be agreeable to sharing their audio-video data if only asked. When participants have been given the opportunity to consent to such use by the research community, they often have consented. This has encouraged efforts to open access to data sources that would have been unavailable in the past (see Section 2). The U.S. National Institutes of Health [78] among others supports data-sharing efforts.

The current state of automated measurement presents limits. First, automatic feature extraction typically results in moderate rates of missing data, such as when head rotation exceeds the operational parameters of the system or face occlusion occurs. This is particularly germane when applying algorithms to participants much different than ones on which they were trained [52]. Second, while communication is multimodal, automated feature extraction typically is limited to one or few modalities. Despite advances in Natural Language Processing (NLP) [115], sampling and integration of speech with nonverbal measures remains a challenge. Third, optimal approaches to multimodal fusion are an open research question and may hinge on specific applications. In manual measurement, coders often use multimodal descriptors [29]. Comparable descriptors for automated feature extraction have yet to appear. In part for these reasons, some investigators have considered a combination of automatic and manual measurement [52, 54] or combination of overlapping algorithms for feature extraction [90].

A further key challenge is to propose statistical and computational approaches suitable for content and temporal structure of dyadic interactions. Various sequential learning models, such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs), are typically used to characterize the temporal structure of social interactions. Further approaches of this type will be of great benefit for automatic analysis and understanding of interpersonal communication in social interaction.

4 Multimodal Embodiment

In the past ten years, significant amount of effort has been dedicated to explore the potential of Social Signal Processing in human interaction with embodied conversational agents and social robots. The *social interaction capability* of an artificial agent may be defined as the ability of a system to interact seamlessly with humans. This definition implies the following:

- the human produces and expects responses to social signals in the communication with the agent;
- the agent is not only perceptive to the social signals emitted by the human, but also uses social signals to further its own purposes.

Especially, the latter point implies a rich internal representation of humans and human-human interactions for the agent.

Needless to say, specific aspects of embodied social interaction cannot be studied under laboratory conditions alone; naturalistic social settings and people's daily environments are needed to situate the user-agent communication (see Section 2 for issues related to the collection of data in naturalistic settings). Looking at recent literature, current goals in multimodal embodied interaction are focused on implementing sets of social communication skills in the agent, including detection of humor, empathy, compassion, and affect [130]. Basic skills like facial emotion recognition, gaze detection, dialogue management, and nonverbal signal processing are still far from being effective. Similarly, synthesis and timing of nonverbal signals, and appropriate ways of signaling apparent social cues are studied.

This section identifies two major challenges in this area. The first is that in these studies, typically, the cultural context is held fixed. One may argue that even humans have troubles selecting correct responses when the cultural setting is not familiar, but studies on artificial agents typically take place in very restricted domains, and naturalistic contexts are absent. The second problem is that the social behaviour of the agent is often not grounded in a rich internal representation, and lacks depth [7, 108]. When an agent shows signs of enjoying humor, it does that according to an internal rule triggered to display amusement as the appropriate response to a certain number of interactional situations. This way of modeling social exchanges is very rudimentary, and while it can be the initial step for implementing a social agent, it is very far from implementing the complexity and richness of social communications in real-life. The two issues mentioned above are strongly connected; without a proper internal representation, shallow models cannot be expected to adapt to different social contexts.

4.1 Approaches to Multimodal Embodiment

Social signals are strongly contextualized. For example, in a situation of bereavement, a gesture that is performed close to the interacting party can easily be interpreted as showing sympathy. The same gesture could be entirely inappropriate in a different context. The interpretation of social signals depends not only on the correct perception and categorization of the signal, but also on the evaluation and active interpretation of the interacting parties. While humans are adept at this, artificial agents lack the semantic background knowledge to deal with subtleties. Subsequently, the human-agent interaction needs to assume that the technology is limited, and compensate for its shortcomings by structuring the interaction in a way that the exchange follows signals that are clear and simple, tailored to the capabilities of the agent, but still rich enough to convey the internal states of the agent to the human, and vice versa.

Technologies for realizing individual components of a social agent have reached a great level of advancement. Progress in the area of social signal interpretation has been considerably fostered by a number of international benchmarking campaigns, such as the series of the *Audio/Visual Emotion Challenges* (AVEC) [1]. This is important for artificial agents that need to understand their users in naturalistic settings, but it is obvious that human-agent interactions do not necessarily need to use the same signals as those used in human-human interaction. The archetypical example is a domesticated cat, which produces a different set of social signals than a human, but seamlessly communicates over this set. The contribution of benchmarking campaigns is essential to the development of new solutions. Realistic data, naturalistic behaviours, and real-time processing are key aspects for these campaigns. The latter aspect is particularly important, as most challenges focus on offline processing, but the online mode, which is essential for real, situated social interactions, is a much more difficult setting [110]. Candace Sidner and Charles Rich [102] coined the term *always-on relational agents* to describe the vision of a robotic or virtual character that lives as a permanent member in a human household, which remains a grand challenge. In a related perspective, Barbara Grosz [50] stated that: "*Is it imaginable that a computer (agent) team member could behave, over the long term and in uncertain, dynamic environments, in such a way that people on the team will not notice it is not human*." The perception, negotiation, and generation of social cues in a context is necessary to achieve this condition.

[1] `http://sspnet.eu/avec2014/`

4.2 Open Issues and Challenges

As tools become more diversified and layered, it becomes possible to create agents with more depth. Work done in the SEMAINE project [113] has shown that simple backchannel signals, such as "*I see*", may suffice to create the illusion of a sensitive listener. However, to engage humans over a longer period of time, a deeper understanding of the dialogue would be necessary. While a significant amount of work has been done on the semantic/pragmatic processing in the area of NLP, work that accounts for a close interaction between the communication streams required for semantic/pragmatic processing and social signal processing is rare (see Section 5). The integration of Social Signal Processing with semantic and pragmatic analysis may help to resolve ambiguities. Especially short utterances tend to be highly ambiguous when solely the linguistic data is considered. An utterance like "*right*" may be interpreted as a confirmation, as well as a rejection, if intended cynically, and so may the absence of an utterance. Preliminary studies have shown that the consideration of social cues may help to improve the robustness of semantic and pragmatic analysis [21] (see Section 5).

Finding the right level of sensitivity is very important in creating seamless interaction, and this requires strong adaptation skills for the agent. Mike Mozer's early experiments on the adaptive Neural Network House established that people tolerate only to a limited extent the mistakes of an "*intelligent*" system [82]. This is true for social signals as well; agents that act and react inappropriately will most likely irritate users [6]. Treating all user behaviours as possible input to the agent (called the "*Midas Touch Problem*" [55]) will result in poor interactions, and confused users.

Recent work by the "*Natural Interaction with Social Robots Topic Group*" [2] (NISR-TG) proposes to use several levels to describe the social ability of an agent:

- *Level 0*: the agent does not interact with the human;
- *Level 1*: the agent perceives the human as an object (useful for orienting and navigating);
- *Level 2*: the agent perceives the human as another agent that is represented explicitly, and can be re-identified time and again;
- *Level 3*: a two-way interaction is possible, provided that the interacting human knows and obeys some conventions and behaviours required by the agent's system;
- *Level 3a*: a two-way interaction is possible with the ability of spoken language interaction;
- *Level 4*: the agent adapts its behaviour to the interaction partners during the interaction;
- *Level 5*: the agent recognizes different users and adjusts its behaviour accordingly;
- *Level 6*: the agent is capable to interact with more than one users;
- *Level 7*: the agent is endowed with personality traits that can be recognised as such by the users and result into displaying different behaviours in the same situations;
- *Level 8*: the agent is capable to learn and accumulate experience over multiple interactions;
- *Level 9*: the agent is capable to build and sustain relationships with its users.

Progressing through the levels, the agent is expected to gain one-way and two-way interaction capabilities, followed by a more advanced set of skills including adaptation, multiparty interaction management, and the incorporation of social constructs like personality.

Humans adapt their social behaviours during interactions based on explicit or implicit cues they receive from the interlocutor. In order to establish longer lasting relationships between artificial companions and human users, artificial companions need to be able to adjust their behaviour on the basis of previous interactions. That is, they should remember previous interactions and learn from them [10]. To this end, sophisticated mechanisms for the simulation of self-regulatory social behaviours will be required. Furthermore, social interactions will have to be personalized to individuals of different gender, personality and cultural background. For example, cultural norms and values determine whether it is appropriate to show emotions in a particular situation [76] and how they are interpreted by others [75]. While offline learning is prevalent in current systems exploiting SSP techniques, future work should explore the potential of online learning in order to enable continuous social adaptation processes. For the integration of context, novel sensor technologies can be used by the agent in ways that are not available to humans in an ordinary interaction [35, 129]. Multimodality can also be harnessed in expressing social signals in novel ways, for instance by adding haptic cues to visual displays [15, 43].

At a finer level, a single interaction between two agents also involves an *interactive alignment* (also see Section 3), where the interacting parties converge on similar representations at different levels of linguistic processing [45, 94]. The alignment at higher-levels (e.g. common goals), relies on the alignment of lower-levels (e.g. objects of joint attention). This requires that the agents model their interaction partners, anticipate interaction directions, align their communication acts, as well as actions [105]. We can safely assume that research in cognitive science and linguistics will be essential in achieving these goals (see Section 5).

[2] http://homepages.stca.herts.ac.uk/comqkd/TG-NaturalInteractionWithSocialRobots.html

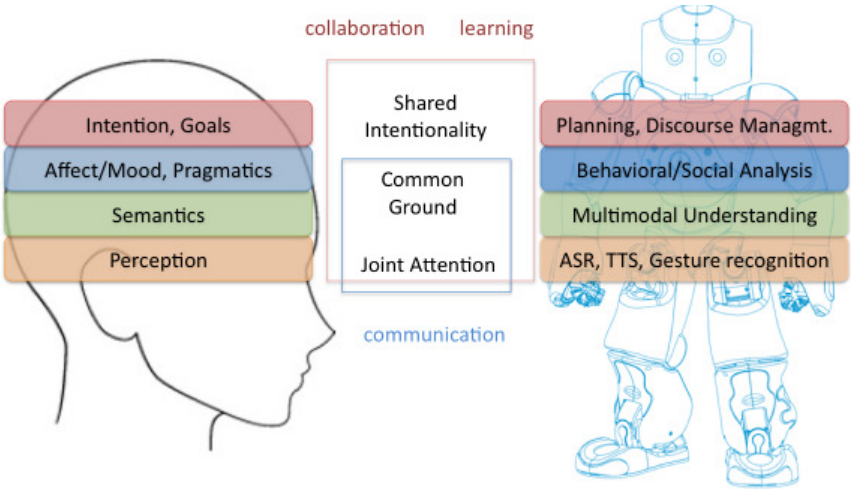

Fig. 1 A layered view of human-machine interaction (ASR and TTS stand for Automatic Speech Recognition and Text to Speech, respectively).

5 Computational Models of Interaction

Broadly based on the work of Tomasello [120, 121] (and others) human-human interaction can be represented as a three-step process: sharing attention, establishing common ground and forming shared goals (a.k.a. *joint intentionality*). Two prerequisites for successful human-human communication via joint intentionality are:

- the ability to form a successful model of the cognitive state of people around us, i.e., decoding not only overt, but also covert communication signals also referred to as "*recursive mind-reading*";
- establishing and building trust, a truly human trait.

Affective computing, SSP and Behavioural Signal Processing (BSP) address the first prerequisite, building machines that can understand the emotional, social and cognitive state of an individual. A layered view of human-machine interaction from the cognitive and computational perspectives are shown in Figure 1. This section reviews computational models and associated challenges for each layer.

5.1 Joint Attention and Saliency

Unlike computers, humans are able to process only the most salient parts of an image, a sound or a brochure, literally ignoring the rest. Being able to model and predict what a human sees and hears in an audio-visual scene is the first step towards forming a cognitive representation of that scene, as well as establishing common ground in interaction scenarios.

Saliency- and attention-based models have played a significant role in multimedia processing in the past decade [22, 33, 42, 58, 63, 70, 83, 98, 118], exploiting low-level cues from the (mostly) visual, audio, and spoken language (transcription) modalities: they have proved very successful in identifying salient events in multimedia for a variety of applications. However, attention-based algorithms typically use only perceptually motivated low-level (frame-based) features and employ no high-level semantic information, with few exceptions in very specific cases (mostly in the visual domain) [83].

Challenges still remain on: 1) mid- and high-level feature extraction including incorporating semantics (scenes, objects, actions) and 2) computational models for the *multimodal* fusion of the bottom-up (gestalt-based) and top-down (semantic-based) attentional mechanisms. Also applying these multimodal salient models to realistic human-human (especially) and human-computer interaction scenarios remains a challenge. The most promising research direction for these challenges seems to be deep learning, where the integration among levels of diverse granularity of knowledge is the core skill [14]. Finally, identifying the dynamics of attention, i.e., constructing joint (interactional) attention models remains an open problem in this area.

5.2 Common Ground and Concept Representations

While interacting, humans process and disambiguate multimodal cues, integrating low-, mid- and high-level cognitive functions, specifically using the low-level machinery of cue selection (as discussed above) via (joint) attention, the mid-level machinery of semantic disambiguation via common ground and shared conceptual representations, and the high-level machinery of intention reading.

Since establishing common ground is a prerequisite for successful communication, an essential module of an ideal interacting machine should model an extensive cognitive semantic/pragmatic representation, that is, a network of con-

cepts and their relations that form the very essence of common ground, and in this sense formal ontologies may help.

Formal Ontologies [51] are a top-down (knowledge-based) semantic representation that has been used for interaction modeling mainly by the research community, e.g., [132]. The main advantages of ontologies are description clarity (via mathematical logic) and inference power. However, the following challenges remain to make ontologies a viable representation for practical interactional systems:

- mapping between the semantic and lexical/surface representations, a.k.a. the "*lexicalization*" of ontologies necessary both for natural language (NL) understanding [28] and for NL generation [8]; the same problem holds also in the visual domain, where a proper "*visual ontology*" is missing, or available in very restricted domains [122];
- representing ambiguous semantics [4];
- representing complex semantics, e.g., time relationships [13];
- combining ontology-driven semantics with bottom-up (data-driven) approaches, e.g., for grammar induction [46] and in general for computer vision.

Grounding exists only in the context of our semantic, affective and interactional cognitive representations and should be addressed as such. This poses the grand challenge of using "*big data*" to construct such cognitive representations, as well as defining the "*topology*" (unified vs. distributed) and processing logic (parallel/serial) of these representations. Cognitively motivated conceptual representations and novel machine learning algorithms (representation and transfer learning [89]) can be used to face these issues by designing algorithms that 1) achieve rapid learning and adaptation to new concepts and situations from very few examples (situational learning and understanding), and 2) provide grounding in interaction and problem-solving settings (negotiating common ground).

5.3 From Semantics to Behaviour and Interaction

Even if it was possible to solve the multimodal understanding problem by mapping from signal(s) to semantics (a monumental task by itself), it would still be only half of the way. Assuming that a conceptual representation is in place (see above), this section discusses how to model jointly semantics and affect.

Given that the cognitive semantic space is both distributed and fragmented into subspaces, the mapping from semantics to affective labels should also be distributed and fragmented. Semantic-affective models (SAM) [72, 123] are based on the assumption that semantic similarity implies affective similarity. Thus affective models can be simply constructed as mappings from semantic neighborhoods to affective scores. In the SAM model proposed in [72, 73], the affective label of a token can be expressed as a map (trainable linear combination) of its semantic similarities to a set of seed words and the affective ratings of these words. The model can be extended to also handle many-to-many mappings between multiple layers of cognitive representations. The model is consistent with (and implementable via) the multi-layered cognitive view of representation and deep learning models.

Although good performance can be obtained for language and image processing applications, the challenge remains on how to apply this model to audio and video, where the segmentation of the stream into tokens is not straightforward. Also, the model works very well at estimating the affective content of single tokens (words, images); going from a single token to a sequence of tokens (e.g., word to sentences) is a hard open problem. Last but not least, generalizing this model to other behavioural labels remains a grand challenge.

5.4 Open Issues and Challenges

The previous sections have identified major challenges that lie ahead in the fields of affective, social and behavioural signal processing as it pertains to interaction modeling. The sections have also argued that it is very improbable that one can successfully address these major challenges without taking into account the peculiarities of human cognition.

The proposition of this paper is that the solution of these problems should be grounded on human cognition, including modeling the errors (cognitive biases) and non-linear logic of humans [101]. Although "*pure*" machine learning algorithms often achieve good performance for classification of low- and mid-level labels, they are less successful with higher-level behavioural classification tasks. This can be partially attributed to the ambiguity, abstraction, subjectivity and representation depth inherent in high-level cognitive tasks. Cognitively inspired models can represent the very errors, biases, subjective beliefs and attitudes of a human. Thus, adopting a human-centered approach becomes increasingly important as we move from signals to behaviours and interaction. The recent achievements of cognitively motivated machine learning paradigms such as representation, transfer and deep learning further validate this view. Interaction modeling poses new challenges and opens up fruitful research directions for the years to come.

6 Applications

Effectiveness in real-world applications is the ultimate test for any technology-oriented research effort. While being an opportunity for methodological progress and acquisition of key-insights about human psychology and cognition, research on modeling, analysis and synthesis of human behaviour

aims at achieving impact in terms of both commercial exploitation, i.e. development of products that reach the market and result into jobs creation, and solutions to societal problems, i.e. development of systems that improve the quality of life, especially when it comes to disadvantaged categories.

Addressing the issues and challenges presented in this work will certainly advance the state-of-the-art, but it will increase the chances of success for a wide spectrum of real-world technologies as well (the list is not exhaustive):

- Analysis of agent-customer interactions at call centres[3] with the goal of improving the quality of services [44];
- Improvement of tutoring systems aimed at supporting students in individual and collective learning processes [111];
- Creation of speech synthesizers[4] that convey both verbal and nonverbal aspects of a text [114];
- Enrichment of multimedia indexing systems with social and affective information [6];
- Recommendation systems that take into account stable individual characteristics (e.g., personality traits) and transient states (e.g., emotions) [119];
- Socially intelligent surveillance and monitoring systems [34];

The rest of this section focuses on three application domains that address crucial issues and aspects of everyday life, namely healthcare, human-machine interactions and human-human conversations. The three cases account for three major steps in the process that leads from the laboratory to the real-world:

- The development of a vision based on current state-of-the art and major technological trends in the case of healthcare personal agents (see Section 6.1);
- The realization of a prototype that addresses one specific application (intelligent control centres), but results into the definition of principles that can be transferred to other areas (see Section 6.2);
- The definition of concrete steps bridging the gap between research, industry and society in the case of conversational technologies (see Section 6.3).

The description of the case studies above will provide insights regarding the interdependency between the challenges outlined so far and application-driven needs.

6.1 The Healthcare Personal Agent: a Vision for the Future of Medicine

Advances in mobile technologies such as voice, video, touchscreens, web 2.0 capabilities and integration of various onboard sensors and wearable computers, have rendered mobile devices as ideal units for delivery of healthcare services [11]. At the same time the dawn of the data-driven economy has stirred the innovation of processes and products. Unfortunately, the innovation has been slow in the healthcare sector where much innovation is needed to improve the quality of the service at various end-points (hospitals, healthcare professionals, patients) and reduce costs.

The 2012 survey in [84] reports that in Europe there were more than one hundred health apps in a variety of languages (Turkish, Italian, Swedish, etc..) and domains (mental problems, self-diagnosis, heart-monitoring, etc.). Such growing number of smartphone applications can track user activity, sleeping and eating habits and covert and overt signals such as blood pressure, heart rate, skin temperature, speech, location, movement, etc. by either using the on-board sensors of the smartphone or interacting with various wearable and healthcare monitoring devices.

In the recent years there has been a growing research interest in creating such applications which can interact with people though context-aware multimodal interfaces and have been used for various healthcare services ranging from monitoring and accompanying the elderly [11, 79], to providing healthcare interventions for long-term behaviour changes [81]. Such agents can be useful in keeping track of patient activity in-between visits or to ensure the patients are taking their medicines on time, or that they follow their advised health routine (see Section 4 for challenges related to "*always on*" agents).

In the future, Healthcare Personal Agent research and development should plan for an agenda where current limitations are addressed and new avenues are explored. Such agenda can directly impact the quality of life and health of people by disrupting current models of delivering healthcare services. Agents will have different physical and virtual appearance (see Section 4 for challenges in embodiment) ranging from avatars to robots (e.g. [79]). Covert signal streams from wearable and mobile sensors may be effectively used to model user state in terms of his/her physiological responses to external stimuli, events and medical protocol he/she is following (see Section 3 for challenges related to behaviour analysis).

Personal Agents need to be able to handle basic and complex emotions such as empathy. In the healthcare domain, the ability to handle emotions is critical to manage and support, for instance, daily healthcare routine. The affective signals and communication need to be adapted for target patient groups such as children, elderly people, etc. By far one of the most important social and cognitive skills of a conversational agent is the ability to carry out a dialogue with a human (see Section 6.3). Different models of user interaction might be needed for different users/user-groups and different application domains (e.g. robotic surgery vs bank fund transfer vs information seeking). An application tracking brushing habits of kids might achieve better results with

[3] See `http://www.cogitocorp.com` for a company working on the analysis of call centre conversations.

[4] See `https://www.cereproc.com` for a company active in the field.

Fig. 2 *Left:* Comfortable sensor-equipped chair. Micro gestures allow for natural interaction during lengthy passive monitoring periods. *Middle:* Operators at workstations are tracked and an acoustic interface targets sound at a particular operator without disturbing others. *Right:* Collaboration and distribution of urgent tasks via hand gestures and shared screens.

gamification, while an obesity monitoring agent should use motivational feedback to improve user compliance.

6.2 Building a Working Prototype: The Example of Intelligent Control Centres

Human-Computer Interaction is one of the domains that directly benefit from multimodal technologies for human behaviour understanding. This applies in particular to applications where machines must adapt as intelligent as possible to the natural and spontaneous behaviour of their users because these need to concentrate their attention and cognitive efforts on difficult and demanding tasks.

Reducing the cognitive load and enabling immediate reaction to alarms in idle times are key requirements that have driven the development of the innovative control centre described in [61]. Comparable efforts on concrete applications have worked on ship bridges [66] and crises response control rooms [56].

In control centers, teams of human operators collaborate to monitor and manipulate external processes, such as in industrial production, IT and telecommunication infrastructure, or public infrastructure such as transportation networks and tunnels. In this domain, innovation towards user interfaces has been picked up slowly since it is limited by governmental regulation or short-term return-on-investment considerations. Surprisingly, many of the systems in use were first built decades ago and have been extended iteratively without proper re-design of their user-interfaces until today. Recent generations of operators, however, are digital natives and hence familiar with mobile devices, gesture interfaces and touch screens, for example. While considerable business opportunities can be expected in the next decade to redesign the interfaces in such control rooms, many research challenges remain to be addressed.

Most current systems feature redundant input devices, little context awareness, and expose operators to information overflow. The support for distribution of tasks and collaboration in general leaves to be desired. One key enabling factor in the re-design of such complex systems is the dynamic interpretation of the operators' actions and interactions as a team while taking the current situation (goal, alarm and stress level, etc.) into account (see Section 3 for the challenges related to understanding the behaviour of groups). Inspired by a human-centered design approach, the concept recently proposed in [61] experiments with the combination of visual cues, micro (i.e. fingers and hands only) and macro gestural interaction, an acoustic interface with individualized sound radiation, and intelligent data processing (Semantic Lifting, see [62]) into a single, universal interface. The concept is considering specific needs of the operators and the length of work shifts, which for example led to the omission of wearable devices such as headsets. Figure 2 illustrates several components of this multimodal interaction concept [61]. The work made clear that while research has been addressing the combination of input and output devices of multiple modalities, a lot more applied research is required on their interplay regarding specific tasks in real industry settings.

In a safety critical environment, user interaction requires different levels of robustness and precision according to the tasks. Control center operators conduct very specific tasks that call for different interaction devices and concepts. Their integration and dynamic adaptation is a challenge. An underlying aim is to actively manage the cognitive load of the operators, mainly to ensure quick reaction in alarm situations. There are idle times where operators essentially take a break but don't leave their workplace, lengthy passive monitoring tasks, and very urgent alarm handling situations. A significant impact can be expected in this domain by improved user behaviour analysis.

6.3 Roadmapping Research and Innovation in Conversational Interaction Technologies

The research community in multimodal conversational interaction has advanced significantly in recent years, however –

despite the fast growth of multimodal smartphone technologies, for example – innovation and commercial exploitation is not always closely connected to research advances. To develop and integrate research and innvovation in this area, it is thus important to identify the key innovation drivers and most promising elements across science, technology, products, and services on which to focus in the future.

Technology roadmapping is a process to lay out a path from science and technology development through integrated demonstration to products and services that address business opportunities and societal needs. Often performed by individual businesses, it can also be used to put together all of the different viewpoints and information sources available in a large stakeholder community as a way of helping them work together and achieve more. The EU ROCKIT project, driven by a broad vision for conversational interaction technologies, has constructed a technology roadmap for Conversational Interaction Technologies (`http://www.citia.eu`).

In consultation with researchers and companies of every size (including several workshops involving about 100 resesarchers and technologists), the ROCKIT support action constructed a technology roadmap for Conversational Interaction Technologies. Since research and business environments can change rapidly, the resultant roadmap is structured to enable stakeholders to steer through change and understand how they can achieve their goals in a changing context. For this reason, the roadmap is not just a series of steps that go from current science and technology outcomes to future profitable products and services, but conveys the relationships among societal drivers of change, products and services, use cases for them, and research results.

The ROCKIT roadmap connects the strong research base with commercial and industrial activity, and with policy makers. To develop the roadmap, and and to make tangible links between research and innovation a small number of *target scenarios* have been developed. Each scenario includes its societal and technological drivers, research aspects, market and business drivers, and potential testbeds. We identified a number of common themes coming out of ROCKIT's consultations with stakeholders, in particular accessibility, multilinguality, the importance of design, privacy by design, systems for all of human-human, human-machine, and human-environment interactions, robustness, security, potentially ephemeral interactions, and using the technology to enable fun.

Building on these themes, together with the different social, commercial, and technological drivers, we have identified five possible target application scenarios:

- *Adaptable interfaces for all*: Interfaces which recognize who you are, where you are, and eventually what you want, by drawing on a profiled knowledge base about your habits and preferences. They will therefore be able to adapt to your disability, language, visual competency, specific need for speech or typed input depending on whether you are driving/working with two hands on a repair job or are seated in front of a keyboard, physical or virtual, or are prostrate in bed (see Section 4 for challenges related to agents with internal representations of users and ability to adapt to context and interactions).
- *Smart personal assistants*: Multisensory agents able to integrate heterogeneous sources of knowledge, display social awareness, and behave naturally in multiuser situations (see Section 4 for challenges related to synthesis of social behaviour).
- *Active access to complex unstructured information*: Linking knowledge to rich interaction will enable the development of agents which can search proactively and can make inferences from their (possibly limited) knowledge, to enable people to be notified of relevant things faster, and to help people reach understanding of complex situations involving many streams of information (see Section 5 for challenges in representing knowledge and cognitive processes).
- *Communicative robots*: Embodied agents able to display personality and to generate and interpret social signals (see Sections 3 and 4 for related challenges).
- *Shared collaboration and creativity*: Empowering and augmenting communication between people. This will include new approaches to social sharing (across languages), design platforms which enable people to build their own tools, and scalable systems that enable groups to collaborate with shared goals, facilitating problem solving, and providing powerful mechanisms for engagement.

7 Conclusions

This article has described some of the most important challenges and issues that need to be addressed in order to achieve substantial progress in technologies for the modeling, analysis and synthesis of human behaviour, especially for what concerns social interactions. Section 2 has shown that data, while being a crucial resource, cannot become an asset for the community without widely accepted practices for design, collection and distribution. Section 3 has proposed to move the focus of analysis approaches from individuals involved in an interaction to phenomena that shape groups of interacting people (e.g., interpersonal influence and social contagion). Section 4 highlighted the need of endowing machines, in particular embodied conversational agents, with an internal representation of their users. Section 5 has focused on the possibility of integrating models of human cognitive processes and semantics in technologies dealing with human behaviour. Finally, Section 6 has overviewed application domains that can benefit, or are already benefiting, from technologies aimed at modeling, analysis and synthesis of behaviour.

While addressing relatively distinct problems, the challenges above have a few aspects in common that might guide at least the first steps required to address them. The first is that human behaviour is always situated and context dependent. Therefore, technologies for dealing with human behaviour should try to address highly specific aspects of the contexts where they are used rather than trying to be generic. Conversely, it should be always kept in mind that an approach effective in a given situation or context might not work in others. The second is the need of considering both verbal and nonverbal aspects of human-human and human-machine interactions. So far, verbal content and semantics tend to be neglected, the reason being that nonverbal aspects are more honest and, furthermore, taking into account what people say violates the privacy. The third is to model explicitly the processes that drive interactive behaviour in humans like, e.g., the development of internal representation of others.

The last part of the article has considered three application case studies that account for different steps of the process that leads from laboratory to real-world applications. Healthcare personal agents have been proposed as a case of research vision that builds upon current technology trends (in particular the diffusion of mobile devices and the availability of large amounts of data) to design new applications of technologies for analysis of behaviour. The case of the intelligent control centres has shown that the implementation of an application-driven prototype provides insights on how technologies revolving around behaviour should progress. Finally, the case of conversational technologies has given an example of how a roadmapping process can contribute to bridge the gap between research and application.

Needless to say, the issues proposed in this article do not necessarily cover the entire spectrum of problems currently facing the community. Furthermore, new challenges and issues are likely to emerge while the community addresses those described in this work. However, dealing with the problems proposed in this article will certainly lead to substantial improvements of the current state-of-the-art.

Acknowledgments

This paper is the result of the discussions held at the "*International Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*" (`sspnet.eu/2014/06/rfmir/`), held in conjunction with the ACM International Conference on Multimodal Interaction (2014).

Alexandros Potamianos was partially supported by the BabyAffect and CogniMuse projects under funds from the General Secretariat for Research and Technology (GSRT) of Greece, and the EU-IST FP7 SpeDial project. Dirk Heylen was supported by the Dutch national program COMMIT. Giuseppe Riccardi was partially funded by EU-IST FP & SENSEI project. Mohamed Chetouani was partially supported by the Labex SMART (ANR-11-LABX-65) under French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02. Jeffrey Cohn and Zakia Hammal were supported in part by grant GM105004 from the U.S. National Institutes of Health.

References

1. Allwood J, Björnberg M, Grönqvist L, Ahlsén E, Ottesjö C (2000) The Spoken Language Corpus at the department of linguistics, Göteborg university. In: Forum Qualitative Social Research, vol 1
2. Allwood J, Cerrato L, Jokinen K, Navarretta C, Paggio P (2007) The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation 41(3-4):273–287
3. Altmann U (2011) Studying movement synchrony using time series and regression models. In: Esposito A, Hoffmann R, Hübler S, Wrann B (eds) Program and Abstracts of the Proceedings of COST 2102 International Training School on Cognitive Behavioural Systems
4. Ammicht E, Fosler-Lussier E, Potamianos A (2007) Information seeking spoken dialogue systems—part I: Semantics and pragmatics. IEEE Transactions on Multimedia 9(3):532–549
5. Anderson A, Bader M, Bard E, Boyle E, Doherty G, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, et al (1991) The HCRC map task corpus. Language and speech 34(4):351–366
6. André E (2013) Exploiting unconscious user signals in multimodal human-computer interaction. ACM Transactions on Multimedia Computing, Communications, and Applications 9(1s):48
7. André E (2014) Challenges for social embodiment. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 35–37
8. Androutsopoulos I, Lampouras G, Galanis D (2014) Generating natural language descriptions from owl ontologies: the naturalowl system. arXiv preprint arXiv:14056164
9. Ashenfelter KT, Boker SM, Waddell JR, Vitanov N (2009) Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. Journal of Experimental Psychology: Human Perception and Performance 35(4):1072
10. Aylett R, Castellano G, Raducanu B, Paiva A, Hanheide M (2011) Long-term socially perceptive and in-

teractive robot companions: challenges and future perspectives. In: Bourlard H, Huang TS, Vidal E, Gatica-Perez D, Morency LP, Sebe N (eds) Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, ACM, pp 323–326
11. Baig MM, Gholamhosseini H (2013) Smart health monitoring systems: An overview of design and modeling. Journal of Medical Systems 37(2)
12. Baker R, Hazan V (2010) LUCID: a corpus of spontaneous and read clear speech in british english. In: Proceedings of the DiSS-LPSS Joint Workshop 2010
13. Batsakis S, Petrakis EG (2011) SOWL: a framework for handling spatio-temporal information in owl 2.0. In: Rule-Based Reasoning, Programming, and Applications, Springer Verlag, pp 242–249
14. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1798–1828
15. Bickmore TW, Fernando R, Ring L, Schulman D (2010) Empathic touch by relational agents. IEEE Transactions on Affective Computing 1(1):60–71
16. BNC-Consortium (2000) http://www.hcu.ox.ac.uk/BNC
17. Boker SM, Xu M, Rotondo JL, King K (2002) Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series,. Psychological Methods 7(3):338 – 355
18. Boker SM, Cohn JF, Theobald BJ, Matthews I, Spies J, Brick T (2009) Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars. Philosophical Transactions of the Royal Society B 364:3485–3495
19. Bonin F, Gilmartin E, Vogel C, Campbell N (2014) Topics for the future: Genre differentiation, annotation, and linguistic content integration in interaction analysis. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 5–8
20. Bos J, Klein E, Lemon O, Oka T (2003) DIPPER: Description and formalisation of an information-state update dialogue system architecture. In: Proceedings of SIGdial Workshop on Discourse and Dialogue, pp 115–124
21. Bosma W, André E (2004) Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the International Conference on Intelligent User Interfaces, pp 85–92
22. Boujut H, Benois-Pineau J, Ahmed T, Hadar O, Bonnet P (2011) A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp 1–5
23. Breazeal CL (2004) Designing sociable robots. MIT press
24. Bunt H (1995) Dialogue control functions and interaction design. NATO ASI Series F Computer and Systems Sciences 142:197–197
25. Campbell N (2007) Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. In: Proceedings of the International Congress of the Phonetic Sciences, pp 343–348
26. Cassell J (2000) Embodied conversational agents. MIT press
27. Chetouani M (2014) Role of inter-personal synchrony in extracting social signatures: Some case studies. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 9–12
28. Cimiano P, Buitelaar P, McCrae J, Sintek M (2011) Lexinfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web 9(1):29–51
29. Cohn J, Tronick E (1988) Mother-infant face-to-face interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior. Developmental Psychology 34(3):386–392
30. Cohn JF, Ekman P (2005) Measuring facial action by manual coding, facial emg, and automatic facial image analysis. In: Harrigan J, Rosenthal R, Scherer K (eds) Handbook of nonverbal behavior research methods in the affective sciences, Oxford University Press, pp 9–64
31. Core M, Allen J (1997) Coding dialogs with the DAMSL annotation scheme. In: AAAI Fall Symposium on Communicative Action in Humans and Machines, pp 28–35
32. Cristani M, Ferrario R (2014) Statistical pattern recognition meets formal ontologies: Towards a semantic visual understanding. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 23–25
33. Cristani M, Bicego M, Murino V (2004) On-line adaptive background modelling for audio surveillance. In: Proceedings of the International Conference on Pattern Recognition, vol 2, pp 399–402
34. Cristani M, Raghavendra R, Del Bue A, Murino V (2013) Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing 100:86–97

35. Damian I, Tan CSS, Baur T, Schöning J, Luyten K, André E (2014) Exploring social augmentation concepts for public speaking using peripheral feedback and real-time behavior analysis. In: Proceedings of the International Symposium on Mixed and Augmented Reality
36. Delaherche E, Chetouani M (2010) Multimodal coordination: exploring relevant features and measures. In: Proceedings of the International Workshop on Social Signal Processing, pp 47–52
37. Delaherche E, Chetouani M, Mahdhaoui M, Saint-Georges C, Viaux S, Cohen D (2012) Interpersonal synchrony : A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing 3(3):349– 365
38. Delaherche E, Dumas G, Nadel J, Chetouani M (2015) Automatic measure of imitation during social interaction: a behavioral and hyperscanning-EEG benchmark. Pattern Recognition Letters (to appear)
39. DuBois JW, Chafe WL, Meyer C, Thompson SA (2000) Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium
40. Edlund J, Beskow J, Elenius K, Hellmer K, Strömbergsson S, House D (2010) Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture. In: Proceedings of Language Resources and Evaluation Conference
41. Ekman P, Huang T, Sejnowski T, Hager J (1992) Final report to NSF of the planning workshop on facial expression understanding. URL `http://face-and-emotion.com/dataface/nsfrept/nsf_contents.html`
42. Evangelopoulos G, Zlatintsi A, Potamianos A, Maragos P, Rapantzikos K, Skoumas G, Avrithis Y (2013) Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. IEEE Transactions on Multimedia 15(7):1553–1568
43. Gaffary Y, Martin JC, Ammi M (2014) Perception of congruent facial and haptic expressions of emotions. In: Proceedings of the ACM Symposium on Applied Perception, pp 135–135
44. Galanis D, Karabetsos S, Koutsombogera M, Papageorgiou H, Esposito A, Riviello MT (2013) Classification of emotional speech units in call centre interactions. In: Proceedings of IEEE International Conference on Cognitive Infocommunications, pp 403–406
45. Garrod S, Pickering MJ (2009) Joint action, interactive alignment, and dialog. Topics in Cognitive Science 1(2):292–304
46. Georgiladakis S, Unger C, Iosif E, Walter S, Cimiano P, Petrakis E, Potamianos A (2014) Fusion of knowledge-based and data-driven approaches to grammar induction. In: Proceedings of Interspeech
47. Godfrey JJ, Holliman EC, McDaniel J (1992) SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 1, pp 517–520
48. Gottman J (1981) Time series analysis: A comprehensive introduction for social scientists. Cambridge University Press
49. Greenbaum S (1991) ICE: The International Corpus of English. English Today 28(7.4):3–7
50. Grosz BJ (2012) What question would Turing pose today? AI Magazine 33(4):73–81
51. Guarino N (1998) Proceedings of the International Conference on Formal Ontology in Information Systems. IOS press
52. Hammal Z, Cohn J (2014) Intra- and interpersonal functions of head motion in emotion communication. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, in conjunction with the 16th ACM International Conference on Multimodal Interaction ICMI 2014. 12-16 November 2014, pp 19–22
53. Hammal Z, Cohn JF, Messinger DS, Masson W, Mahoor M (2013) Head movement dynamics during normal and perturbed parent-infant interaction. In: Proceedings of the biannual Humaine Association Conference on Affective Computing and Intelligent Interaction
54. Hammal Z, Cohn JF, George DT (2014) Interpersonal coordination of head motion in distressed couples. IEEE Transactions on Affective Computing 5(9):155–167
55. Hoekstra A, Prendinger H, Bee N, Heylen D, Ishizuka M (2007) Highly realistic 3D presentation agents with visual attention capability. In: Proceedings of International Symposium on Smart Graphics, pp 73–84
56. Ijsselmuiden J, Grosselfinger AK, Münch D, Arens M, Stiefelhagen R (2012) Automatic behavior understanding in crisis response control rooms. In: Ambient Intelligence, Lecture Notes in Computer Science, vol 7683, Springer, pp 97–112
57. ISO (2010) Language resource management: Semantic annotation framework (SemAF), part 2: Dialogue acts
58. Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research 40(10):1489–1506
59. Jaffe J, Beebe B, Feldstein S, Crown CL, Jasnow M (2001) Rhythms of dialogue in early infancy. Monographs of the Society for Research in Child Development 66(2)

60. Janin A, Baron D, Edwards J, Ellis D, Gelbart D, Morgan N, Peskin B, Pfau T, Shriberg E, Stolcke A (2003) The ICSI meeting corpus. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 1, pp 1–364
61. Kaiser R, Fuhrmann F (2014) Multimodal interaction for future control centers: Interaction concept and implementation. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges
62. Kaiser R, Weiss W (2014) Virtual director. In: Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media, Wiley, pp 209–259
63. Kalinli O (2009) Biologically inspired auditory attention models with applications in speech and audio processing. PhD thesis, University of Southern California
64. Kenny D, Mannetti L, Pierro A, Livi S, Kashy D (2002) The statistical analysis of data from small groups. Journal of Personality and Social Psychology 83(1):126
65. Koutsombogera M, Papageorgiou H (2014) Multimodal analytics and its data ecosystem. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 1–4
66. Kristiansen H (2014) Conceptual design as a driver for innovation in offshore ship bridge development. In: Maritime Transport VI, pp 386–398
67. Larsson S, Traum DR (2000) Information state and dialogue management in the TRINDI dialogue move engine toolkit. Natural Language Engineering 6(3&4):323–340
68. Lemke JL (2012) Analyzing verbal data: Principles, methods, and problems. In: Second International Handbook of Science Education, Springer, pp 1471–1484
69. Levenson R, Gottman J (1983) Marital interaction: Physiological linkage and affective exchange. Journal of Personality and Social Psychology 45(3):587–97
70. Liu T, Feng X, Reibman A, Wang Y (2009) Saliency inspired modeling of packet-loss visibility in decoded videos. In: International Workshop on Video Processing and Quality Metrics for Consumer Electronics, pp 1–4
71. Madhyastha TM, Hamaker EL, Gottman JM (2011) Investigating spousal influence using moment-to-moment affect data from marital conflict. Journal of Family Psychology 25(2):292–300
72. Malandrakis N, Potamianos A, Iosif E, Narayanan S (2013) Distributional semantic models for affective text analysis. IEEE Transactions on Audio, Speech, and Language Processing 21(11):2379–2392
73. Malandrakis N, Potamianos A, Hsu KJ, Babeva KN, Feng MC, Davison GC, Narayanan S (2014) Affective language model adaptation via corpus selection. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing
74. Martin P, Bateson P (2007) Measuring behavior: An introductory guide (3rd ed.). Cambridge University Press.
75. Matsumoto D (1989) Cultural influences on the perception of emotion. Journal of Cross-Cultural Psychology 20(1):92–105
76. Matsumoto D (1990) Cultural similarities and differences in display rules. Motivation and Emotion 14(3):195–214
77. McCowan I, Carletta J, Kraaij W, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V (2005) The AMI meeting corpus. In: Proceedings of the International Conference on Methods and Techniques in Behavioral Research, vol 88
78. Messinger DS, Mahoor MH, Chow SM, Cohn JF (2009) Automated measurement of facial expression in infant-mother interaction: A pilot study. Infancy 14(3):285–305
79. Minato T, Nishio S, Ogawa K, Ishiguro H (2012) Development of cellphone-type tele-operated android. In: Proceedings of the Asia Pacific Conference on Computer Human Interaction, pp 665–666
80. Morency L (2010) Modeling human communication dynamics. IEEE Signal Processing Magazine 27(5):112–116
81. Moschitti A, Chu-Carroll J, Patwardhan S, Fan J, Riccardi G (2011) Using syntactic and semantic structural kernels for classifying definition questions in jeopardy! In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp 712–724
82. Mozer MC (1998) The neural network house: An environment hat adapts to its inhabitants. In: Proceedings of AAAI Spring Symposium on Intelligent Environments, pp 110–114
83. Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2049–2056
84. Nead C (ed) (2012) European Directory of Health Apps 2012-2013. Patent View
85. Nishida T, Nakazawa A, Ohmoto Y, Mohammad Y (2014) Conversational Informatics. Springer
86. Oertel C, Cummins F, Edlund J, Wagner P, Campbell N (2010) D64: A corpus of richly recorded conversational interaction. Journal on Multimodal User Interfaces pp 1–10

87. Oullier O, de Guzman GC, Jantzen KJ, S Kelso JA, Lagarde J (2008) Social coordination dynamics: Measuring human bonding. Social Neuroscience 3(2):178–192
88. Paggio P, Allwood J, Ahlsén E, Jokinen K (2010) The NOMCO multimodal nordic resource - goals and characteristics. In: Proceedings of the Language Resources and Evaluation Conference
89. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359
90. Pantic M, Rothkrantz L (2004) Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34(4):1449–1461
91. Pardo J (2006) On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America 119
92. Pentland A (2008) Honest signals. MIT Press
93. Picard RW (2000) Affective computing. MIT press
94. Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. Behavioral and brain sciences 27(02):169–190
95. Piperidis S (2012) The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In: Proceedings of Language Resources and Evaluation Conference, pp 36–42
96. Potamianos A (2014) Cognitive multimodal processing: From signal to behavior. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 27–34
97. Ramseyer F, Tschacher W (2011) Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. Journal of Consulting and Clinical Psychology 79(3):284 – 295
98. Rapantzikos K, Avrithis Y, Kollias S (2009) Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1454–1461
99. Renals S, Carletta J, Edwards K, Bourlard H, Garner P, Popescu-Belis A, Klakow D, Girenko A, Petukova V, Wacker P, Joscelyne A, Kompis C, Aliwell S, Stevens W, Sabbah Y (2014) ROCKIT: Roadmap for conversational interaction technologies. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 39–42
100. Riccardi G (2014) Towards healthcare personal agents. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 53–56
101. Riccardi G, Hakkani-Tür D (2005) Grounding emotions in human-machine conversational systems. In: Intelligent Technologies for Interactive Entertainment, Lecture Notes in Computer Science, Springer Verlag, pp 144–154
102. Rich C, Sidner CL (2010) Collaborative discourse, engagement and always-on relational agents. In: Proceedings of the AAAI Fall Symposium on Dialog with Robots, vol FS-10-05
103. Richardson D, Dale R (2005) Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science 29(6):1045–1060
104. Richardson M, Marsh K, Isenhower R, Goodman J, Schmidt R (2007) Rocking together: Dynamics of intentional and unintentional interpersonal coordination. Human Movement Science 26(6):867 – 891
105. Rickheit G, Wachsmuth I (2006) Situated Communication, vol 166. Walter de Gruyter
106. Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language pp 696–735
107. Saint-Georges C, Mahdhaoui A, Chetouani M, Cassel RS, Laznik MC, Apicella F, Muratori P, Maestro S, Muratori F, Cohen D (2011) Do parents recognize autistic deviant behavior long before diagnosis? taking into account interaction using computational methods. PLoS ONE 6(7):e22,393
108. Salah A (2014) Natural multimodal interaction with a social robot: What are the premises? In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 43–45
109. Scassellati B, Admoni H, Mataric M (2012) Robots for use in autism research. Annual Review of Biomedical Engineering 14:275–294
110. Scherer S, Glodek M, Schwenker F, Campbell N, Palm G (2012) Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. ACM Transactions on Interactive Intelligent Systems 2(1):4:1–4:31
111. Scherer S, Weibel N, Morency L, Oviatt S (2012) Multimodal prediction of expertise and leadership in learning groups. In: Proceedings of the International Workshop on Multimodal Learning Analytics
112. Scherer S, Hammal Z, Yang Y, Morency L, Cohn J (2014) Dyadic behavior analysis in depression severity assessment interviews. In: Proceedings of the ACM International Conference on Multimodal Interaction
113. Schröder M, Bevacqua E, Cowie R, Eyben F, Gunes H, Heylen D, ter Maat M, McKeown G, Pammi S, Pantic

M, Pelachaud C, Schuller B, de Sevin E, Valstar MF, Wöllmer M (2012) Building autonomous sensitive artificial listeners. IEEE Transactions on Affective Computing 3(2):165–183
114. Schroeder M (2009) Expressive speech synthesis: Past, present, and possible futures. In: Affective Information Processing, Springer, pp 111–126
115. Schuller B, Batliner A (2014) Computational paralinguistics: Emotion, affect, and personality in speech and language processing. John Wiley & Sons
116. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2013) Paralinguistics in speech and language—state-of-the-art and the challenge. Computer Speech & Language 27(1):4–39
117. Shockley K, Santana MV, Fowler CA (2003) Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance (29)
118. Shokoufandeh A, Marsic I, Dickinson SJ (1999) View-based object recognition using saliency maps. Image and Vision Computing 17(5):445–460
119. Tkalčič M, Burnik U, Košir A (2010) Using affective parameters in a content-based recommender system for images. User Modeling and User-Adapted Interaction 20(4):279–311
120. Tomasello M (2008) Origins of human communication. MIT Press
121. Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences 28(05):675–691
122. Town C (2006) Ontological inference for image and video analysis. Machine Vision and Applications 17(2):94–115
123. Turney PD, Littman ML (2003) Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4):315–346
124. Valstar M (2014) Automatic behaviour understanding in medicine. In: Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges, pp 57–60
125. Van Engen KJ, Baese-Berk M, Baker RE, Choi A, Kim M, Bradlow AR (2010) The Wildcat Corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. Language and Speech 53(4):510–540
126. Varni G, Volpe G, Camurri A (2010) A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Transactions on Multimedia 12(6):576 –590
127. Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Transaction on Affective Computing 5(3):273–291
128. Vinciarelli A, Pantic M, Bourlard H (2009) Social Signal Processing: Survey of an emerging domain. Image and Vision Computing 27(12):1743 – 1759
129. Vinciarelli A, Murray-Smith R, Bourlard H (2010) Mobile Social Signal Processing: vision and research issues. In: Proceedings of the ACM International Conference on Mobile Human-Computer Interaction, pp 513–516
130. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schroeder M (2012) Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. IEEE Transactions on Affective Computing 3(1):69–87
131. W-S Chu FZ, la Torre FD (2012) Unsupervised temporal commonality discovery. Proceedings of the European Conference on Computer Vision
132. Wahlster W (ed) (2000) Verbmobil: Foundations of Speech-to-Speech Translation. Springer Verlag