Int J Soc Robot
DOI 10.1007/s12369-013-0179-x

Self-talk Discrimination in Human–Robot Interaction Situations for Supporting Social Awareness

Jade Le Maitre · Mohamed Chetouani

Accepted: 12 January 2013

Abstract Being aware of the presence, activities and is fundamental for Human–Robot Interaction and assistive applications. In this paper, we describe (1) designing triadic situations for cognitive stimulation for elderly users; (2) characterizing social signals that describe social context: system directed speech (SDS) and self-talk (ST); and (3) estimating an interaction efficiency measure that reveals the quality of interaction. The proposed triadic situation is formed by a user, a computer providing cognitive exercises and a robot that provides encouragement and help using verbal and non-verbal signals. The methodology followed to design this situation is presented. Wizard-of-Oz experiments have been performed and analyzed through eye-contact behaviors and dialog acts (SDS and ST). We show that users employ two interaction styles characterized by different prosody features. Automatic recognition systems of these dialog acts is proposed using k-NN, decision tree and SVM classifiers trained with pitch, energy and rhythmic-based features. The best recognition system achieves an accuracy of 71 %, showing that interaction styles can be discriminated on the basis of prosodic features. An Interaction Efficiency (IE) metric is proposed to characterize interaction styles. This metric exploits on-view/off-view discrimination, semantic analysis and ST/SDS discrimination. Experiments on collected data prove the effectiveness of the IE measure in evaluating the individual's quality of interaction of elderly patients during the cognitive stimulation task.

Keywords Social signal processing · Social engagement · Measuring interaction · Prosodic cues

1 Introduction

Interest in service robotics has recently grown partially due to human assisting applications. The proposed robotic systems are designed to address various supports: physical, cognitive or social. Human–Robot Interaction (HRI) plays a major role in these applications, making Socially Assistive Robotics (SAR) [1] a promising field. Indeed, SAR aims to aid patients through social interaction with several applications, including motivation and encouragement during exercises [1–3].

Providing social signals during interaction is continuously performed during human–human interaction [4], and their absence is identified in pathologies such as [5]. Interpreting and generating of social signals allow sustaining and enriching interactions with conversational agents and/or robots. In [6], an early system that realizes the full action–reaction communication cycle by interpreting multimodal user input and generating multimodal agent behaviors is presented. The importance of feedback in regulating interaction has been highlighted in several situations [7, 8].

The ROBADOM project [9] is devoted to designing a robot-based solution for assistive daily living aids: managing shopping lists, meetings, medicines, and appointment reminders. Within the project, we are developing a specific robot to provide verbal and non-verbal help, i.e., encouragement and coaching during cognitive stimulation exercises. Cognitive stimulation is a methodology that alleviates the

This work has been supported by French National Research Agency (ANR) through TecSan program (Robadom project ANR-09-TECS-012).

J. Le Maitre · M. Chetouani (✉)
ISIR UMR 7222, Université Pierre et Marie Curie, Paris, France
e-mail: mohamed.chetouani@upmc.fr

J. Le Maitre
e-mail: lemaitre@isir.upmc.fr

Published online: 26 January 2013

elderly decline in some cognitive functions (e.g., memory, attention) [10]. The robot would be dedicated to Mild Cognitive Impairment (MCI) patients (i.e., a cognitive impairment that is not severe enough to meet the dementia criteria). Cognitive impairment is a major health problem facing elderly people in the new millennium. This refers not only to dementia but also to less severe cognitive impairments associated with a decreased quality of life and, in many cases, progress to dementia.

In this paper, we study patient-robot interaction schemes during cognitive stimulation exercises. The key idea is to propose evaluation metrics for these interaction schemes to further enhance the collaboration. There is a growing interest in evaluating of Human–Robot Interactions [11, 12]. Achieving efficient interaction requires several elements, among which engagement plays a fundamental role. A previous study [13], reports ten challenges related to efficient joint human–agent activity, and entering into a basic compact and being able to detect a partner's intents and actions are identified as research challenges. Successful cooperation also requires a commitment between the partners and general dynamics ([14] presents an overview on interpersonal synchrony). Engagement is a complex phenomenon, and the literature has used numerous terms to describe it. Sidner et al. [15] consider engagement to be the process in which partners establish, maintain and end interactions. Poggi [16] defines engagement as the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing interaction.

Engagement detection is identified as a key element in socially assistive robotics, which usually require more attention to interface design [17, 40]. In this work, we propose studying an important component of engagement: social awareness. This component requires that the robot is able to detect features that are relevant to social interaction: human presence, communicative intents... Quality of Human–Robot interactions should be improved by the introduction of social awareness mechanisms [12, 14]. We investigate engagement, social awareness and the quality of interaction in a triadic framework: user-computer (providing cognitive exercises) -robot (providing encouragements and backchannels). We identify social signals, including system-directed speech and self-talk, as indicators of the user's engagement level during interaction.

Our overall research agenda is to investigate human–robot communication dynamics by proposing a framework including engagement characterization, interpersonal synchrony and affective management. In this paper, we first present a social awareness framework based on the detection of interaction styles by employing eye-contact, semantic and prosodic features. Then, durations of these interactions styles are combined in a metric, Interaction Efficiency capturing the time required to achieve a task. Specifically, our contributions include the following:

- An automatic recognition system to detect both system-directed speech and self-talk. Self-talk provides insights about the involvement of the patient with MCI. Self-talk includes out-of-task utterances. We also propose relevant rhythmic features to characterize speech registers
- The definition and evaluation of a metric characterizing the quality of interaction based on the features of interaction styles. This measure is employed to understand patient strategies during cognitive stimulation exercises.

The remainder of this paper is organized as follows. Section 2 describes the related works in human–human and human–machine interactions. We summarize our approach in Sect. 3. Sections 4 and 5 give an overview of the cognitive stimulation situation including the design of the robot and the Wizard-of-Oz experiment. Section 6 describes the analysis of the manually labelled data for the extraction of dialog acts from the Wizard-of-Oz experiment: self-talk (ST) and system directed speech (SDS). Section 7 shows and discusses the social activity characterization framework. The experimental results are discussed in Sect. 8. Finally, Sect. 9 concludes our work.

2 Related Work

2.1 Social Engagement Cues

The engagement detection problem has been studied using verbal and non-verbal cues, but many existing approaches attempt to estimate engagement using gazes [15, 19–21, 30], considering eye-contact a prominent social signal. Eye-contact is usually employed to regulate communication between humans [27]: initial contact, turn-taking and triggering backchannels.

Mutual gaze has been shown to contribute to smooth turn-taking [21, 22]. Goffman [23] mentioned that eye-contact during interaction tends to signal that each partner agrees to engage in social interaction. Gaze deficiency or failure during interaction may be interpreted as a lack of interest or attention, as Argyle and Cook noticed [24]. In face-to-face communication, initiation, regulation and/or disambiguation can be achieved through eye-gaze behaviors. Interaction efficiency is based on the ability of shifting roles, which is again possible in eye-gaze behaviors [25, 26]. During interaction, gaze might be combined with speech. Kendon [27] analyzed these situations and identified how speakers look away from their partners at the beginning of an utterance and look at their partners at the end of an utterance. This procedure might be useful, as it serves to avoid cognitive load (i.e., utterance planning) and shift roles with the partner.

In HRI situations, robots are required to estimate the engagement of addressee for efficient communication. Gaze

estimation of gaze is a difficult task in HRI due to greater distances between the robot and addressee; other cues, including head orientation, body posture, and pointing, might also be used to indicate the direction of attention. Most proposed techniques are based on the face engagement concept proposed by Goffman [23] to describe the process through which people employ eye-contact, gaze and facial gestures to interact or engage with each other. The engagement detection framework is based on (1) face detection and (2) facial/head gestures classification. In a previous study [28], the authors proposed combining multiple cues to understand the behaviors of the potential addressee in the human–robot interaction task. A set of utterances is defined and used to start an interaction. In addition to the detection step, the authors estimated the visual focus of user attention. They computed the probability that the partner is looking at a pre-defined list of possible focus targets. Because the focus targets include the robot itself and other potential users, engagement estimation is reinforced and allows benefitting from the eye-gaze functions without an explicit modeling. A similar work [19] on a multi-robot interaction framework, based on face detection and gestures classification, allowed selecting and commanding individual robots.

Robots could also use eye-gaze behaviors to improve of the interactions. Mutlu et al. [29] conducted experiments where their robot, Robovie, employed various eye-gaze behavior strategies to signal specific interaction roles: addressee, bystander, and overhearer. The authors show that gaze direction serves as a moderator, as gaze cues support the conversation by reinforcing human subjects' roles and participation.

The abovementioned works have shown the benefits of using eye-gaze behaviors to measure engagement during interaction. However, this social signal should be more precisely characterized. Rich et al. [30] first investigated relevant cues for recognizing engagement in human–human interaction and proposed an automatic system for HRI situations. Regarding eye-gaze behaviors, they identified directed and mutual facial gaze. Directed gaze characterizes events when one person looks at some object, after which the other person looks. Mutual gaze refers to events when one person looks at the other person's face. Various features are thus employed to describe communicative functions of eye-gaze behaviors. Ishii et al. [21] also extract various features from eye-gaze behaviors, including gaze transitions, mutual gaze occurrences, gaze duration, distance of eye movement, and pupil size. These features are employed to predict users' conversational engagement. The statistical modeling and feature combinations have shown the features' relevance in recognizing users' attitudes (engagement vs. disengagement).

Other cues can be employed to detect engagement in social interactions. Castellano et al. [20] proposed combining eye contact with smiling, which their game scenario considers an engagement indicator. The authors enriched the characterization by adding contextual features, such as the game state and behavior of the robot used (iCat facial expressions). A Bayesian network is employed to model the cause-effect relationships between the social signals and contextual features. The evaluation allows identifying a set of actions performed by users correlated with engagement. Spatial movements have been used to initiate conversations [31] and more generally for engagement characterization ([32] presents an interesting discussion on relevant social cues).

These results show that eye-gaze behaviors could be combined with others cues in various tasks to improve detection rates. Other strategies can be followed by not estimating eye-gaze behaviors. A previous study [33] employed physiological signals, including skin response and temperature, to estimate engagement. The work was undertaken because extracting implicit information on the patient has seminal importance in the rehabilitative context. Physiological signals are more correlated to a user's internal state and can thus be employed to infer emotional information [34]. A robot can use the engagement detection from these signals, e.g., coaching or assistance, to alter the interaction scenario.

Peters et al. [35] highlighted the various elements related to engagement. They proposed a simplified model based on an action-cognition-perception loop, which makes it possible to differentiate between several aspects of engagement: perception (e.g., detecting cues), cognition (e.g., internal state: motivation), action (e.g., display interest). They also identified a dimension called experience, which covers subjective experiences felt by individuals. This work shows that engagement is not a simple concept, and, as with other social signals, characterization, detection and understanding must still be investigated to design adaptive interfaces such as robots.

In this paper, we consider the characterization of the user's engagement level by the evaluation of social activities (interaction styles) in the assistive context, specifically cognitive stimulation situations. Research works on related topics are usually devoted to acceptability [36–38]; maintaining engagement is, however, a key component of socially assistive robots [39, 40]. Xiao et al. [41] have investigated how seniors deal with multimodal interfaces and they found that elderly people require more time and make errors, as expected. More interestingly, they demonstrated that older users employ an audible speech register called self-talk, which is a type of think-aloud process produced during difficult tasks. Our rationale for expecting self-talk as an indicator of social activity is motivated because producing self-talk would confuse robots, which are generally based on spoken dialog systems [42, 57]. Discriminating self-talk (ST) from system/robot-directed speech (SDS) holds great importance for two main reasons: (1) intentional interactions

with the agent (robot/computer) are produced during SDS, and (2) ST production can be continuously evaluated to provide insights on out-of-task situations (e.g, self-regulation), thus allowing socially aware machine designs. The next section aims to describe more precisely this specific speech register.

2.2 Self-talk in Machine Interaction

Following Oppermann [18], self-talk, or private speech, refers to the audible or visible talk people use to communicate with themselves. This register can be considered part of off-talk, which is a special dialog act characterizing "every utterance that is not directed to the system as a question, a feedback utterance or an instruction". Off-talk is a problem for automatic speech recognition systems, and distinguishing it from on-talk (or system-directed speech) will clearly improve recognition rates. However, its characteristics make the task difficult. If a user is reading instructions, lexical information is not discriminant and other features should be employed. One relevant strategy is to try to combine audio and visual features [43, 44]. Batliner et al. formulated the problem by defining on-talk vs. off-talk and on-view vs. off-view strategies. Combining them leads to on-focus (on-talk + on-view) and off-focus, where on-view is not discriminant, i.e., listening to someone and looking away. The authors employ an audio-visual framework to classify on-talk and various off-talk elements (e.g., read, paraphrasing and spontaneous off-talk). Prosodic, part-of-speech (POS) features and visual features (a simple face detection system) are employed. Detecting a user's focus on interaction yields 76.6 % using prosodic features, and combining it with linguistic and visual features achieves 80.8 % and 84.5 %, respectively.

From a conceptual perspective, open-talk (robot-directed speech) is considered social speech because it is produced with the objective of communication, while self-talk is a means for thinking, planning and behavioral self-regulation [45]. Lunsford et al. [44] investigated audio-visual cues and reviewed some self-talk functions. Among the most interesting functions, the authors reported that self-talk supports task performance and self-regulation. When seniors completed a spatial task, a high amount of self-talk was observed in [70]: 80 % of the subjects in their study engaged in ST at some point during their session. This amount increases with task difficulty, which has strong correlation with the cognitive load of the person: the ST amount increased from low to high difficulty tasks (26.9 % versus 43.7 %, respectively).

As cognitive stimulation experiments requiring explicit dialog were completed, as in our research (Sect. 4), the user's verbal production is considered a unique source of information on the user's engagement level. However, the amount of conversations does not totally reflect the quality of interaction, and a characterization of how these utterances are produced may provide insight into the communicative functions employed. This paper focuses on one aspect of this characterization by considering self-talk as out-of-task utterances [46]. The proportion of SDS and ST are exploited to identify phases when the patient is intentionally communicated with the system. A global metric reflecting the durations of these interaction styles (directed or not) is proposed. Although, each interaction style may be the support of various communicative functions (examples: direct attention for SDS and thinking-aloud for ST), and a characterization of these functions is beyond the scope of this paper. However, by being aware of these interaction phases, the coach robot can produce useful feedback, encouragement or help. Potential benefits of improving social awareness include (1) designing adaptive social interfaces (including robots), (2) improving the impact of assistive devices and (3) understanding individuals' strategies and behaviors.

3 Overview of Our Approach

The purpose of this paper is the introduction of social awareness mechanisms for the overall improvement of interactions between an MCI patient and a robot using verbal utterance classification as Self-Talk or System-Directed Speech. To analyze the role of the aforementioned speech registers, we collected two datasets with the following situations: patient-therapist and patient-robot. The first set includes audio-visual recordings of a patient completing cognitive exercises on a computer while being coached by a therapist. Manual corpus analysis allows identifying interactive patterns, mainly the therapist's behaviors (e.g., encouraging or answering questions) (Sect. 4), which are then employed to develop socially relevant coach-robot behaviors. The second set includes patients completing cognitive stimulation exercises with the help of a controlled robot during a Wizard-of-Oz experiment. From the recordings, we automatically extracted non-verbal cues to characterize individual participants: eye-contact (on-view vs. off-view), prosodic and rhythmic features and dialog acts studied using the Latent Semantic Analysis (LSA) method. This method groups semantically similar keywords into meaningful clusters, structuring dialog acts in a non-supervised way, and giving a semantic signification.

Another manual annotation occurs in these clusters, splitting keywords into two categories: whether they were spoken to the robot and/or computer (System-Directed-Speech) or to the patient himself (Self-Talk). We developed methods to automatically detect self-talk and with all labeled keywords (cluster, ST or SDS), we can perform social activity characterization, as described in Sect. 7. Figure 1 details the stages of our work, which are described below.

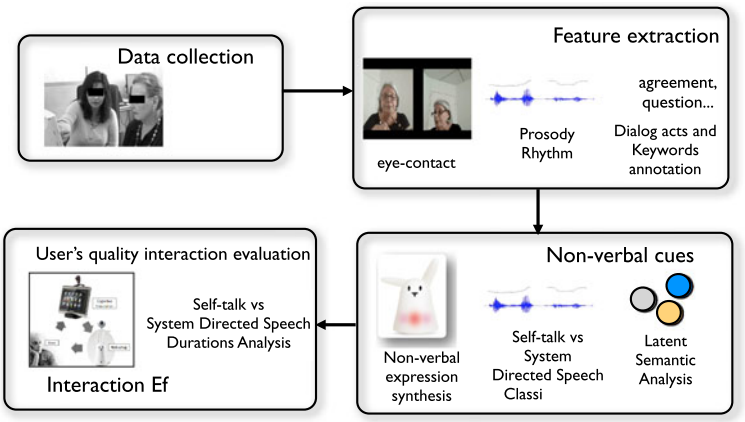

Fig. 1 Visualization of our approach for social activity characterization

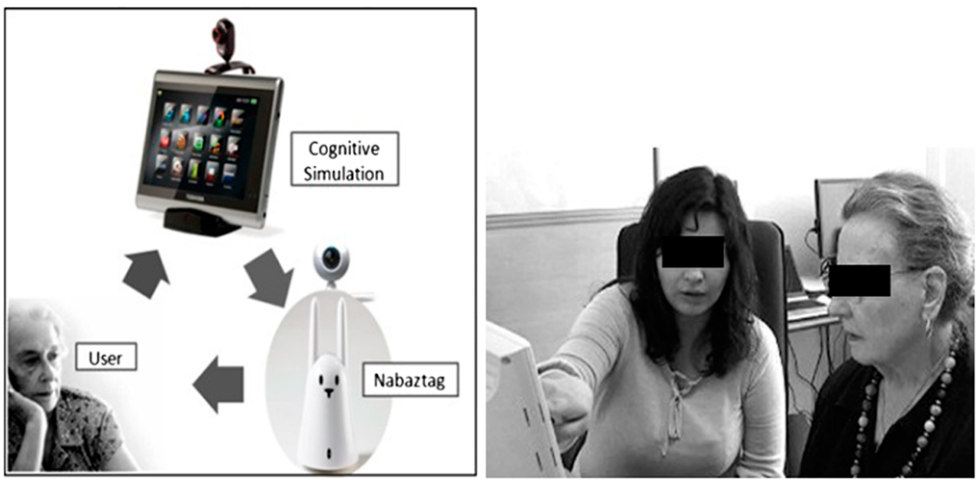

Fig. 2 Triadic situation either with robot or therapist

4 Patient–Therapist Interaction

Social interactions are by definition dynamic [14], and we are interested by the social cues that allow evaluating and sustaining interactions. Traditional cognitive stimulation sessions were thus recorded. The patient had to solve exercises on a tactile screen, while the therapist, seated near the patient, observed the situation and provided help whenever the patient needed it. The therapist could help with the technical setup, i.e., indicating how to deal with the tactile screen, or provide help for a particular exercise, i.e., how to correct an answer or say whether the patient answered correctly. Backchannel feedback (i.e., head nods or sounds such as "uh-huh") [49] from the therapist is important for patients to gain confidence and sustain the interaction. Backchannels are signals directed to the speaker, indicating that the communication is working and they should continue speaking. Machine learning techniques can automatically predict these social signals, which were previously optimized on face-to-face data [8].

The interaction between the patient, therapist and cognitive stimulation exercise is a triadic situation, as shown in Fig. 2. Psychologists at the Broca Hospital (Paris, France) thus organized recorded cognitive stimulation sessions, which led to our analysis of the interactions between patient and therapist. Thanks to these sessions, we determined the interaction phases of the therapist and later duplicated them for the robot. The therapist is always at the patient's side, measuring and evaluating his attention and engagement. The presence of the therapist is important for the patient to gain confidence. As Fig. 2 shows, the patient sits in front of the tactile screen with the therapist at his right during the cognitive stimulation exercises. This triadic configuration is kept for the human–robot interaction, because the current robot is designed as a companion that allows stimulating joint attention situations. Because robot acceptance is our second concern, after appropriate robot reactions, we and our colleagues investigated which type of robot the targeted end-users with Mild-Cognitive Impairments might accept.

5 Patient–Robot Interaction

5.1 Designing Robot for MCI Patients

Focus group sessions were conducted at the Broca Hospital to identify how the elderly perceive a robot's appearance. There, fifteen adults over the age of sixty-five, divided in three groups, took part in the sessions, thirteen of whom were recruited from the Memory Clinic at the Broca Hospital (Paris, France); two were recruited from an association for the elderly. Seven participants had been diagnosed with Mild Cognitive Impairment, according to previous criteria [47]. The focus group results indicate that the robots considered attractive were small robots, often with a modern design, shaped like animals or objects they could use in their daily life [48], such as Mamoru or Paro. Various robots that fit these characteristics can be employed, but we selected the rabbit shaped Violets Nabaztag, type Nabaztag: tag. This electronic device has enabled Wi-Fi and can connect to the internet to process specific services using a distant server, located at http://www.nabaztag.com. The Nabaztag has motorized ears, 4 color LEDs, a speaker and a microphone. As described in Sect. 5.2, ears and LEDs are employed to enhance the expressiveness of the Nabaztag. Regarding robot acceptability, experimental results can be found in other projects, such as SERA (Social Engagement with Robots and Agents), where social engagement is investigated [36, 37].

5.2 Description of the Wizard-of-Oz Experiments

5.2.1 Technical and Experimental Design

Human–robot communication differs from human–human communication. To gather reliable information about human–robot communication, it is thus important to observe human behavior in a situation in which humans believe they are interacting with a real robotic system. The user should think that he or she is communicating with the system, not a human [69]. Our Wizard-of-Oz experiments aim to record interactions between the patient and robot, using the interaction schemes observed between the patient and therapist. After analyzing videos of the sessions between the therapist and patient, interaction patterns were detected (e.g., *therapist encouraging, therapist answering a question, backchannels*) and adapted for the Nabazatg. The purpose was to give the Nabaztag an interaction panel, leaving the Wizard the responsibility of choosing the right pair [*answer+backchannel*] for each situation.

The patient sits in front of the tablet-PC, with the Nabaztag at his right as shown in Fig. 2. During a sequence of cognitive exercises solved on the tablet PC, the robot interacts with the patient. The Wizard gathers information about the situation using two cameras and a screen capture of the tactile screen. The Wizard can hear the patient, but the reversed situation is impossible. The Wizard remotely controls the Nabaztag, activating the pair [*answer+backchannel*] at the same time. The total duration of the WOZ experiments is 96 min, and the mean session duration was 7 min 30 s, though the individual sessions varied.

5.2.2 Verbal and Nonverbal Behaviors for the Nabaztag

Nonverbal behaviors should be defined for the Nabaztag. As with other robots, including Paro [50], Emotirob [51] or Aibo [8], the Nabaztag can exploit movements and sounds as social communicative signals. According to Lee and Nam [52], concerning the relationship between physical movement and emotional interaction augmentation, the expressions of the Nabaztag are correlated with both movement speed and LEDs blinking. As Fig. 3 shows, slow movements or blinking LEDs express unpleasant expressions, such as sadness or annoyance, when the user is getting lost or does not know how to solve the exercise. Positive expressions are related to active movements and blinking, which are employed to encourage the user.

The nonverbal behaviors for the Nabaztag were implemented using the NabazFrame, developed by the University of Bretagne Sud.[1] Nonverbal choreographies contain ear movements and different sets of blinking LEDs, where the color depends on the mood or feedback to transmit. Green or yellow fast blinking LEDs express pleasant expressions, while slow movements with blue or violet LEDs express unpleasant signals (Fig. 3). End-users are currently testing these behaviors in another work [48].

5.3 Participants Analysis

Therapists at the Broca Hospital selected eight total participants, seven females and one male, aged 64 to 82, to participate in the experiments. Two participants had a slight MCI. All the users didn't have any communication, hearing, or vision impairments. Sessions have been completed in French. Transcriptions and translations of interactions between the participants and the Nabaztag are shown in the following two paragraphs, one dedicated to the SDS, the second to ST. Clearly these two speech registers are useful for the user. However, correct interpretations by the robot are only possible by the introduction of social awareness mechanisms (SDS vs ST discrimination).

Example 1 Dialog Between a User and the Nabaztag: System-Directed-Speech

[1] www-valoria.univ-ubs.fr.

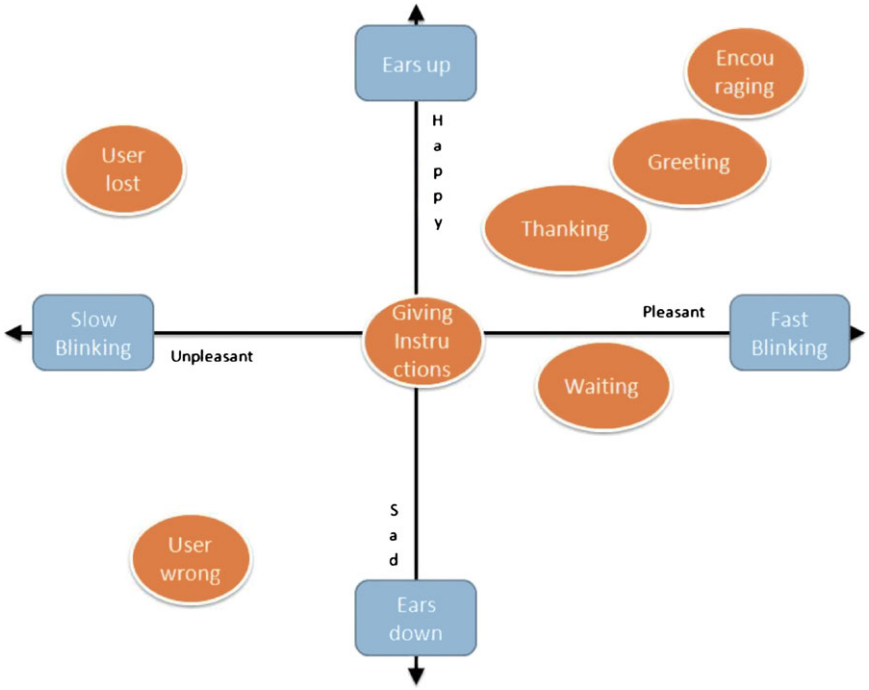

Fig. 3 Relationship between movements and expressions

Nab.: Good Morning, my name is Carole, what's your name?
User: My name is Bob.
Nab.: Hello, Bob. I'm here to help you to solve your exercises. Do you want to start them now?
User: Yes!
Nab.: Let's go! First, you have to drag the images into the box corresponding to their name.
User: How do I do that?
Nab.: Press on the image, then drag it to the box.

Example 2 Dialog Between a User and the Nabaztag: Self-Talk

User: And I drag the little image with the tree to autumn box... oh, it's not coming! It's gone, uh, found, I drag it to the box, oh, it escaped...
Nab.: When dragging the image to the correct box, you must always touch the screen.
User: Uh, yes, I drag it, I drag the image, the tree is in the box. The image with the flowers can't be the summer, these flowers grow in the spring, oh, I don't know, I drag the image to the spring, It's not coming, ah, yes, let's go to the next image.

6 Analysis of the WOZ Experiment Annotations

6.1 Manual Annotation of the WOZ Experiment Content

For each participant in the recorded experiments, the first step was to manually annotate the videos taken during the Wizard-of-Oz experiments. Before the annotation, videos from the two cameras were synchronized and edited together, so the annotator could view the patient from the two different angles: computer and robot.

First, the gaze was annotated without paying attention to the speech. The annotator carefully annotated when the patient looked at the robot (the setup made clearly visible when the patient was looking at the robot: from the computer's perspective, the person moves her head, and from the robot's perspective, the person's gaze is focused on the camera) and when the patient looked directly at the computer (same perspectives, but reversed). The second step was speech annotation.

The annotator listened to and annotated the relevant keywords found in the patient's utterances. Keywords can be a single word or an expression of words with close signification: "*I'm doing well*" is annotated as one keyword. Filled pauses were not annotated, as the dialog system processed them. The spoken keywords were primarily divided into eight simple different categories, defined by the annotator after seeing all eight films. A general structure is proposed.

- **Agreement** "*Let's start*", "*Yes*", "*You're right*".
- **Technical Question** (often about using of the tactile screen).
- **Contextual Question** (about the cognitive exercise itself, i.e., "*What should I do with this picture?*", "*Should I put it here, or there?*").
- **Non-Obligatory Turn-Taking**: Comment without requiring an answer from the robot (Users with MCI of-

ten comment about what they are doing, "*And I move the picture over there…*").
- **Obligatory Turn-Taking**: Comment requiring an answer from the robot (i.e., "*Tell me, this is right*"), which indicates the difficulty felt by the user.
- **Support Needed:** User is confused and needs more support (e.g., "*I don't remember*", "*I don't know what I should do?*").
- **Thanks**: (a user thanks the Nabaztag when they receive help; one complimented it about its color and shape).
- **Disagreement** ("*No, I don't want*").

The gaze annotation shows that users look at the computer 89.76 % of the time, because they were asked to complete the exercises. Other explanations could be found in the specific design of the triadic situation, as eye contact with the robot is not required for effective engagement: the robot is only here to help and encourage, but the patient must gaze at the computer to solve an exercise. As previously discussed, eye-gaze behaviors might not be discriminant for engagement detection on some tasks, especially for seniors. Similar results have been obtained in [44], where 99.5 % of the self-talk utterances were associated with a gaze behavior directed to the system, and 98.1 % for system-directed speech. Eye contact is not always discriminant for social activity detection, and evaluation using dialog acts as if self- and open-talk are relevant. Eye contact is not discriminant in our case, but we found that patients' verbal production could provide insight into their difficulties. Because the robot needs to know when the patient encounters difficulties to produce the appropriate feedback and help the patient, we focused our work on verbal utterances.

6.2 Latent Semantic Analysis of the Annotation Content

Understanding the annotation content in interactive databases is made difficult by individuals' strategies in a cognitive stimulation task. Indeed, several keywords or utterances correspond to one of the height categories. To provide insights on the annotation databases and thus the exhibited behaviors, we performed a latent semantic analysis (LSA) [53]. LSA is based on a Singular Value Decomposition (SVD) of the term-document matrix and results in a reduced dimension of the feature space. Interpreting the reduced space allows identifying semantic concepts. Although LSA is usually performed on texts, we decided to use it here because the patients' verbal production is not spontaneous. Verbal utterances are structured (the exercise or technical difficulties). Even in the ST parts of the interaction, verbal production is structured and limited. The term-by-documents matrix becomes, in our case, a keyword-by-clusters matrix. To train LSA, we estimated the occurrence of each of the 247 different keywords or utterances produced by the participants in the height categories. Using a Kaiser criterion (80 % of the information), the reduced rank was set to five. We carefully analyzed the content (keywords) of the obtained clusters, and they were interpreted thus:

Table 1 Distribution of self-talk and system-directed speech over the clusters

Semantic cluster	Self-talk	System-directed speech
Positive feeling	91	9
Comments	21	24
Social etiquette	43	69
Request information	27	33
Others	130	96

- **Positive Feeling**: positive feeling expressed towards the robot or exercises.
- **Comments**: useful keywords, expressing nothing more than a simple statement about the exercise.
- **Social Etiquette**: the keywords in this cluster all expressed agreement with the robot ("*Okay*", "*I'm in*", "*Thanks*"). This cluster is based on Light's original works, analyzing communication and characterizing communication messages into different categories [71].
- **Request Information**: all keywords expressing a question on the exercises were put into this cluster. The questions are context-based and depend on the part of the exercises in which the patient experiences difficulties.
- **Others**: the cluster in which words are relevant but the amount of words is too small to form a specific cluster. This last cluster covers comments useful for the cognitive exercise but totally relevant for engagement investigations due to the amount of self-talk (see Table 1).

6.3 Speech Corpus

Our speech corpus contains 543 utterances or keywords, and each corresponds to a semantic cluster. Each utterance or keyword has been carefully annotated: (1) self-talk or (2) system/robot-directed speech. The annotator simply answered these two questions: "*According to you, does the person speak to herself?*" (ST) or "*speak to the robot or the computer?*" (SDS).

We evaluated the subjectivity of the annotation process by evaluating inter-judge agreement. We chose a second naive annotator who had no contact with the participants and did not previously know the differences between ST and SDS. After watching the videos, the annotator was requested to annotate the verbal utterances as ST or SDS. We then performed an inter-annotator score between the first labeling and that of the naive annotator. Using the kappa method, the result was a score of 0.68, showing a sustainable agreement.

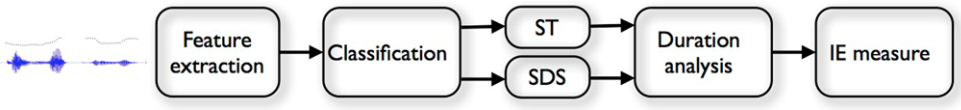

Fig. 4 Description of the proposed system for social awareness measure

Table 2 Number of self-talk and system-directed speech

Users	Self-talk	System-directed speech
1	20	19
2	1	7
3	106	85
4	14	2
5	30	6
6	37	20
7	58	37
8	49	55

Table 1 gives the distribution of utterances over the clusters. As expected, system-directed speech is more present on semantic clusters characterized by a direct relationship to the cognitive exercise: comments, social etiquette and information requests. These clusters express direct speech to the robot because the person is asking or requesting something precise; for a discussion between two people, the patient instinctively directs his speech to the system.

Most self-talk verbalizations are present in the most heterogeneous semantic cluster (Others), reflecting the semantic variability of this speech register. SDS and ST are the support of various communicative functions and semantic information is not enough of a discriminant for self-talk detection. A positive feeling mostly contains self-talk elements, because these patients are not used to interacting with robots, and they provide comments that are considered out-of-task utterances, i.e., self-talk.

In this paper, we investigate global behaviors during the overall interaction. But, individual evolution of amount of both SDS and ST are possible during the cognitive stimulation depending on several parameters among them task difficulty, user's skill and motivation but and it may also be impacted by pathology.

Table 2 details the distribution of self-talk and system-directed speech utterances. Our research is motivated by the goal of designing metrics that capture individual behaviors. We expect that automatic ST and SDS characterization will provide insight into these behaviors.

After an acoustic analysis of all keywords and utterances annotated as self-talk or system-directed speech, we selected 293 utterances for ST and 223 for SDS. The removed utterances were mainly due to their shorter duration (less than 1 s). Utterance durations are between 1 and 2.5 s.

7 Using Self-Talk Detection for Social Activity Characterization

In this section, we describe the system developed to characterize social activities by detecting out-of-task utterances (i.e., self-talk). Figure 4 depicts the proposed system. It requires discriminating self-talk from system-directed speech. We then combine the duration of both acting dialogs in an social awareness measure, which aims to characterize the degree of interaction efficiency.

As mentioned above, self-talk is produced with a lower intensity than open-talk or robot-directed speech. In this paper, we propose investigating prosodic features, including pitch, energy and rhythm, because most classification frameworks of specific speech registers, such as emotions [54], infant-directed speech [55], or robot-directed speech [56], are mainly based on supra-segmental feature characterizations. Batliner et al. [57] have also shown the relevance of prosodic and duration features to discriminate on-talk from read off-talk. However, researchers on speech characterization and feature extraction show that is difficult to reach a consensus on relevant features to characterize emotions, intentions, and personality.

7.1 Feature Extraction

Several studies have shown the relevance of fundamental frequency (F0) and energy-based features for emotion recognition applications [58]. F0 and energy were estimated every 20 ms using Praat [59], and we computed several statistics for each voiced segment (segment-based method) [60]: maximum, minimum, mean, standard deviation, interquartile range, mean absolute deviation, and quantiles corresponding to the cumulative probability values 0.25 and 0.75, resulting in a 16 dimensional vector.

Rhythmic features were obtained by applying a set of perceptual filters to the audio signal dedicated to characterizing prominent events in speech, called p-centers [61, 62]. We then estimated the spectrum of the prominent signal to characterize the speaking rate. We estimated 3 features: mean frequency, entropy and barycenter of the spectrum. Differences in rhythm are indicators of efficiency and clarity in delivering speech (fluency) [63].

7.2 Classification

After extracting features, supervised classification is used to categorize data into classes corresponding to (1) self-talk or

(2) system/robot-directed speech. This can be implemented using standard machine learning methods. This study investigated three different classifiers: decision trees, k-nearest neighbor (k-NN) (k is experimentally set to 3) [64], and Support Vector Machines (SVM) with Radial Basis Function [65].

A k-fold cross-validation scheme was used for the experimental setup (k is set to 10), and the performance is expressed using of the overall accuracy: average of the accuracies obtained over the k partitions.

7.3 Evaluating Social Awareness

Engagement is an interactive process [15] in which two participants decide when to establish, maintain and end their connection. Due to the subjective nature of engagement, manual annotation of the engagement level is usually required to train classifiers [66]. In this paper, indicators of engagement and disengagement are restricted to verbal productions of the user characterized by SDS and ST reflecting two different interaction styles: directed to the system or self directed. We propose a global evaluation of the interaction styles using a metric able to reflect individual behaviors.

After detecting SDS and ST utterances, we propose combining them to estimate a dimension called interaction efficiency (IE). Interaction efficiency is a unitless measure of the time needed to complete the task [67]. Depending on the task, IE can be more or less difficult to measure. In our application, where interaction is conducted through dialog, we argued that IE is related to the quantity of robot/system-directed speech, which is characterized by (1) intentional communication towards the system and (2) on-view state (gazing to the system). Self-talk, conversely, indicates out-of-task situations resulting from various factors: task difficulty, cognitive load [70]. We propose estimating the IE of a given human robot interaction during a cognitive stimulation situation using the following expression:

$$IE = \frac{SDS}{SDS + ST} \tag{1}$$

SDS and ST refer to the duration of system-directed speech and self-talk (in seconds). The numerator is the amount of time of intentional interaction, and the denominator is the amount of time of interaction time (speaking). IE is a unitless measure ($0 \leq IE \leq 1$). If SDS is small relative to ST, IE is also small. The unique information source is the verbal productions; consequently, the IE metric can only characterize dialog acts by considering them as intentional (SDS) or out-of-task (ST). Self-talk can positively affect task completion, depending on the individuals. Future works will investigate the relationship between performance and verbal production.

The IE measure proposed in this paper allows evaluating IE by capturing the ratio between system-directed and self-directed social activities. In future works, this measure will allow changing the verbal and non-verbal behaviors of the robot and adapting both the cognitive exercises and encouragements provided.

Table 3 Accuracy of classifiers using 10 folds validation

Features	Decision tree	k-NN	SVM
Pitch-based	49.8 %	53.35 %	52.16 %
Energy-based	55.54	54.29 %	59.51 %
Rhythm-based	52.78 %	56.58 %	56.97 %
Pitch + Energy	57.42 %	59.28 %	64.31 %
Pitch + Energy + Rhythm	55.46 %	58.20 %	71.62 %

8 Experimental Results

This section describes the experiments and subsequent results for characterizing social activity based on self-talk detection. We first present the performance of our self-talk detection system; we then propose to derive a social awareness measure.

8.1 Self-talk Detection

Table 3 classifiers trained on different feature sets. Energy discriminates system-directed speech from self-talk [44]. Compared to pitch-based features, classifiers trained with energy are more efficient, and the best results are obtained using SVM. One possible explanation is that extracting pitch might be more complex for self-talk, which users produce for themselves, and thus has less energy and intelligibility.

In addition to the literature about energy, we argue that rhythm should be relevant to our application because of the change in speaking rates observed during self-talk. The experimental results show that using only rhythm-based features achieves acceptable performance (56.97 %), which is between those obtained by energy (59.51 %) and pitch features (52.16 %). Energy and rhythmic features are more robust. Rhythmic features are related to the vocalic energy of signals [68] and have similar characteristics to short-term energy; perceptual filters, however, are employed before computing energy (from acoustic prominence enhancement). An SVM classifier trained on three feature sets outperforms (71.62 %) all configurations. Adding features does not exhibit the same performance for all classifiers. Adding features for the decision tree and k-NN classifiers decreases performance.

8.2 Social Activity Characterization

This section describes estimating the IE measure (Eq. (1)). To evaluate only the IE measure, we exploited the manually

Table 4 Interaction efficiency (IE) measure estimation

Users	From annotation	From automatic self-talk detection
1	0.5	0.62
2	0.83	0.78
3	0.45	0.53
4	0.13	0.08
5	0.20	0.26
6	0.43	0.46
7	0.42	0.38
8	0.57	0.63

labeled data; Table 4 presents the results. The best IE measure is obtained for user 2 (0.83), but one should be careful with this result, as he produced only 1 self-talk and 7 system directed speech utterances (Table 2). For the most talkative user (patient 3), the IE measure provides insights about his behavior: a relative balance between ST and SDS.

Patients' 4 and 5 IE estimations are under 0.20, because these patients did not talk directly to the robot. They addressed the system directly at the beginning of the exercise, showing they understood the instructions. The other verbal utterances were only comments addressed to themselves. Patient 5 had few difficulties that required her to address directly to the robot for an answer, but the number of her self-talk utterances was too considerable to balance these SDS utterances.

For automatic estimation, we followed the framework described in Fig. 4. A Vocal Activity Detector (VAD), suitable for real-time detection in robotics [8], is employed for speech segmentation. The self-talk/system-directed speech discrimination system is based on the SVM classifier trained on pitch, energy and rhythmic features, as previously designed. After classifying speech utterances, we extracted their duration and estimated the IE measure at the end of the experiment. Table 4 shows that the IE measures computed using the automatic approach capture each user's individual behaviors. High and low IE measures are also efficiently characterized. However, for low IE measures (i.e., user 4), the automatic approach underestimated the performance. This may be due to factors such as the small number of verbalizations from this user. Due to imperfect classifications, some IE measures are over- or underestimated but do characterize a trend.

9 Conclusions

In this paper, we have demonstrated promising results in automatically estimating the interaction styles of patients during a coaching experiment with cognitive stimulation exercises. After analyzing WOZ experiments (triadic situation), we identified relevant social signals characterizing the social activity (directed or not to the system) of elderly patients: speech-directed talk and self-talk. Users employ the latter during difficult tasks for planning, thinking and self-regulation, and it is an indicator of out-of-task situations. We have shown that prosodic features are discriminant. We proposed a system to automatically detect these two social signals based on extracted relevant features: pitch, energy and rhythm, using three different classifiers. The experimental results have shown the discriminative function of energy, as described in the literature.

Future work in this area should investigate multimodal cues for extensions to off-talk situations. Off-talk is the act of not speaking to an addressee, and it includes self-talk and talking to a third addressee. In this case, automatically detecting on-view states could have great importance. Among the automatic cues that should be developed, eye-gaze social signal detection remains a challenge. Metrics of quality of interaction can also include touching and/or manipulation, and a more general definition of multimodal and integrative engagement characterization should be proposed.

Furthermore, we intend to use the IE measure to characterize users and give the opportunity to adapt a robot's behaviors. In our specific task, this includes encouragements and potentially the cognitive exercise difficulty. In future work, we will exploit questionnaires to understand and estimate users' engagement awareness during interaction (engagement experience).

Acknowledgements The authors would like to thank the Broca hospital for their work: Ya-Huei Wu, Christine Fassert, Victoria Cristancho-Lacroix and Anne-Sophie Rigaud.

References

1. Feil-Seifer DJ, Mataric MJ (2005) Defining socially assistive robotics. In: International conference on rehabilitation robotics, Chicago, IL, pp 465–468
2. Fasola J, Mataric MJ (2010) Robot exercise instructor: a socially assistive robot system to monitor and encourage physical exercise for the elderly. In: 19th IEEE international symposium in robot and human interactive communication (Ro-Man 2010), Viareggio, Italy
3. Mataric MJ, Tapus A, Winstein CJ, Eriksson J (2009) Socially assistive robotics for stroke and mild TBI rehabilitation. In: Gaggioli A, Keshner EA, (Tamar) Weiss PL, Riva G (eds) Advanced technologies in rehabilitation. IOS Press, Amsterdam, pp 249–262.
4. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vis Comput J 27(12):1743–1759
5. Saint-Georges C, Cassel RS, Cohen D, Chetouani M, Laznik M-C, Maestro S, Muratori F (2010) What studies of family home movies can teach us about autistic infants: a literature review. Res Autism Spectr Disord 4(3):355–366
6. Cassell J, Bickmore J, Billinghurst M, Campbell L, Chang K, Vilhjàlmsson H, Yan H (1999) Embodiment in conversational interfaces: rea. In: CHI'99, Pittsburgh, pp 520–527

7. Wrede B, Kopp S, Rohlfing K, Lohse M, Muhl C (2010) Appropriate feedback in asymmetric interactions. J Pragmat 42(9):2369–2384
8. Al Moubayed S, Baklouti M, Chetouani M, Dutoit T, Mahdhaoui A, Martin J-C, Ondas S, Pelachaud C, Urbain J, Yilmaz M (2009) Generating robot/agent backchannels during a storytelling experiment. In: ICRA'09, IEEE international conference on robotics and automation, Kobe, Japan
9. Chetouani M, Wu YH, Jost C, Le Pevedic B, Fassert C, Cristancho-Lacroix V, Lassiaille S, Granata Tapus A, Duhaut D, Rigaud AS (2010) Cognitive services for elderly people: the ROBADOM project. In: ECCE 2010 workshop: robots that care, European conference on cognitive ergonomics
10. Yanguas J, Buiza C, Etxeberria I, Urdaneta E, Galdona N, González MF (2008) Effectiveness of a non-pharmacological cognitive intervention on elderly factorial analysis of Donostia Longitudinal Study. Adv Gerontol 3:30–41
11. Young J, Sung JY, Voida A, Sharlin E, Igarashi T, Christensen H, Grinter R (2011) Evaluating human-robot interaction. Int J Soc Robot 3(1):53–67
12. Sciutti A, Bisio A, Nori F, Metta G, Fadiga L, Pozzo T, Sandini G (2012) Measuring human-robot interaction through motor resonance. Int J Soc Robot 4(3):223–234
13. Klein G, Woods DD, Bradshaw JM, Hoffman RR, Feltovich PJ (2004) Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intell Syst 10(6):91–95
14. Delaherche E, Chetouani M, Mahdhaoui M, Saint-Georges C, Viaux S, Cohen D (2012) Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans Affect Comput 3(3):34–365
15. Sidner CL, Kidd CD, Lee C, Lesh N (2004) Where to look: a study of human-robot engagement. In: Proceedings of the 9th international conference on intelligent user interfaces (IUI'04)
16. Poggi I (2007) Mind, hands, face and body. A goal and belief view of multimodal communication. Weidler, Berlin
17. Kulyukin V (2006) On natural language dialog with assistive robots. In: Proceedings of the 2006 ACM conference on human-robot interaction (HRI 2006), Salt Lake City, Utah, pp 164–171
18. Oppermann D, Schiel F, Steininger S, Beringer N (2001) Off-talk—a problem for human-machine-interaction? In: EUROSPEECH-2001, pp 2197–2200
19. Couture-Beil A, Vaughan R, Mori G (2010) Selecting and commanding individual robots in a vision-based multi-robot system. In: Seventh Canadian conference on computer and robot vision (CRV)
20. Castellano G, Pereira A, Leite I, Paiva A, McOwan PW (2009) Detecting user engagement with a robot companion using task and social interaction-based features. In: Proceedings of the 2009 international conference on multimodal interfaces (ICMI-MLMI'09), pp 119–126
21. Ishii R, Shinohara Y, Nakano T, Nishida T (2011) Combining multiple types of eye-gaze information to predict user's conversational engagement. In: 2nd workshop on eye gaze on intelligent human machine interaction
22. Nakano YI, Ishii R (2010) Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: 2010 international conference on intelligent user interfaces (IUI2010)
23. Goffman E (1963) Behavior in public places: notes on the social organization of gatherings. The Free Press, New York
24. Argyle M, Cook M (1976) Gaze and mutual gaze. Cambridge University Press, Cambridge
25. Duncan S (1972) Some signals and rules for taking speaking turns in conversations. J Pers Soc Psychol 23(2):283–292
26. Goodwin C (1986) Gestures as a resource for the organization of mutual attention. Semiotica 62(1/2):29–49
27. Kendon A (1967) Some functions of gaze direction in social interaction. Acta Psychol 26:22–63
28. Klotz D, Wienke J, Peltason J, Wrede B, Wrede S, Khalidov V, Odobez JM (2011) Engagement-based multi-party dialog with a humanoid robot. In: Proceedings of SIGDIAL 2011: the 12th annual meeting of the special interest group on discourse and dialog, pp 341–343
29. Mutlu B, Shiwa T, Kanda T, Ishiguro H, Hagita N (2009) Footing in human-robot conversations: how robots might shape participants roles using gaze cues. In: Proc of ACM conf human robot interaction
30. Rich C, Ponsler B, Holroyd A, Sidner CL (2010) Recognizing engagement in human-robot interaction. In: Proc of ACM conf human robot interaction
31. Shi C, Shimada M, Kanda T, Ishiguro H, Hagita N (2011) Spatial formation model for initiating conversation. In: Proceedings of robotics: science and systems
32. Michalowski MP, Sabanovic S, Simmons R (2006) A spatial model of engagement for a social robot. In: IEEE international workshop on advanced motion control, pp 762–767
33. Mower E, Mataric MJ, Narayanan S (2011) A framework for automatic human emotion classification using emotional profiles. IEEE Trans Audio Speech Lang Process 19(5):1057–1070
34. Zong C, Chetouani M (2009) Hilbert-Huang transform based physiological signals analysis for emotion recognition. In: IEEE symposium on signal processing and information technology (ISSPIT'09)
35. Peters C, Castellano G, de Freitas S (2009) An exploration of user engagement in HCI. In: Proceedings of AFFINE'09
36. Payr S, Wallis P, Cunningham S, Hawley M (2009) Research on social engagement with a rabbitic user interface. In: Tscheligi M, de Ruyter B, Soldatos J, Meschtscherjakov A, Buiza C, Streitz N, Mirlacher T (eds) Roots for the future of ambient intelligence. Adjunct proceedings, 3rd European conference on ambient intelligence (AmI09). ICT&S Center, Salzburg
37. Klamer T, Ben Allouch S (2010) Acceptance and use of a social robot by elderly users in a domestic environment. In: ICST PERVASIVE health 2010
38. Heerink M, Krose BJA, Wielinga BJ, Evers V (2006) The influence of a robot's social abilities on acceptance by elderly users. In: Proceedings RO-MAN, Hertfordshire, September 2006, pp 521–526
39. Mataric MJ (2005) The role of embodiment in assistive interactive robotics for the elderly. In: AAAI fall symposium on caring machines: AI for the elderly, Arlington, VA
40. Tapus A, Tapus C, Mataric MJ (2009) The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In: Proceedings, international conference on rehabilitation robotics (ICORR-09), Kyoto, Japan
41. Xiao B, Lunsford R, Coulston R, Wesson M, Oviatt S (2003) Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In: Proceedings of the 5th international conference on multimodal interfaces
42. Lunsford R (2004) Private speech during multimodal human-computer interaction. In: International conference on multimodal interfaces (ICMI'04), p 346
43. Batliner A, Hacker C, Kaiser M, Mogele H, Noth E (2007) Taking into account the user's focus of attention with the help of audio-visual information: towards less artificial human-machine communication. In: Auditory-visual speech processing (AVSP 2007)
44. Lunsford R, Oviatt S, Coulston R (2005) Audio-visual cues distinguishing self- from system-directed speech in younger and older adults. In: Proceedings of the 7th international conference on multimodal interfaces (ICMI'05), pp 167–174
45. Diaz R, Berk LE (eds) (1992) Private speech: from social interaction to self regulation. Erlbaum, Hillsdale
46. ten Bosch L, Boves L (2004) Survey of spontaneous speech phenomena in a multimodal dialog system and some implications for

ASR. In: Proceedings, international conference on spoken language processing, October 2004, South Korea
47. Petersen RC, Doody R, Kurtz A, Mohs RC, Morris JC, Rabins PV, Ritchie K, Rossor M, Thal L, Winblad B (2001) Current concepts in mild cognitive impairment. Arch Neurol 58:1985–1992
48. Wu YH, Fassert C, Rigaud AS (2012) Designing robots for the elderly: appearance issue and beyond. Arch Gerontol Geriatr 54(1):121–126
49. Yngve VH (1970) On getting a word in edgewise. In: Proceedings of the sixth regional meeting of the Chicago linguistic society
50. Shibata T, Wada K, Saito T, Tanie K (2001) Mental commit robot and its application to therapy of children. In: IEEE/ASME international conference on AIM'01
51. Saint-Aime S, Le Pevedic B, Duhaut D (2008) EmotiRob: an emotional interaction model. In: IEEE RO-MAN 2008, 17th international symposium on robot and human interactive communication
52. Lee J, Nam T-J (2006) Augmenting emotional interaction through physical movement. In: UIST2006, the 19th annual ACM symposium on user interface software and technology
53. Steinberger J, Jarek K (2004) Using latent semantic analysis in text summary evaluation. In: Proceedings of ISIM'04 2004, pp 93–100
54. Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2007) The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of interspeech, pp 2253–2256
55. Mahdhoui A, Chetouani M (2011) Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. Speech Commun 9–10:1149–1161
56. Breazeal C, Aryananda L (2002) Recognizing affective intent in robot directed speech. Auton Robots 12(1):83–104
57. Hacker C, Batliner A, Noth E (2006) Are you looking at me, are you talking with me: multimodal classification of the focus of attention. In: Sojka P, Kopêcek I, Pala K (eds) TSD 2006. LNAI, vol 4188. Springer, Berlin, pp 581–588
58. Truong K, van Leeuwen D (2007) Automatic discrimination between laughter and speech. Speech Commun 49:144–158
59. Boersma P, Weenink D (2005) Praat, doing phonetics by computer. Tech. rep, Institute of Phonetic Sciences, University of Amsterdam, Pays-Bas. URL www.praat.org
60. Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions. Speech Commun 49(3):201–212
61. Tilsen S, Johnson K (2008) Low-frequency Fourier analysis of speech rhythm. J Acoust Soc Am 124(2):EL34–EL39
62. Ringeval F, Chetouani M, Schuller B (2012) Novel metrics of speech rhythm for the assessment of emotion. In: Interspeech 2012
63. Zellner-Keller B, Keller E (1998) The chaotic nature of speech rhythm: hints for fluency in the language acquisition process. Integrating speech technology in language learning. Swets Zeitlinger, Lisse
64. Duda R, Hart P, Stork D (2000) Pattern classification, 2nd edn. Wiley, New York
65. Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin
66. Sanghvi J, Castellano G, Leite I, Pereira A, McOwan PW, Paiva A (2011) Automatic analysis of postures and body motion to detect engagement with a game companion. In: Proc of ACM conf human robot interaction
67. Olsen DR, Goodrich M (2003) Metrics for evaluating human-robot interaction. In: PERMIS 2003
68. Delaherche E, Chetouani M (2010) Multimodal coordination: exploring relevant features and measures. In: Second international workshop on social signal processing. ACM Multimedia, New York
69. Dahlbaeck N, Joensson A, Ahrenberg L (1993) Wizard of oz studies ? Why and how. In: Proceedings of the 1993 international workshop on intelligent user interfaces (IUI193). ACM Press, New York, pp 193–200
70. Xiao B, Lunsford R, Coulston R, Wesson M, Oviatt S (2003) Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In: Proceedings of the 5th international conference on multimodal interfaces. Vancouver, British Columbia, Canada
71. Light J (1997) Communication is the essence of human kife: reflections on communicative competence. Augment Altern Commun 13(2):61–70

Jade Le Maitre received an M.S. degree in Engineering jointly from the EPF Paris (Grande Ecole d'Ingenieurs) and the Hochschule Munich in 2010. She then started a Ph.D. in the Artificial Perception research group at the Institute for Intelligent Systems and Robotics. The topic of her thesis was the analysis of Social Signal in Human–Robot Interaction. Her research interests include social signal processing, HRI, and machine learning. She currently works at Provaltis, a French agency specialized in scientific and technic communication, providing support for the animation of research networks.

Mohamed Chetouani is the head of the IMI2S (Interaction, Multimodal Integration and Social Signal Processing) research group. He received the M.S. degree in Robotics and Intelligent Systems from the UPMC, Paris, 2001. He received the Ph.D. degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling (UK). Dr. Chetouani was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona (Spain). He is currently an Associate Professor in Signal Processing and Pattern Recognition at the UPMC. His research activities, carried out at the Institute for Intelligent Systems and Robotics, cover the areas of non-linear signal processing, feature extraction, pattern classification and fusion for social signal processing. He was member of the Management Committee for the COST action 2102: "Cross-Modal Analysis of Verbal and Non-verbal Communication". He is an Associate Editor of the Cognitive Computation Journal (Springer). He is also the co-chairman of the French Working Group on Human-Robots/Systems Interaction (GDR Robotique CNRS). In 2008, he leaded the project titled: Multi-Modal Communication with Virtual Agents and Robots for the 4th international summer workshop on Multi-Modal Interfaces (eNTERFACE'08).