

# Rapport with Virtual Agents: What do Human Social Cues and Personality Explain?

Aleksandra Cerekovic, Oya Aran, *Member, IEEE*, and Daniel Gatica-Perez, *Member, IEEE*

**Abstract**—Rapport has been recognized as an important aspect of relationship building. While rapport in the context of human-human interaction has been widely studied, how it can be established and maintained in human-agent interaction has been studied only recently. Our study investigates how social cues and personality of a human interacting with an agent can be used for automatic prediction of rapport in this context. We conduct experiments with two emotional virtual agents. Alongside the audio-visual data, we also collect human personality measures and two measures of rapport: self-reported rapport and rapport judged by observers. The social cues, such as turn-taking patterns and facial expressions are extracted from audio-visual data. Our results show that the most significant cues that infer the rapport judgments are the number of turn-taking cues and pauses. We also find that some of the significant social cues related to rapport are similar to those reported in previous psychology literature. We also confirm previous findings on how human personality plays an important role in perceiving the interaction with agents - people who score high in extraversion and agreeableness report higher rapport with both agents. Finally, the rapport prediction results suggest that automatic analysis of social phenomena in human-agent interaction could be a feasible method for agent evaluation.

**Index Terms**—Human-agent interaction, rapport prediction, human personality, nonverbal behavior analysis, social signal processing

## 1 INTRODUCTION

A number of sophisticated virtual agents that accomplish different tasks and exhibit a variety of interpersonal behavior are appearing. The Mach agent coaches for communication skills in a simulated job interview [1], the Rapport agent tries to establish and maintain the rapport [2], while some other agents aid people with difficulties, or make education more enjoyable. Despite efforts in the field, the reasoning functionality of state-of-the-art agents is limited to the specific purpose they have been designed for. As a result, the usual ways to evaluate human-agent interaction include measuring the agent’s usage and acceptance, and measuring the user’s experience and reactions during the interaction. For this purpose, questionnaires and interviews are indispensable assessment tools.

Our study investigates how a variety of visual, acoustic, and social cues displayed by a human user can be used to automatically measure rapport in human-agent interaction. Rapport has been recognized as one of the major aspects for building human-agent relationships [3] and is the focus of several studies (overview in Section 2). Motivated by recent technological advances in audio-visual processing [4], we explore how audio-visual data and features obtained from low-cost and user-friendly equipment and software (e.g. depth cameras) can be used to predict rapport. We conduct experiments and investigate behaviors of humans who interact with Sensitive Artificial Listeners (SALs) [5]. Automatic SALs are publicly available virtual agents designed to induce a specific emotional conversation with a

human by exhibiting four different moods: happy, angry, sad and neutral. Designed for active listening, they provide an emotional response to a human user through visual and acoustic back-channels. A storytelling scenario is the best fit for achieving realistic interaction with SALs; i.e. a happy agent may approve a user by nodding, smiling and saying: “You must have a good time.”, or may encourage a user to talk: “Do you have any gossip to tell?”. SALs reasoning is based on nonverbal user behavior analysis and management rules that select appropriate emotional responses from a lexicon. SALs have been carefully designed [5], and despite limited verbal skills and understanding, they can sustain a realistic interaction, as shown in evaluation studies [6].

In our experiment, we use two SALs (sad Obadiah and cheerful Poppy) for two conversation scenarios. For each scenario we study differences in the user’s social displays and whether human personality plays a role in the way a person expresses behavior and attitudes. For each scenario we collect two measures of rapport: a self-reported feeling of rapport, and the rapport judged by external observers. In the absence of self-reported measures, the external annotations could be collected as relevant measures, and in this study, we investigate how rapport is judged for the human-agent setting. Automatic rapport prediction is done with regression and classification models trained with self-reported personality traits and extracted social cues. A preliminary version of this work has been presented in [7]. In the current study, we extend our previous study by looking at judged rapport in addition to self-reported one. We also explore and propose additional features, such as facial expressions, linguistic content, and motion cues from depth data as rapport descriptors. We investigate the correlation of these features with rapport as well as their prediction power on both self-reported and judged rapport.

Our study has three contributions. First, we investigate

- A. Cerekovic is affiliated with the University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia (e-mail: acerekovic@fer.hr); O. Aran is affiliated with the Idiap Research Institute, Martigny, Switzerland (e-mail: aran@idiap.ch); D. Gatica-Perez is affiliated jointly with the Idiap Research Institute, Martigny, Switzerland, and Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland (e-mail: gatica@idiap.ch)

the links between extracted social cues and self-reported human personality to two measures of rapport. We also explore the use of an inexpensive motion tracking system and depth images for automatic interaction analysis. Hereby, we complement the findings in agent studies that have investigated the relation of rapport to human personality [2] [8]. Second, we investigate external judgments of rapport; first, we investigate how a crowdsourcing platform can be utilized to collect judgments of rapport in human-agent interaction, and then we study what social cues infer those judgments. Finally, we report results on prediction of rapport measures based on extracted social cues and self-reported personality traits. We extract a large set of audio-visual and social cues as features for building rapport models. To our knowledge, our study represents a first attempt to predict rapport in human-agent interaction from audio-visual data. There are a few works similar to ours in the literature, measuring postural congruence [9], and gaze and voicing of interactants [10] to predict rapport in human-human interaction. In comparison to these works, our study targets human-agent interaction and uses a larger number of social cues for prediction.

This paper is organized as follows. In Section 2, we discuss the related work on rapport in psychology and computational science. In Section 3 we describe the data collection and in Section 4 computational methods to automatically extract social cues. Results, in which we discuss the quality of external judgments, and present a statistical analysis of the rapport measures and their relationships with social cues and human personality, as well as the prediction results, are given in Section 5. Finally, we conclude in Section 6.

## 2 RELATED WORK

### 2.1 Related work in psychology

Rapport in dyadic interaction manifests itself when two people “click” [11], perceive interaction as enjoyable [12] [13], and feel connected and close to each other [14]. As rapport leads to success in interpersonal interactions, many works in psychology, mainly coming from marketing and education, investigate this social construct.

The most well-known theoretical model of rapport is proposed by Tickle-Degnen and Rosenthal [11]. Based upon meta-analysis of the literature, their framework decomposes rapport into three components: positivity, attentiveness, and coordination. Positivity refers to positive attitudes that interactants show to each other, including smiling, forward leaning and eye-contact. Attentiveness means that subjects feel involved in the interaction, displaying behaviors such as forward lean, uncrossed arms, eye contact, listening, openness, empathy and friendliness. Coordination, or being ‘in-sync’ is manifested through smooth turn-taking and unconscious mimicry and it often happens spontaneously. In practice however, this framework is not applicable to every conversational context [12]. In a study on rapport in “strangers meet” scenario, Bernieri et al. [15] failed to decompose rapport into the three components upon factor analysis, most likely because these are highly correlated. Capella further raises a question about the need for “coordination” as a necessary component [16] - in situations of

conflict between husbands and wives, partners are coordinated, but lack of rapport. He also argues that the presence of positivity is the most evident component of building rapport. Some other studies have also shown how nonverbal displays, that Tickle-Degnen and Rosenthal associate to their components, can not be taken for granted. In a debate study on controversial topics, subjects who argued a lot leaned forward often, which is contrary to the belief that forward leans indicate a high rapport [13]. In another situation in which people receive bad news from a doctor, informing the illness state of their relative, there could be a high presence of negative displays and a high presence of positivity [17].

To measure rapport in face-to-face interaction, researchers have proposed different scales, mainly constructed upon the Tickle-Degnen and Rosenthal’s components, such as [18], also with examples in virtual agent studies [19], [2]. Measured rapport is likely to differ between the interactants and external observers. In a debate study between mixed sex dyads, the correlation between the male and female rapport was only .34 [13]. Observer judgments, on the other hand, are likely to correlate highly; in a number of studies Bernieri and his colleagues have confirmed that observers can judge the rapport reliably by watching short interaction clips, lasting 30-50s, even without hearing the conversation. This finding led them to conclusion that rapport judgments seem to be more of an automatic perception process than a cognitive one ([20] pp. 83). A summary of rapport studies and theoretical models can be found in ([12] pp. 83; [20] pp. 67). Instead, we conclude this brief overview with the case of rapport in the “strangers meet” scenario, which matches our study.

In the “strangers meet” scenario, mimicry and postural congruence have been found to be significant in the construction of liking ([21], [22]). Psychological literature also shows how, depending on the social context, there might be a high discrepancy between displays that interactants use to communicate liking towards each other, and displays that infer observer judgments. Researchers agree that frequent eye contact, relaxation, leaning and orienting towards, less fiddling, moving closer, touching, more open arm and leg positions, smiling and more expressive face and voice are behaviors that observers interpret as signs of liking [23], [24]. Acoustic cues are as important as visual cues in judging other people [25]. On the other hand, in communication of positive attitude via posture cues, Mehrabian shows how interactant’s displays vary with respect to his/her gender and status [26]. A study on initial same-sex dyad interactions [22] shows how amount of mutual gaze and the total percentage of looking time are indicators of high liking between subjects. Other significant behaviors are: expressiveness of the face, synchrony of movement and speech, and expressiveness of the gesturing. Another cross-study on displays of liking [27] examined vocalic behaviors, showing how an increased pitch variety as indicator of liking is only shown by females. Bernieri and colleagues have also investigated the nonverbal cues that communicate rapport and infer rapport judgments in two different contexts: debate and trip planning between two mixed-sex strangers ([20] pp. 77). It is not strange that different cues have been found in those two studies. In the latter one, interactants encoded rapport mainly via mimicry, which is then followed by proximity.

On the other hand, observers interpreted situations in which interactors were smiling a lot and were more expressive (face, hand gestures) as situations of high rapport.

## 2.2 Related work in computational science

Nonverbal social cues often occur unconsciously in a face-to-face context, but are very significant in describing attitudes of interactants. Recently, a new research field that addresses automatic face-to-face interaction analysis, called social signal processing, has emerged [4]. Social signal processing addresses 5 categories of social cues: physical appearance; gesture and posture; gaze, facial behaviors and mimics; vocal cues; and proxemics and environment. Among vocal cues, prosody and vocal quality have been analyzed the most [28], with more recent works on emotion recognition from speech [29]. Visual cues include motion space [30], facial expressions [31], hand and body movements [32], and gaze and visual focus of attention [33]. Consumer depth cameras and available pose estimation algorithms are used as input for activity recognition [32] and greeting pattern analysis [34]. Yet when it comes to rapport and mimicry modeling from audio-visual data, we found several notable works. A database on mimicry in human-human interaction is proposed in [35]. This dataset provides an interesting benchmark for studies on mimicry, which is a component of rapport. A few other approaches address the problem of detecting mimicry in task-oriented interaction by means of motion energy analysis [36] and by means of body tracking information [37]. In [9], self-reported feelings of rapport are modeled by measuring postural congruence. In [10], rapport is classified by fusing audio-visual features, namely gaze and voicing features, with 66% - 81% accuracy for different prediction models.

Automatic human behavior analysis from audio-visual data is crucial for developing the reasoning functionality in virtual agents. Yet, we found no studies on automatic human-agent rapport analysis. The vast majority of related rapport studies focus on developing human abilities in an agent, i.e. the Rapport agent [2], [19]. In [8] the Rapport agent is used to measure how human personality affects human feelings towards the agent. The results show how Big Five extraversion and agreeableness are more relevant than features like gender or age. Another study compared human self-reported feeling of rapport in interaction with the Rapport agent to the rapport in interaction with another human [2]. Results indicate that people who score higher in conscientiousness perceive strong rapport with a human, whereas people high in agreeableness report a preference for the agent. More recently, a theoretic model of rapport development in human-agent setting has been proposed [14]. The model is based upon a meta-review of the existing literature and its functionality is successfully tested on a corpus of human-human interaction.

## 3 HINT RAPPORT STUDY

The methodology of our study, depicted in Figure 1, has been inspired by studies on prediction of dominance in human-human interaction [33], and prediction of human personality from social media [38]. We study prediction

of two measures of rapport: self-reported rapport from the subjects who interact with agents, and judged rapport, assessed by the external observers. Self-reported rapport is collected in an experiment in which subjects interact with two virtual agents. Subjects are recorded during interaction, and also asked to report their personality traits. Experiment design and data collection are detailed in Section 3.1.

Judged rapport is collected in a crowdsourcing setting (Section 3.3), in which short 1-minute interaction videos are shown to non-trained workers (Figure 2). To investigate features for rapport prediction models, we extract social cues from audio-visual data. The majority of audio-visual features are extracted automatically, as described in Section 4. The results in Section 5 are divided in three different parts. First, we analyze the quality of rapport judgments from the crowdsourcing experiment. Second, we deliver the descriptive statistics of two measures of rapport and investigate relation between them. Next, we investigate significant social cues for the rapport measures, and also whether human personality is correlated to the rapport. Finally, we address the problem of predicting the rapport measures from the social cues and human personality traits.

### 3.1 Experimental design and data collection

The HINT (Human INTeraction) dataset obtains audio-visual recordings of 33 subjects (14 females and 19 males), with mean age 26.7, ranging from 18 to 43 years, with different cultural backgrounds (85% subjects are Caucasians). Each subject is interacting with two SALs: sad Obadiah, and cheerful Poppy. These SALs are selected based on the findings of [6], showing how Poppy is the most consistent and familiar character, while Obadiah is the most believable character.

Before the recording session, each subject had to sign the consent form and complete the NEO FFI Big Five personality questionnaire. Then, each subject was briefly introduced to the nature of the experiment. Specifically, subjects were told to have a conversation with two different virtual agents. Subjects were also told that agents behave like a human listener and they are supposed to tell them a story, and not to ask questions. To encourage the interaction, a list of five suggested conversation topics were placed in front of the subject (i.e. weekend plans, things that a subject did yesterday). However, subjects were allowed to speak of any topic of their choice. Furthermore, SALs could also ask questions not necessarily related to the topic of the conversation. Upon initiating the agent, the experimenter would exit the recording room, and return after 4-minutes to stop the interaction. After interaction, each subject filled out the post-interaction questionnaire. This questionnaire is used to measure a self-reported feeling of rapport and is inspired by [39]. In our setup, every subject interacted first with sad Obadiah, and then with cheerful Poppy without prior knowing the agent personality. Due to the relatively small number of recruited subjects, same experimental conditions were used for every subject. We further discuss this limitation in Section 5.

Several recordings were obtained at the end of each recording session: RGB-D data and audio data from Kinect,

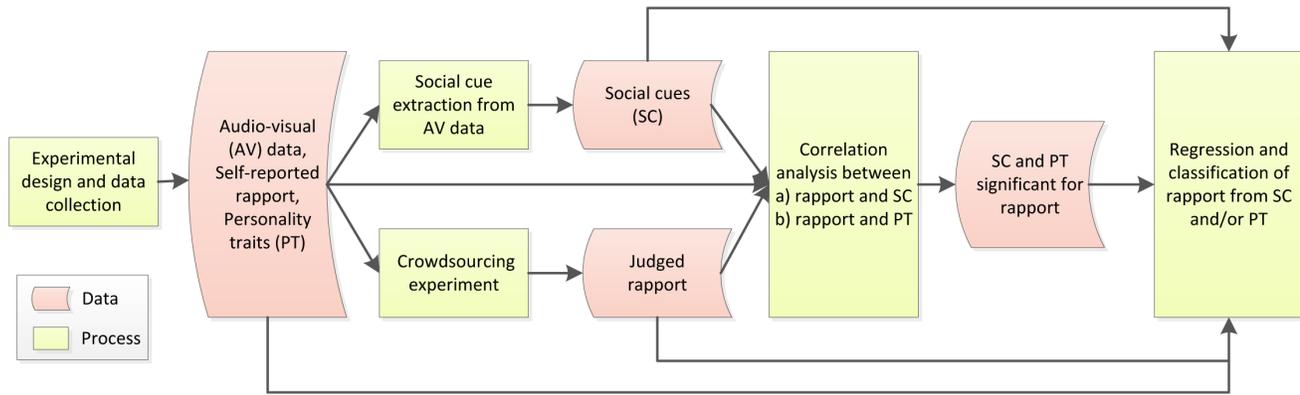


Fig. 1. The methodology used in our study. Prediction of two measures of rapport (self-reported and judged) is investigated using audio-visual features, or social cues (SC), and human personality traits (PT). See Section 3 for details.

and screen captures and log files with a description of the agents behaviour. In total, we performed 66 recordings. One session with missing RGB data for a subject interacting with Poppy, and missing screen captures for two agents in other two sessions were discarded in a crowdsourcing experiment, which comprises 32 interaction sessions with Obadiah and 31 with Poppy. Further details about data collection can be found in [7].

### 3.2 Self-reported rapport

To measure a self-reported feeling of rapport we used the post-interaction questionnaire. In our previous work [7], we used this questionnaire to measure three different aspects of human-agent interaction: quality of interaction, degree of rapport and degree of liking the agent. However, since Tickle Dengen’s and Rosenthal framework suggests that the quality of interaction and liking the agent can be regarded as rapport components, we perform Exploratory Factor Analysis (EFA) with oblique principal-axis transformation on the post-interaction questionnaire.

Rather high Kaiser-Meyer-Okin (KMO) index of 0.89 on the item scores ( $n = 66$ ; 15 items) has shown that our data is suitable for the EFA. In further analysis, Cattell’s scree plot and Revelle and Rocklin Very Simple Structure (R psych package), suggested only one underlying component. This is also confirmed by the EFA, where the single factor EFA explains 58% of the variance, yielding a root mean square of the residuals (RMSR) of 0.07. Using 2 factor EFA, the first factor explains 55% and the second only 5% of the variance, with a RMSR of 0.06. In our case, the EFA could not confirm three rapport components of Tickle Dengen’s and Rosenthal framework. This phenomenon is also found in “strangers meet” human-human interaction study done by Bernieri et al. [15], explaining how components are likely correlated in this context. In the final revision of the post-interaction questionnaire we removed the item “I felt self-conscious during the conversation.” since it had insignificant factor loading (-0.04). The revised questionnaire is hereafter referred as Human-Agent Rapport Questionnaire (HARQ). The EFA factor loadings suggest that items measure one social construct (Table 1, Cronbach’s alpha 0.94). The validity of items for measuring rapport is investigated in the crowdsourcing experiment (Section 3.3).

TABLE 1  
Factor Loadings from EFA with oblique transformation for 1-factor solution for the Human-Agent Rapport Questionnaire (HARQ),  $n = 66$

Item	Factor loadings
I got along with the character pretty good.	0.88
I did not want to get along with the character.	0.74
I was paying attention to way that character responds to me and I was adapting my own behaviour to it.	0.49
I felt that character was paying attention to my mood.	0.73
The interaction with the character was smooth, natural, and relaxed.	0.81
I felt accepted and respected by the character.	0.86
I think the character is likeable.	0.85
I enjoyed the interaction.	0.84
The interaction with the character was forced, awkward, and strained.	0.82
I felt uncomfortable during the interaction.	0.44
The character often said things completely out of place.	0.73
I think that the character finds me likeable.	0.81
The interaction with the character was pleasant and interesting.	0.68
I would like to interact more with the character in the future.	0.59

TABLE 2  
Items in the modified Laken’s Rapport Scale (mLRS) [18]

The person and the agent enjoyed the interaction.
The person and the agent did not like each other.
The person and the agent were aware of and were interested into each other.
The person and the agent did not understand each other.
The person and the agent were coordinated and in rhythm.
The person and the agent were not involved into the interaction.

### 3.3 Rapport judgments: a crowdsourcing experiment

Studies show that crowdsourcing is an inexpensive and fast way to collect quality annotation data for experimental research [40], such as human personality annotations [38]. For that matter, we used Crowdfunder, an aggregating crowdsourcing platform with more than 5 million of workers.

To design a Crowdfunder task we created 1-minute video clips, extracted from 2nd to 3rd minute of human-agent interaction sequence. Thin slices, or short video clips, are commonly used for making judgments in first impression studies. One minute clip length is selected based on psy-

chological finding that observers can assess rapport by watching clips lasting 30- 50 s [13]. Decision about extraction location (2nd to 3rd minute) is less straightforward. While some studies on rapport and personality judgments use clips extracted from the beginning of interaction [13][6], some other studies show how the beginning of interaction contains “degree of awkwardness” and uncoordinated non-verbal behavior [11], that is potentially misleading for judgments on different traits, such as personality (see [6], p.315). Also, because of the interaction order in our experiment (first Obadiah, then Poppy), which implies unfamiliarity of subjects with the experiment when encountering Obadiah, we assumed that the most spontaneous and meaningful part of interaction occurred later in the interaction, so we chose to extract the clip from 2nd to 3rd minute. The final task we designed for Crowdfunder is shown in Figure 2. Synchronized human and agent playbacks, with the agent displayed in the upper left corner, were shown to the workers. For annotation of rapport we included two rapport scales: our Human-Agent Rapport Questionnaire (HARQ), and modified version of Laken’s rapport scale (mLRS) [18]. HARQ items displayed in Table 1 are modified for the purpose of judging impressions from the third perspective. The mLRS scale, which is displayed in Table 2, is added to measure the construct validity of the HARQ scale. The mLRS contains modified items used in the judgments of rapport [18] (Cronbach’s alpha 0.81). The original questions are reformulated into the statements, and three were reversely coded. Finally, one item is replaced with the item used by Bernieri et al. [13]. All items are scored on the five-point Likert scale. For control reasons, we also added a question to collect the workers opinion about the topic of interaction.

We created two different jobs: first job contained videos of interaction with Obadiah, and second with Poppy. Due to missing data during the recording, the job with agent Obadiah contained 32 videos, and the job with agent Poppy 31 videos. In order to avoid influence of judgments between two experimental conditions, and to provide a different pool of workers working on the task, each job is launched at different time. The job with Obadiah is launched at 10 am. CET, and the job with Poppy at 11 pm. CET, four days after. Each video within a job is annotated by 5 workers coming from the following English-speaking countries: USA, Great Britain, Canada and Australia. The cost of one annotation, with completion time estimated on 2 minutes, is set to 30 cents.

## 4 SOCIAL CUE EXTRACTION

Based upon a review of the literature on displays of rapport and personality, the following verbal and nonverbal cues are selected and extracted from audio-visual data: language style, speaking activity, pauses, prosody, body leans, head direction, visual activity, hand activity and facial expressions (see Table 3). Spoken transcripts, which are essential for computing the language style, and hand activity were manually annotated. In this study we focus on individual human behavior. We believe that features obtained by mimicry analysis can improve rapport prediction, but not dramatically in our case. SALs have a “talking head” appearance, so postural congruency and hand gestures, which



Please first watch the 1-minute video completely! Once the video is finished, you can start filling in the questionnaire.

The following questionnaire contains a series of statements related to the interaction you have just watched. Read each statement carefully and tick the box that best reflects your opinion.

1. The person and the agent got along pretty good. (get along with – establish a relationship).

Strongly Disagree	1	2	3	4	5	Strongly Agree
-------------------	---	---	---	---	---	----------------

2. The person did not want to get along with the agent.

Strongly Disagree	1	2	3	4	5	Strongly agree
-------------------	---	---	---	---	---	----------------

Fig. 2. A snapshot of the task designed to collect rapport judgments from the Crowdfunder workers.

are the most powerful way of expressing the mimicry ([21], [13]), can not be considered. Moreover, SALs behaviors are short and similar at every turn. In speaker state, SALs say a simple, emotional comment or question, accompanied with eyebrow and head movements. In listener state, they use blinks, head and eyebrow movements. Obadiah shows sad facial expressions, and Poppy happy ones. This implies that by measuring facial expressions of a human, we can grasp a certain level of mimicry between agent and a human.

### 4.1 Verbal cues

**Language style.** In order to analyze the language style of subjects, we use Linguistic Inquiry and Word Count software (LIWC) [41]. LIWC is a text analysis tool, which computes occurrence of 4 word categories and its 81 subcategories. These are designed to study structural, emotional, and cognitive components of a language. These include: Linguistic Processes (with subcategories such as word count (WC), words longer than six letters, etc.), Psychological Processes (social such as friends or family, affective such as sad, anger), Personal Concerns (such as achievements, or health), and Spoken Categories (assent, fillers). LIWC has been utilized in social studies related to natural language processing and spoken dialogue analysis. For example, in task-oriented scenario leadership skills of an individual can be expressed in a spoken dialogue [42].

We use LIWC to process spoken transcripts per subject. Transcription is done manually because experiments with three state-of-the-art speech recognizers have resulted in relatively poor results, mainly due to majority of non-native English speakers in the HINT dataset. From LIWC analysis we excluded punctuations (periods, dots, exclamation marks, etc.), which resulted with 66 different features.

### 4.2 Nonverbal cues

#### 4.2.1 Auditory cues

To extract nonverbal cues from speech, we first have applied automatic speaker diarization on human-agent audio files

using Idiap Speaker Diarization Toolkit [43]. Based on diarization output, we compute speaking activity, pauses and voice quality measures. The resulting auditory nonverbal cues contain 27 features.

**Speaking activity.** We extract speech segments and compute the following features for each interaction sequence: total speaking length (TSL), total speaking turns (TST), filtered turns (TSTf - turns shorter than 2 seconds were not taken into account), and average turn duration (ATD).

**Pauses.** Based on the speech segments we compute 4 features that measure occurrence of pauses in each sequence. The following features are inspired by [31]: presence of pauses, or proportion of active time (PausesPT), frequency, or number of pause segments (PausesNS), average duration of pauses (PausesAD) and presence of pauses shorter than 0.5 sec (PausesPTS). Computation details can be found in [31], p. 2.

**Voice quality measures.** To export voice quality measures we use MIT Human Dynamics group toolkit ([28]). The voice quality measures are computed on the subject's speech, which was extracted from the recorded audio using the diarization output. We extract the statistics - mean and standard deviation - of the following prosodic features: pitch (F0 (m), F0 (std)), pitch confidence (F0 conf (m), F0 conf (std)), spectral entropy (SE (m), SE (std)), delta energy (DE (m), DE (std)), location of autocorrelation peaks (Loc R0 (m), Loc R0 (std)), number of autocorrelation peaks (# R0 (m), # R0 (std)), and value of autocorrelation peaks (Val R0 (m), Val R0 (std)). Furthermore, five other measures are exported: average length of speaking segment (ALSS), average length of voiced segment (ALVS), fraction of time speaking (FTS), voicing rate (VR), and fraction speaking over (FSO).

#### 4.2.2 Visual cues

One of the aspects we wanted to investigate is the use of an inexpensive motion tracking system (MS Kinect SDK v1.8) and depth images for automatic interaction analysis. Performance of the Kinect SDK upper body tracker was subjectively judged by visually inspecting the aligned skeleton on the subjects body in real-time. The results were quite good for head, neck and shoulder joints, whereas elbow and hand tracking produced poor results. This is a reason why hand activity is manually annotated. We used Kinect SDK upper body and face tracking information to model body leans and head direction classifier.

**Body Leans.** We developed a method for automatic body lean analysis from 3D upper body tracking information. We trained a Support Vector Machine (SVM) classifier with RBF kernel, to recognize the following body leans per frame: neutral, sideways left, sideways right, forward and backward leans. These categories are inspired from psychological findings on affective displays in body posture [26].

The SVMs are trained on a balanced dataset, with 300 manually annotated frames, representing 8 human subjects who had the most expressive torso movements in the HINT dataset. We use the following features: 3D neck and shoulder joints, z-coordinate of head joint, and 3D shoulder joints normalized with respect to the subject's neutral body pose. The subject's neutral body pose is approximately an upright 90-degree sitting position. To compute it, we first annotated frames for each subject containing these poses based on

visual inspection. Then, for each interaction sequence, we find the subject's neutral body pose which is the first frame within the distance of one standard deviation from mean of annotated frames. With the features, we train SVMs with 10-fold cross-validation, to obtain the best parameters for SVMs. The best model is tested on the rest of the annotated images (forward leans: 3937 frames; neutral: 10136; sideways left: 42; sideways right: 87; backwards: 113) with the accuracy of 94% for neutral, 98% forward leans, 70% for sideways left, 62% for sideways right, and 72% for back leans. Using the body leans classifier, we compute distribution of body leans and frequency of shifts between the leans.

**Head Direction.** We implemented a simple heuristic method, which outputs two head directions per frame: screen, or away from screen. The method is using 3D object approximation of the screen and 3D head coordinates from the Kinect face tracker. We tested the performance on manually annotated ground truth data from 10 randomly selected subjects (away: 339 frames; screen: 1249). We obtained accuracy of 72% for away, and 81% for screen. Using the head direction method, we compute distribution of head directions and frequency of shifts from one position to another.

**Visual activity.** The visual activity of the subject is extracted by using two spatial motion cues: weighted motion energy images (wMEI) and depth wMEI images. The following statistics is calculated for each cue: entropy, mean and median value.

wMEI image is a binary image that describes the spatial motion distribution in the video sequence [30]. It is a simple method which highlights regions in the image where any form of motion is present, if a camera is fixed. As such, it can be utilized to observe a space of movements of a person sitting in front of the computer screen. We used the length of the interaction sequence to normalize the wMEIs.

Depth wMEI images are calculated based on depth images and a similar strategy is followed as in WMEI images based on grayscale intensity. One extra step that we needed to follow was to preprocess the depth images to remove the holes, or noise on the depth image that occurs during the Kinect recording. For preprocessing, the depth images are first converted to have depth values in the grayscale range and then we applied morphological operators and inpainting to remove the holes in the depth images. First, we used a closing operator applied with an ellipse shaped structuring element (with size 25 by 25). While the closing operator eliminates most of the holes, some of the holes still remain. To remove those as well, we applied OpenCV's Navier-Stokes based inpainting method on the remaining holes, where the hole is filled with the depth values of the neighboring pixels. Dependent on the neighbouring pixels, these methods generally perform better on smooth surfaces than on boundaries. We manually checked the resulting images after preprocessing and confirmed that the holes are successfully removed without significant distortions on the depth images. After this step all holes have been removed and the resulting depth images are ready for further processing. The rest of the processing is similar to the wMEI calculation [30]. The aim is to have a motion image via frame differencing, where the motion is defined with respect to the

depth change. For the thresholding, we used a threshold of 15, which is determined empirically.

**Hand activity.** For hand activity, every 15th frame in the interaction sequence is manually annotated with one of the following labels: hidden hands (HDH), hand gestures (GES), gestures on table (GOT), hands on table (HOT), and self-touch (ST). Based on these labels, we compute distribution and frequency of hand activity shifts.

**Facial Expressions.** Facial expressions are modeled with an output of the Computer Expression Recognition Toolbox (CERT) [44], which is a state-of-the-art facial expression recognition software. For selected emotion in a video sequence, CERT outputs estimated probabilities of the emotion occurrence per frame. Although CERT toolkit is designed for situations when a person is not speaking, we visually inspected that the HINT dataset contains very few facial expressions on listener segments, so we processed the whole conversational sequences instead.

We use CERT estimations of basic emotions: happiness, sadness, surprise, anger, disgust; plus neutral expression and smile intensity. The way how we model the facial expression cues is inspired by [31], who compute the presence of expressions, their duration, and their frequency from CERT probabilities. The main difference between their method and our method, is that, for some emotions, we adapt the threshold for binarization of probabilities to the human subject. We inspected that emotions expressed by some HINT subjects had constantly fairly high probabilities through sequence, which is presumably due to subjects' facial features (wrinkles, beard, the eyebrows shape). For example, this was inspected for emotions of anger and surprise for some subjects. In this case, a low fixed threshold for binarization of probabilities would yield constant presence of anger and surprise for these subjects, and if raised, anger and surprise would cease to exist for the rest of subjects. To adapt threshold, we assume that there is at least 10% of data in each sequence in which a subject is not showing expression of anger, sadness, disgust and surprise. For these expressions, threshold is initialized to 0.015, and then increased with the mean of subset of probabilities containing 10% of the lowest values. Neutral expression and expression of joy may be constantly present during interaction, so for these emotions we leave the fixed threshold of 0.015. For smiles, we use 0. Once the emotion probabilities are converted to binary values, morphological filters dilation and erosion are used for smoothing. From smoothed values we extract three different facial features: proportion of active time (PT), number of active segments (NS) and average duration of expressions (AD). Computation of these features is explained in [31], p. 2.

## 5 RESULTS AND DISCUSSION

### 5.1 Analysis of crowdsourcing data

To collect annotations of rapport we created two jobs with 63 tasks; the job with Obadiah contained 32 tasks, and the job with Poppy 31 tasks. Five workers were assigned to each task. Tasks were completed by a total of 117 workers from 4 countries. Tasks with Obadiah were completed by 70 workers in 2 hours, and tasks with Poppy were completed by 56 workers in 1 hour and 10 minutes. Among 117 workers, 6

TABLE 3  
Social cues extracted from audio-visual data. See details in Section 4

Lang. Style	LIWC categories (see [41]): Linguistic Processes, Psychological Processes and Personal Concerns # Features: 64
Sp. Activity	Total speaking length (TSL), Total speaking turns (TST), Filtered turns (TSTf), Average turn duration (ATD) # Features: 4
Pauses	Proportion of active time (PausesPT), Number of segments (PausesNS), Average duration (PausesAD), Presence of pauses <0.5 sec (PausesPTS) # Features: 4
Voice Quality Measures	Means(m) and standard deviations (std): Pitch (F0 (m), F0 (std)), Pitch confidence (F0 conf (m), F0 conf (std)), Spectral entropy (SE (m), SE (std)), Delta energy (DE (m), DE (std)), Location of autocorrelation peaks (Loc R0 (m), Loc R0 (std)), Number of autocorrelation peaks (#R0 (m), #R0 (std)), Value of autocorrelation peaks (Val R0 (m), Val R0 (std)) Average length of speaking segment (ALSS), Average length of voiced segment (ALVS), Fraction of time speaking (FTS), voicing rate (VR), Fraction speaking over (FSO) # Features: 19
Body Leans	Neutral leans (NLS), Sideways left leans (SLLS), Sideways right (SRLS), Forward leans (FLS), Backward leans (BLS), Frequency of lean shifts (FQLS) # Features: 6
Head Direction	Head directed towards screen (HDS), Head directed away from screen (HDA), Frequency of head direction shifts (FQHD) # Features: 3
Visual Activity	Median (med)), mean(m), and entropy (e): wMEI (wMEI (med), wMEI (mn), wMEI (e)), depth wMEI (dwMEI (med), dwMEI (mn), dwMEI (e)) # Features: 6
Hand Activity	Hidden hands (HDH), Hand gestures (GES), Gestures on table (GOT), Hands on table (HOT), Self-touch (ST), Frequency of hand shifts (FQHS) # Features: 6
Facial Expressions	Proportion of active time (PT), Number of active segments (NS), Average duration (AD) of the expressions; Neutral (NeutralPT, NeutralNS, NeutralAD), Anger (AngerPT, AngerNS, AngerAD), Sadness (SadnessPT, SadnessNS, SadnessAD), Joy (JoyPT, JoyNS, JoyAD), Disgust (DisgustPT, DisgustNS, DisgustAD), Surprise (SurprisePT, SurpriseNS, SurpriseAD), Smile (SmilePT, SmileNS, SmileAD) # Features: 21

workers annotated videos of both Poppy and Obadiah. Each worker completed on average 2.7 tasks. The average time of task completion was 2 mins and 36s for Obadiah, and 2 mins and 57s for Poppy. The price of the whole experiment is \$133.

After jobs completion we visually inspected annotations. Since the workers did not have any restrictions on the number of tasks, we paid attention to cases when a worker completed 6 or more tasks, and when a worker's performance in some of the previous Crowdfunder jobs was not good. We found 6 suspicious workers who annotated 38 videos in total, and who scored all items similarly. On the 5-point Likert scale, with some reverse coded items, scores of 3 workers were all 3s, scores of 2 workers were 2s and 3s, and

scores of one worker 4s and 5s. These workers completed their tasks in a very short period of time (approx. 70-75 secs), and it is very likely that they did not pay attention to the content. The usage of reverse coded items was a good way to spot the spammers. Contribution of suspicious workers was excluded, and another Crowdfunder job is used to collect the missing annotations.

On the final annotations, we first investigated the control question about conversation topic. Workers could choose among three offered topics, and the category “None of the above”, accompanied with an explanation box. For some interactions 2 or more topics were present, so it is expected that the workers agreement will differ highly in these cases. Fleiss’ Kappa for the control question was moderate for the job with Obadiah ( $\kappa=39$ ), and fair for the job with Poppy ( $\kappa=29$ ). Further, we studied consistency of the interrater reliability by computing the Intra-Correlation Coefficient (ICC) and interrater Cronbachs  $\alpha$ . We computed two Intraclass Correlation Coefficients (ICCs) for each item: the ICC(1,1) to measure the extent to which two workers agree with each other, and the ICC(1,k) to measure the agreement level in rating the targets, when the annotations are aggregated across the five workers to obtain a final rapport score. The ICC and interrater Cronbachs  $\alpha$  results are given in Table 4. The agreement is moderate for both rapport scales (HARQ and mLRS) on both jobs. Besides, we also measured the internal consistency of individual items in HARQ and mLRS scale. Cronbachs  $\alpha$  are given in Table 4 (2nd field, column 3). The lowest Cronbachs  $\alpha$  of 0.7 which is obtained for mLRS scale, agent Obadiah, is still considered acceptable measure of scale reliability.

To study whether the HARQ scale measures rapport, we computed the Spearman’s correlation coefficient between individual scores from the HARQ scale and scores from the control mLRS scale. Scatter plot displayed on left side of Figure 3 reveals strong correlation between the scores ( $n = 315$ ;  $\rho = 0.83$ ,  $p$ -value  $<0.001$ ). Spearman’s  $\rho$  between the HARQ and the mLRS scores, computed distinctly for Obadiah ( $n = 160$ ;  $\rho = 0.8$ ,  $p$ -value  $<0.001$ ), and for Poppy ( $n = 155$ ;  $\rho = 0.85$ ,  $p$ -value  $<0.001$ ) are also on the high side. Right side of Figure 3 displays two box plots of individual rapport scores on two scales, whereas Table 4 shows the descriptive statistics. All scores range from minimum 1 to maximum 5. For agent Poppy, there is less visible difference between HARQ and mLRS scores - difference between the average of scores is only 0.04 (1%). For agent Obadiah, there is a more obvious difference: the box plot shows how HARQ scores are slightly skewed to the right when compared to mLRS scores, which has normal distribution. However, the difference between the average of scores is also low - 0.12 (3%). Based on the results, we find the rapport validity of the HARQ scale to be acceptable. To compute judged rapport per interaction, we averaged judged rapport, computed on the HARQ scale, from the five workers.

## 5.2 Analysis of self-reported and judged rapport

Distribution of the self-reported feelings of rapport and distribution of judged rapport are given in Figure 4, and descriptive statistics in Table 5. We observe how self-reported rapport with Obadiah is lower than self-reported rapport

TABLE 4

Descriptive statistics of individual worker’s scores (Obadiah (O),  $n = 160$ ; Poppy (P),  $n = 155$ ) and worker’s agreement (\*ICC:  $p < 0.001$ ) on two rapport scales: our Human-Agent Rapport Questionnaire (HARQ) and the modified Laken’s Rapport Scale (mLRS)[18]

Scale	M	SD	MD	$\alpha$	ICC(1,1)	ICC(1,k)	Interrater $\alpha$
O	HARQ	2.84	0.75	3.00	0.89	0.23*	0.60*
	mLRS	2.92	0.72	3.00	0.70	0.27*	0.65*
P	HARQ	2.96	0.79	3.00	0.92	0.31*	0.69*
	mLRS	3.00	0.81	3.00	0.82	0.21*	0.56*

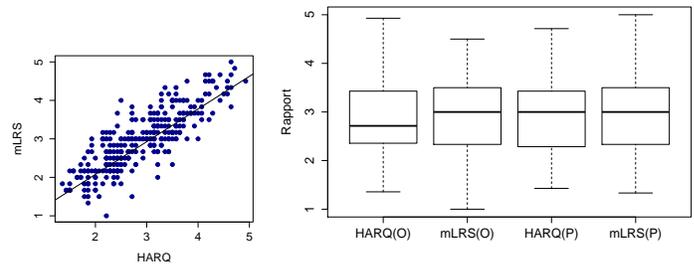


Fig. 3. Comparison of scores on two rapport scales, HARQ and mLRS: correlation between all individual scores (left), and distribution of scores for Obadiah (HARQ(O) and mLRS(O)) and Poppy (HARQ(P) and mLRS(P)) (right)

with Poppy, and is right-skewed (Figure 4). Indeed, the paired t-test ( $t(32) = -6.31$ ,  $p < 0.05$ ) has confirmed that the HINT subjects on average report higher self-report rapport with Poppy ( $M = 3.36$ ,  $SD = 0.83$ ), than with Obadiah ( $M = 2.35$ ,  $SD = 0.69$ ). During the experiment, several subjects reported that “Poppy seems to be dull”. Besides, findings on SALs show how Obadiah is the most believable, and Poppy ‘possibly the archetypal character’ [6]. Based on these a priori findings, we did not expect any significant preference for Poppy shown by the HINT subjects. On the other hand, if we observe interaction with Obadiah under the lens of psychological findings (Section 2.1), it is expected that emotional support via subject’s self-disclosure about some sad topic yields high rapport with Obadiah. However, this situation is very unlikely to happen in 4 minutes of recorded interaction. A few of the subjects who reported preference for Obadiah also reported how they wanted to cheer him up, and they viewed the task as a game. This leads us to speculation how SALs may be considered less “human-like.” (notion tackled in [6], p. 319). However, these results should be taken with caution due to potential bias introduced by our experimental design.

On the other hand, the paired t-test has shown no significant difference between judged rapport with Obadiah and judged rapport with Poppy. We observed some subjects whose intentions to cheer up Obadiah lead to frustration towards the end of interaction and accordingly, with low self-reported rapport. Interestingly, some of these examples are judged as interactions with high rapport. We designed the crowdsourcing study under the assumption that judges can access the rapport by watching short interaction clips [13]. To prove whether this applies for human-agent interaction,

TABLE 5

Descriptive statistics of self-reported rapport and rapport judged by external observers.

	M	SD	MD	Min	Max
Self-reported, Obadiah	2.38	0.69	2.28	1.28	3.93
Judged, Obadiah	2.82	0.47	2.77	1.68	3.54
Self-reported, Poppy	3.36	0.83	3.42	1.28	4.64
Judged, Poppy	2.95	0.53	2.84	1.91	4.24

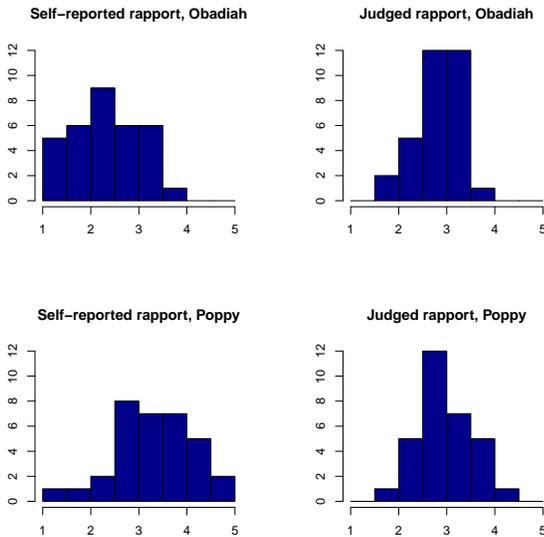


Fig. 4. Histograms of self-reported rapport and rapport judged by external observers.

a more thorough study has to be performed. Nevertheless, this could be an explanation why the paired t-test revealed that self-reported rapport with Obadiah ( $M = 2.35$ ,  $SD = 0.69$ ) is on average significantly lower than the judged one ( $M = 2.82$ ,  $SD = 0.47$ );  $t(31) = -2.98$ ,  $p < 0.05$ . On the contrary for Poppy, the t-test indicated that self-reported rapport ( $M = 3.36$ ,  $SD = 0.83$ ) is on average significantly higher than the judged rapport ( $M = 2.95$ ,  $SD = 0.53$ );  $t(30) = 2.3$ ,  $p < 0.05$ . These results show that: a) after watching videos with Obadiah, external observers judge higher rapport than rapport reported by the HINT subjects, and b) after watching videos with Poppy, external observers judge lower rapport than rapport felt by the HINT subjects.

Pearson's correlation analysis between the self-reported rapport and the judged one, which is distinctly done for Obadiah and for Poppy, did not reveal any significant results. Findings on rapport judgments done by Bernieri and his colleagues on human-human debate interaction [13], reveal weak relationship between external judgments and self-reported rapport. They found that a level of judge's accuracy in which group of 50 trained observers judged 50 video clips was below meaningful. Only the subset of 14 observers was able to achieve the accuracy significantly higher than chance. One possible explanation given by researchers are individual differences of personality of judges, meaning that people who are motivated to understand others, and have good social skills, perform better on interpersonal communication task.

### 5.3 Correlation analysis between rapport and extracted social cues and personality

#### 5.3.1 Verbal cues and rapport

A conversation with Obadiah on average results with fewer words spoken by a subject ( $M = 275.18$ ,  $SD = 112.3$ ,  $MD = 250$ ), than a conversation with Poppy ( $M = 310.24$ ,  $SD = 109.19$ ,  $MD = 296$ ), as confirmed by the paired t-test ( $t(32) = -3.65$ ,  $p < 0.001$ ). To test whether a subject's language style characterizes the presence of rapport, we computed Pearson's correlation between the LIWC categories and rapport scores per interaction, the self-reported and the judged scores. Table 6 reports the significant results. When it comes to judgments of rapport, the most significant LIWC category is word count (WC) ( $p < 0.005$ ). Subjects who talk more are judged to have higher rapport, or being more involved in interaction with both agents. This is in direct relation to our findings on turn-taking (see Section 5.3.2). Other LIWC categories for agent Obadiah suggest how interactions in which people are oriented towards the future and their achievements (low I, high Future and Achieve), while keeping the positive attitude (high Conjunction and Tentative, complemented with facial expressions of Joy), and perhaps also trying to cheer up the agent (high Discrepancy and Sad, complemented with facial expressions of Joy) are judged to have high rapport. When it comes to LIWC categories that characterize judgments of rapport with Poppy, less pronouns (Total pronouns and Personal pronouns), less present tense (Present), less words like "observe" or "feeling" (Perceptual Processes), more prepositions and words related to spatial movements (Relativity and Space) are found to be significant. These categories suggest that interactions in which a subject is focused on describing the present events and plans, are judged to have high rapport by external observers.

For the self-reported rapport with Obadiah, one can notice very low cue utilization, not only with LIWC categories, but also with other social cues (Table 6). One possible, albeit speculative, explanation could be that subjects freely expressed themselves while interacting with Poppy because they knew what is expected from them. As mentioned in Section 3.3, a certain degree of awkwardness may be contained in an early phase of "strangers meet" interaction. For Obadiah, only one significant LIWC cue was found - Cognitive processes (.35), which indicates a high utilization of words like "cause", "know", "ought". This weakly supports the notion that people who aim to cheer up Obadiah report high rapport. For the self-reported rapport with Poppy, more significant categories were found. Subjects who say "I" more frequently, do not negate, swear less, and use less angry words, report higher rapport with Poppy. These are all indicators of positive interactions with cheerful Poppy.

Except for LIWC, we have also investigated the most frequent words spoken per interaction. For each individual interaction 30 most frequent words are computed and combined within each conversation scenario. Then, two word clouds are generated. Results, shown in Figure 5, show almost no difference between interactions with Obadiah and with Poppy. Words like "yeah", "well", "good", "like", "happy" are the most spoken words in both scenarios.



TABLE 6

Significant Pearson correlation effects between rapport and different features in interactions with Obadiah and Poppy ( $p < 0.05$ ,  $*p < 0.01$ ). See social cue acronyms in Table 3

	Obadiah		Poppy	
	Self-reported rapport	Judged rapport	Self-reported rapport	Judged rapport
Verbal Cues	Cognitive processes (.35)	Word Count (.51), Words >6 letters (.41) I (-.67)* Future (.52) Prepositions (-.35) Conjunctions (.45) Causation (-.38) Number (-.38), Sad (.45) Discrepancy (.51) Tentative (.66)* Achieve (.46)	I (.39), Negate (-.46)  Swear (-.49) Anger (-.50), Feel (-.45) Discrepancy (-.35) Body (-.35)	Word Count (.47)  Total pronouns (-.36) Personal pronouns (-.39) Present (-.39) Prepositions (.43)  Perceptual processes (-.36) Relativity (.42) Space (.42)
	# Features: 1	# Features: 12	# Features: 7	# Features: 8
Nonverbal Cues	ATD (-.39), ALSS (-.48)	TSL (.38), TST (.49) PausePT (-.41) PauseNS (.43) PauseAD (-.50), PausePTS (.43) F0 (m) (-.41), Loc R0 (m) (.44), HOT (.36), ST (-.45) wMEI (m) (-.35), wMEI (md) (-.44) dwMEI (e) (-.35) JoyPT (.37) JoyNS (.35)	TST (.43) PauseNS (.60)* PauseAD (-.39) Val R0 (m) (.38) ALVS (-.34) BL (-.45)  DisgustPT (-.45) DisgustNS (-.42) SurpriseNS (-.35)	TSL (.43), TSTf (.49) PausePT (-.38) PauseAD (-.37) PausePTS (.37)  GOT (.53)  SmileAD (.43) AngerAD (-.38)
	# Features: 2	# Features: 15	# Features: 9	# Features: 8
Personality traits	Extr. (.55)* Agree. (.35)	Neur. (-.35)  Consc. (-.40)	Neur. (.47) Extr. (.40) Agree. (.39)	
	# Features: 2	# Features: 2	# Features: 3	# Features: 0

TABLE 7

Regression results for Obadiah and Poppy with different feature sets (Personality traits (PT), social cues (SC), all vs. significant cues). For each feature set we only show results of the best regression model.

	Feature Set	Meth.	R2	RMSE	
Obadiah	Self-reported rapport	SC+PT (sig.)	SVR	0.359	0.136
		PT (sig.)	KRR	0.177	0.155
	Judged rapport	SC (sig.)	SVR	0.126	0.160
		SC (all)	SVR	0.622	0.071
Poppy	Self-reported rapport	SC+PT (sig.)	KRR	0.618	0.075
		SC (sig.)	SVR	0.321	0.096
		SC+PT (all)	SVR	0.590	0.131
	Judged rapport	SC+PT (sig.)	KRR	0.561	0.136
		PT (all.)	SVR	0.406	0.158
		PT (all.)	SVR	0.152	0.121
			0.094	0.125	

To investigate features relevant for prediction of rapport, we trained the models with different feature sets. We experimented with (1) all extracted social cues (SC), verbal and nonverbal (2) all SC cues and all personality traits (PT), (3) all PT, (4) significant SC cues, (5) significant PT, and (6) significant SC and PT. Significant SC cues and PT are shown in Table 6. Missing SC cues for one subject, obtained from color images, are replaced with mean values (4 significant features).

Table 7 shows the results of our experiments, where we

report the  $R^2$  and Root Mean Square Error (RMSE). Among different feature sets and regression models that we have experimented with, we report the best results for each rapport measure. To stress the difference between experimented feature sets, we only show the results of the best regression model for a specific feature set. One can first notice that the best results are obtained when personality traits and social cues are combined, which boost the performance of each individual input source (PT and SC used alone).

For classification task we first segmented normalized rapport scores into two binary classes using the interval mean as threshold (0.5). We trained support vector machines (SVMs) in the manner same to regression; using double CV, leave-one-out CV approach for the outer fold, and 5-fold CV approach for the inner fold. We experimented with the same feature sets as in regression task, except new significant cues are obtained with Pearson's correlation between the exported cues and two binary rapport classes. To balance the feature sets we applied random oversampling technique. Classification Kappa statistics for all rapport measures is depicted in Figure 6. To compare results between the feature sets, we display Kappa values for classification task when we use all SC cues and PT, only significant SC, and only significant SC and PT (for judged rapport in interaction with Poppy, no significant personality is found). We observe how regression findings are translated to classification task; the best results for predicting rapport measures are obtained when both SC and PT are used together. The confusion

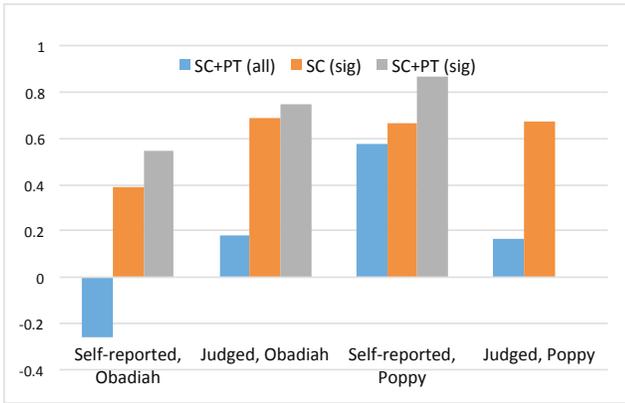


Fig. 6. The Kappa statistic of SVM models trained with different feature sets.

TABLE 8  
Confusion matrices of the best classification models for rapport measures.

Self-reported, Obadiah				Judged, Obadiah			
		Predicted				Predicted	
		(+)	(-)			(+)	(-)
True	(+)	6	4	12	2	12	2
	(-)	2	21	2	16	2	16
OA		82%		OA		88%	

Self-reported, Poppy				Judged, Poppy			
		Predicted				Predicted	
		(+)	(-)			(+)	(-)
True	(+)	21	2	11	3	11	3
	(-)	0	10	2	15	2	15
OA		94%		OA		84%	

matrices for the best classification models are given in Table 8. The results are on the stronger side, showing that prediction accuracy for each rapport measure is higher than 80%.

## 6 CONCLUSION

We have presented an experimental study on rapport prediction in two scenarios where human subjects are conversing with two virtual agents: sad Obadiah and cheerful Poppy. Two measures of rapport, a self-reported feeling of rapport, and rapport judged by external observers, are collected and correlated to self-reported personality traits and social cues displayed by a human subject. Non-trained observers are recruited in a crowdsourcing setting, in which they judged 1-minute video clips of human-agent interaction. Clips are extracted from longer sequences; based on psychological findings [13], rapport impressions on these short segments can be transferred to whole interaction.

To our knowledge, this is the first study about external judgments of rapport in human-agent interaction. Although the crowdsourcing results suggest that this could be a low-cost and fast approach for collecting annotations in human-agent studies, a further investigation is necessary to validate the annotations. Correlation analysis between the observers' judgments and social cues extracted from audio-visual data, has revealed that observers infer rapport from turn-taking and pauses in both interaction scenarios. Interactions with

Obadiah and Poppy in which there are more turn-takes, less pauses, and where a subject speaks longer are judged to have high rapport. When it comes to visual cues, subjects who are smiling more often and making more hand gestures are judged as highly connected to cheerful Poppy. For Obadiah, despite the fact that subjects with passive body movements, which are a sign of sadness [46], are judged to have a high rapport with sad Obadiah, a number of positive and joyful social cues are also found to be correlated. These are frequent and highly present facial expressions of joy displayed by the subjects, spoken content that is characterized by achievements, tentative words, and talk about future. Moreover, subjects who are low in conscientiousness and neuroticism are also judged to have high rapport with Obadiah. Based on these cues, we find that interactions in which people are not affected by Obadiah's gloomy state are judged as interactions with high rapport. These people try to cheer up sad Obadiah by talking joyfully about their future and achievements.

When it comes to a self-reported feeling of rapport, and its relationship to a human personality traits, agreeableness and extraversion are found to be significant. Subjects who scored high in agreeableness and extraversion have reported high rapport with both agents. This behavior is also found in human dyadic interaction; in a study that inspired our work [39], presence of at least one agreeable subject in a dyad resulted with mutual stronger feelings of rapport, and furthermore, extraverted subjects perceived the interaction as more enjoyable. In a previous study on human-agent interaction [2], agreeable subjects reported strong rapport with the Rapport agent. Related to this finding, our result suggests how human agreeableness plays a key role in self-perception of rapport while communicating with different virtual agents. Among social cues correlated to self-reported rapport, paralinguistic cues and turn-taking are found to be the most significant for both scenarios. Several significant visual cues are only inspected for agent Poppy. Some of those are related to psychological literature, such as back leans, which are indicators of lack of interest. Related to visual cue scarcity, it is important to note that, due to limited resources, subjects are assigned to the same experimental conditions, in which they first interacted with Obadiah and then with Poppy. This introduces a potential bias in the self-reported rapport with Poppy. Moreover, the unfamiliarity of subjects with the experiment when encountering Obadiah could also be the reason of high discrepancies between social cues correlating to the self-reported rapport with the two agents.

In the rapport prediction task the best results for all measures are obtained when social cues and personality traits are used together as input features for machine learning models. Best classification models have accuracy higher than 80%. Among these input features, several are computed from data gathered from low-cost markerless motion capture system (MS Kinect v1.8). More specifically, we used the MS Kinect upper body and face tracking information to build body lean and head direction classifiers, which yielded good and modest results respectively. Based on depth data, we have also proposed depth wMEI images, which are spatial motion images [30] extended with 3D information. Depth wMEI information was found to be

significant for the judged rapport.

There are several implications for future work. First, our rapport models are based on the whole interaction sequence: each of our features is extracted to summarize the overall interaction. Recent advances on multimodal fusion utilize sophisticated models that grasp the temporal dynamics, such as augmented conditional random fields (CRFs) [50], or temporal deep networks [51]. These models are trained on per-frame level, so the next step would be to address a methodology that is able to provide rapport labels on per-frame basis. Besides, temporal information about rapport and social cues can also be used to explore causality between those two variables. Second, despite the modest agreement between the crowdsourcing coders, rapport annotations obtained by non-experts have to be validated by hiring professional annotators. Finally, with regards to vision-based social cues, eye gazing is rather neglected. Head direction classifier built upon Kinect facial tracker did not yield any significant features, however, advances on eye gazing with depth data [52] suggest that eye patterns of our subjects can be estimated and integrated in prediction models.

To conclude, we hope that this study can generate research in somewhat two distinct areas. First, a variety of social cues can be explored for automatic communication analysis, such as for prediction of affect or depression labeled in the AVEC 2013 dataset [53]. Second, in addition to questionnaires and interviews, researchers working on agent evaluation studies can also explore potentials of automatic measurements.

## ACKNOWLEDGMENTS

This work was partly conducted while the first author visited Idiap. This work was partly funded by the Ministry of Science, Education and Sports of the Republic of Croatia, the Swiss National Science Foundation (SNSF) Ambizione project 'Multimodal Computational Modeling of Nonverbal Social Behavior in Face to Face Interaction' (PZ00P2-136811) and by grants from the Croatian Science Foundation (CSF), and Pascal 2 Network of Excellence.

## REFERENCES

- [1] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York, NY, USA: ACM, 2013, pp. 697–706.
- [2] S.-H. Kang, J. Gratch, N. Wang, and J. Watt, "Agreeable people like agreeable virtual humans," in *Intelligent Virtual Agents*. Springer Berlin Heidelberg, 2008, vol. 5208, pp. 253–261.
- [3] S. Kang, J. Watt, and J. Gratch, "Associations between interactants' personality traits and their feelings of rapport in interactions with virtual humans." *8th Internat. Conf. on Independent Component Analysis and Signal Separation*, 2009.
- [4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vision Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.
- [5] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. F. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners." *T. Affective Computing*, vol. 3, pp. 165–183, 2012.
- [6] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin, and C. Pelachaud, "Evaluation of four designed virtual agent personalities." *T. Affective Computing*, vol. 3, no. 3, pp. 311–322, 2012.
- [7] A. Cerekovic, O. Aran, and D. Gatica-Perez, "How do you like your virtual agent: Human-agent interaction experience through nonverbal features and personality traits," in *Human behavior understanding*, 2014.
- [8] A. M. von der Putten, N. C. Kramer, and J. Gratch, "How our personality shapes our interactions with virtual characters - implications for research and development," in *IVA'10*, 2010.
- [9] J. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *Third International Conference on Social Computing (SocialCom)*, Oct 2011, pp. 613–616.
- [10] Z. Yu, D. Gerritsen, A. Ogan, A. Black, and J. Cassell, "Automatic prediction of friendship via multi-model dyadic features." in *Proceedings of the 14th annual SIGdial Meeting on Discourse and Dialogue*, 2013.
- [11] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychological Inquiry*, vol. 1, pp. 285–293, 1990.
- [12] D. D. Gremler and K. P. Gwinner, "Customer-employee rapport in service relationships," *Journal of Service Research*, vol. 3, pp. 82–104, 2000.
- [13] F. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis." *Journal of Personality and Social Psychology*, vol. 71(1), pp. 110–129, 1996.
- [14] A. Papangelis, R. Zhao, and J. Cassell, "Towards a computational architecture of dyadic rapport management for virtual agents," in *Intelligent Virtual Agents*, 2014, vol. 8637, pp. 320–324.
- [15] F. J. Bernieri, J. Davis, R. Rosenthal, and C. Knee, "Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect," *Personality and Social Psychology Bulletin*, vol. 20, pp. 303–311, 1994.
- [16] J. N. Cappella, "On defining conversational coordination and rapport," *Psychological Inquiry*, vol. 1, no. 4, pp. 303–305, 1990.
- [17] B. M. DePaulo and K. L. Bell, "Rapport is not so soft anymore," *Psychological Inquiry*, vol. 1, no. 4, pp. 305–308, 1990.
- [18] D. Lakens and M. Stel, "If they move in sync, they must feel in sync: Movement synchrony leads to attributions of rapport and entitativity," *Social Cognition*, vol. 29(1), pp. 1–14, 2011.
- [19] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," in *Intelligent Virtual Agents*. Springer Berlin Heidelberg, 2011, vol. 6895, pp. 68–79.
- [20] J. A. Hall and F. J. Bernieri, Eds., *Interpersonal Sensitivity: Theory and Measurement*. Taylor & Francis, 2001.
- [21] T. Chartrand and J. Bargh, "The chameleon effect: the perception-behavior link and social interaction." *Journal of Personality and Social Psychology*, vol. 76(6), pp. 893–910, 1999.
- [22] G. M. Maxwell and M. W. Cook, "Postural congruence and judgments of liking and perceived similarity." *New Zealand Journal of Psychology*, vol. 15, no. 1, pp. 20–26, 1985.
- [23] M. Argyle, *Bodily communication*. Methuen, 1988.
- [24] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal Communication in Human Interaction*, 8th ed. Cengage Learning, 2013.
- [25] P. Ekman, W. V. Friesen, M. O'Sullivan, and K. R. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect," *Journal of Personality and Social Psychology*, vol. 38, pp. 270–277, 1980.
- [26] A. Mehrabian, "Significance of posture and position in the communication of attitude and status relationships." *Psychological Bulletin*, vol. 71, no. 5, pp. 359–372, 1969.
- [27] G. B. Ray and K. Floyd, "Nonverbal expressions of liking and disliking in initial interaction: Encoding and decoding perspectives," *Southern Communication Journal*, vol. 71, pp. 45–65, 2006.
- [28] A. S. Pentland, *Honest Signals: How They Shape Our World*. The MIT Press, 2008.
- [29] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010.
- [30] O. Aran, J.-I. Biel, and D. Gatica-Perez, "Broadcasting oneself: Visual discovery of vlogging styles," *Multimedia, IEEE Transactions on*, vol. 16, no. 1, pp. 201–215, Jan 2014.
- [31] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: Predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 2012, pp. 53–56.

- [32] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, and D. Gatica-Perez, "Capturing upper body motion in conversation: An appearance quasi-invariant approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 327–334.
- [33] D. Sanchez-Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez, "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 39–53, 2013.
- [34] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2012, pp. 28–35.
- [35] X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, and M. Pantic, "A multimodal database for mimicry analysis," in *Affective Computing and Intelligent Interaction*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer Berlin Heidelberg, 2011, vol. 6974, pp. 367–376.
- [36] A. Nelson, J. Grahe, F. Ramseyer, and K. Serier, "Psychological data from an exploration of the rapport / synchrony interplay using motion energy analysis," *Journal of Open Psychology Data*, vol. 2(1), p. 1, 2014.
- [37] A. Won, J. Bailenson, S. Stathatos, and W. Dai, "Automatically detected nonverbal behavior predicts creativity in collaborating dyads," *Journal of Nonverbal Behavior*, vol. 38, no. 3, pp. 389–408, 2014.
- [38] O. Aran and D. Gatica-Perez, "One of a kind: inferring personality impressions in meetings," in *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, 2013, pp. 11–18.
- [39] R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions: personality similarity helps extraverts and introverts, but hurts "disagreeables"." *Journal of Personality and Social Psychology*, vol. 97, no. 4, pp. 667–684, 2009.
- [40] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Evaluating online labor markets for experimental research: Amazon.com's mechanical turk," *Political Analysis*, vol. 20(3), pp. 351–368, 2012.
- [41] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, 2010.
- [42] D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez, "Assessing the impact of language style on emergent leadership perception from ubiquitous audio," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, 2012, pp. 33:1–33:8.
- [43] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features," *Speech Communication*, vol. 54, no. 1, pp. 55 – 67, 2012.
- [44] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 298–305.
- [45] Y. Yabar and U. Hess, "Display of empathy and perception of out-group members." *New Zealand Journal of Psychology*, vol. 36(1), pp. 42–49, 2007.
- [46] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.
- [47] N. Dael, M. Mortillaro, and K. Scherer, "Emotion expression in body action and posture," *Emotion*, vol. 12(5), pp. 1085–1101, 2012.
- [48] D. Funder and C. Sneed, "Behavioral manifestations of personality: an ecological approach to judgmental accuracy." *Journal of Personality and Social Psychology*, vol. 64, no. 3, pp. 479–490, 1993.
- [49] K. Isbister and C. Nass, "Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics," *International Journal of Human-Computer Studies*, vol. 53, no. 2, pp. 251 – 267, 2000.
- [50] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Affect analysis in natural human interaction using joint hidden conditional random fields," in *International Conference on Multimedia and Expo (ICME '13)*, 2013.
- [51] M. R. Amer, B. Siddiquie, A. Tamrakar, D. A. Salter, B. Lande, D. Mehri, and A. Divakaran, "Human social interaction modeling using temporal deep networks," *Computers and Society*, 2015.
- [52] K. Funes Mora and J.-M. Odobez, "Gaze estimation in the 3d space using rgb-d sensors. towards head-pose and user invariance," *International Journal of Computer Vision*, 2015.
- [53] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous

audio/visual emotion and depression recognition challenge," in *Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 3–10.



**Aleksandra Cerekovic** is a Postdoctoral researcher at the University of Zagreb, Croatia. She has been visiting researcher at the Idiap Research Institute (Martigny, Switzerland), working in the area of multimodal behavior modeling and analysis. Her research interests are automatic multimodal communication analysis, social signal processing, pattern recognition and virtual agents.



**Oya Aran** received her PhD degree in Computer Engineering from Bogazici University, Istanbul, Turkey in 2008. She was awarded a EU FP7 Marie Curie IEF fellowship in 2009 and a Swiss National Science Foundation Ambizione fellowship in 2011. Currently, she is a scientific collaborator at the Idiap Research Institute. Her research interests include pattern recognition, computer vision, and social computing. She is a member of the IEEE.



**Daniel Gatica-Perez** (S'01, M'02) directs the Social Computing Group at Idiap Research Institute and is Professeur Titulaire at the Ecole Polytechnique Federale de Lausanne (EPFL) in Switzerland. His research interests include social computing, social media, and ubiquitous computing. He is a member of the IEEE.