Connecting Meeting Behavior with Extraversion—A Systematic Study

Bruno Lepri, Ramanathan Subramanian, *Member*, *IEEE*, Kyriaki Kalimeri, *Student Member*, *IEEE*, Jacopo Staiano, *Member*, *IEEE*, Fabio Pianesi, and Nicu Sebe, *Senior Member*, *IEEE*

Abstract—This work investigates the suitability of medium-grained meeting behaviors, namely, *speaking time* and *social attention*, for automatic classification of the **Extraversion** personality trait. Experimental results confirm that these behaviors are indeed effective for the automatic detection of Extraversion. The main findings of our study are that: 1) Speaking time and (some forms of) social gaze are effective indicators of Extraversion, 2) classification accuracy is affected by the amount of time for which meeting behavior is observed, 3) independently considering only the attention received by the target from peers is insufficient, and 4) distribution of social attention of peers plays a crucial role.

Index Terms—User/machine systems, human-information processing, vision and scene understanding, psychology

1 Introduction

Social and personality psychology has long been concerned with Extraversion, which has emerged as a fundamental dimension of personality [1] because it is 1) capable of explaining a wide range of behaviors, 2) able to predict functioning across a number of domains, ranging from cognitive performance [2] and social endeavors [3] to socio-economic status [4], and 3) useful for assessing the risk of different types of psychopathology [5].

At the same time, the importance of determining people's personality for technology and human-computer interaction in particular has been widely acknowledged. Studies have shown that personality traits influence the basic dimensions of adaptivity [6] and determine people's attitude toward computers in general [7] as well as conversational agents [8], [9]. It has been argued that social networking websites can increase the chances of a successful relationship by analyzing text messages and matching personalities [12], and that tutoring systems would be more effective if they adapt to the learner's personality [10]. Moreover, given its relevance in social settings, information on people's personalities and their Extraversion levels can be useful for providing personalized support to group dynamics [11]. More generally, letting computers understand people's personality can be an important step toward endowing them with the folk-psychological capability [12] of explaining and/or predicting people's behavior.

Recent psycho-social work has emphasized the importance of social attention for Extraversion; Ashton et al. [13] argued that at the core of Extraversion lies the behavioral tendency to engage, attract, and enjoy social attention, i.e., extraverts invest time and energy in activities that attract others' attention. It should therefore be possible to exploit attention patterns to predict a person's Extraversion level. For example, if the social gaze of peers is affected by the target's speaking patterns, one can expect the amount of social attention he/she receives to depend on his/her extraversion level. Furthermore, it is likely that Extraversion influences the social-attention patterns of the target, too, so that the latter can be predictive of his/her Extraversion levels. However, available evidence is mixed on this subject. Some studies [14], [15] have concluded that extraverts gaze more frequently at their peers, and with longer glances while talking; others [16], while confirming that extraverts look more frequently at peers than introverts, do not agree that this necessarily translates to more social attention directed at peers.

In this paper, we take inspiration from the social attention view to inform the task of automatically predicting the Extraversion trait of participants in small group meetings. To this end, we consider two kinds of behavior capable of attracting others' attention: the person's speaking time (ST) and the amount of social attention given to other group members (AG). The latter is further distinguished into attention devoted while the person is speaking ($AG1$) and is silent ($AG2$). This distinction is meant to account for the possibility that differential variations in those quantities are good cues for Extraversion. Together with the target's behaviors, we also consider the total amount of social attention those behaviors elicit from others (AR). As before, this appears in two guises: the amount of attention the person receives while he/she is speaking ($AR1$) and is silent ($AR2$). All the above social attention quantities are measured from

- *B. Lepri is with the MIT Media Lab and FBK-Foundation Bruno Kessler, via Sommarive 18, Povo, Trento 38123, Italy. E-mail: lepri@fbk.eu.*
- *R. Subramanian, J. Staiano, and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento, via Sommarive 14, Povo, Trento 38100, Italy. E-mail: {subramanian, staiano, sebe}@disi.unitn.it.*
- *K. Kalimeri is with the University of Trento and FBK-Foundation Bruno Kessler, via Sommarive 18, Povo, Trento 38123, Italy. E-mail: kalimeri@fbk.eu.*
- *F. Pianesi is with FBK-Foundation Bruno Kessler, via Sommarive 18, Povo, Trento 38123, Italy. E-mail: pianesi@fbk.eu.*

Manuscript received 7 June 2011; revised 17 Feb. 2012; accepted 16 May 2012; published online 29 May 2012.
Recommended for acceptance by C. Lisetti.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2011-06-0042.
Digital Object Identifier no. 10.1109/T-AFFC.2012.17.

the proportion of eye-gaze (over the analyzed time frame) a subject directs toward a particular peer.

As already remarked, personality is not only a theoretical construct but also a dimension routinely used in our daily life. Humans are capable of forming accurate impressions of others' personality upon observing very short sequences of expressive behavior termed *thin slices* [17]. We model our task in a similar manner: We first discretize the continuous scale of Extraversion into High (extravert) and Low (introvert) classes, and then attempt classification by using thin slices of speaking activity and social gaze. In relevant respects, the task is similar to what humans are routinely involved in in judging others' personalities from very short behavioral sequences.

We also investigate whether the thin slice size has an impact on the prediction of Extraversion, i.e., whether the use of longer time slices produces better classification results. Current literature on human personality assessment is not decisive on this point. Meta-analytic studies [17] did not find any significant effects of exposure time on judgment accuracy for many target constructs. Conversely, Blackman and Funder [18] found a significant linear increase in agreement between observers' and targets' assessments as a function of exposure time, with the correlation increasing from $r = 0.22$ to 0.26 when the exposure increased from 5-10 to 25-30 minutes. Furthermore, Carney et al. [19] found a significant linear increase in the assessment accuracy of Extraversion when the slice-length increased from 5 to 300 seconds. In this work, we address the impact of thin slice size on the automatic classification of Extraversion considering 1, 2, 3, 4, 5, and 6 minute-long video slices.

Finally, one also expects the observers' judgment to be influenced by the amount of information they can access. Most thin-slice studies based on human first impressions let judges access the full social context through videos where both the target subject and his/her meeting peers were present. The properties of the social context can actually modulate the behavioral expression of personality traits; hence, the joint consideration of the target's characteristics and social context can contribute to improved Extraversion prediction accuracy. The total attention received by the target from others (AR) inherently provides some information on the social context. However, this information is very coarse grained and fine-grained detail about the speech behavior or the attention patterns of peers could be of some utility for Extraversion prediction. To account for this possibility, we systematically compare several conditions where only target-specific behavior is considered against those also including behavior of each of the peers.

In summary, this paper addresses the automatic prediction of Extraversion, where prediction is modeled as a classification task employing 1) speaking time and 2) social attention (received and given) features extracted from thin meeting video slices. The possible effects of the social context (others' behavior) in modulating the expression of Extraversion are systematically accounted for, as is the possible influence of slice length on classification performance.

The remainder of the paper is organized as follows: The next section reviews related work. Section 3 describes the Mission Survival corpus used in our experiments, and methods adopted for manual and automatic speech and eye-gaze feature extraction, while correlations between some of these features and Extraversion are discussed in Section 4. Section 5 outlines the methodology employed for Extraversion classification, while Section 6 discusses experiments and results. Results are further elaborated in Section 7, and we conclude with key observations and directions for future research in Section 8.

2 Related Work

Pioneering work addressing automatic recognition of personality was done by Argamon et al. [20], who used the relative frequency of function words and word categories based on systemic functional grammar (SFG) to train support vector machines (SVMs) for the recognition of Extraversion and emotional stability. Data concerning the two personality traits were based on self-reports. Mairesse and Walker [21] applied classification, regression, and ranking models for the recognition of the Big Five personality traits. They compared personality self-reports with observed data and showed that 1) Extraversion is the easiest personality trait to model from spoken language, 2) prosodic features play a major role, and 3) obtained results were closer to observed personality than to self-reports.

Olguin et al. [22] collected various behavioral measures from the daily activities of 67 professional nurses in a hospital. The data were collected by means of the sociometer badge, a wearable device integrating a number of sensors measuring aspects such as physical and speech activity, number of face-to-face interactions, level of proximity to relevant objects (people, but also beds, etc.), and social networks parameters. They demonstrated that the acquired signals provided a lot of information about personality through correlation analysis.

Recently, Pianesi et al. [23] and Lepri et al. [24] showed the feasibility of automatically recognizing the Extraversion and LoC personality traits using simple nonverbal features. Kalimeri et al. [25] employed audio-visual features to study classification performance of two models incorporating theoretically motivated hypotheses about personality-behavior relationships. Also, Mohammadi et al. [26] showed how prosodic features extracted from 640 speech samples can be used to assess personality traits with up to 75 percent accuracy.

3 The Mission Survival Corpus

The Mission Survival corpus contains video recordings of 12 round-table meetings spanning 6 hours, where each meeting involves four people engaged in the Mission Survival Task (MST). The MST [27] is often used in experimental and social psychology to elicit decision making in small groups. It promotes group discussion by asking participants to reach a consensus on how to survive a disaster scenario like a ship wreck or a plane crash in the mountains. The group has to rank up to 15 items according to their importance for the crew members' survival.

A consensus decision-making scenario is chosen and enforced because intensive engagement is required to reach an agreement, thereby offering the possibility to observe a

large set of social dynamics and attitudes. In consensus decision-making processes, each participant expresses his/her opinion, which is evaluated by the group. In our case, consensus is enforced as the item a person proposes would become part of the common list only if he/she manages to convince the others of its utility. Also, we added an element of competition by awarding a prize to the individual who proposed the greatest number of appropriate and consensually accepted items.

The meetings were recorded in a specially equipped room by means of four FireWire cameras placed in the corners and four actively driven web cameras (PTZ IP cam) installed on the walls surrounding the table. Four wireless close-talk microphones (one for each participant) and one omnidirectional microphone placed on the table around which the group sat were used to record speech. Forty-eight volunteers (25 males, 23 females; average age: 35 years) participated in the meetings. Besides involving them in the MST, we also asked the participants to fill the Big Five Marker Scales (BFMS) that measures Extraversion [28].

3.1 Manual and Automated Extraction of Speech and Social Attention Features

For every participant in 10 meetings chosen from the Mission Survival corpus, judges manually annotated 1) if the target is speaking or is silent and 2) the direction of the target's social attention in every video frame. The 10 meetings were selected so as to comply with some tracking criteria, such as sufficient facial visibility throughout the meeting. Each frame of the meeting video was labeled as "speaking" or "nonspeaking." Social attention states were labeled according to whether the person's gaze was directed toward the member on his/her left ("L"), right ("R"), directly opposite ("O"), or self ("S"). The social attention label "S" denotes the state where the target is examining his/her list of items and is not looking at others.

From the manual annotations, the following features were computed for each participant p in each time slice:

- Speaking time (ST_p): The percentage of time, over the slice duration, for which individual p is speaking.
- Attention received while speaking ($AR1_p$): The percentage of p's speaking time during which he/she receives social attention from all the other group members.
- Attention received when not speaking ($AR2_p$): The percentage of p's nonspeaking time during which he/she receives social attention from all the peers.
- Attention given (AG_p): The percentage of time for which p is directing his/her social attention at one of the other participants.
- Attention given while speaking ($AG1_p$): The percentage of p's speaking time during which he/she devotes social attention to one of his/her peers.
- Attention given while not speaking ($AG2_p$): The percentage of p's nonspeaking time during which he/she looks at one of the others.

Overlapping windows (thin slices) of length 1/2/3/4/5/6 minutes and shifted by 0.5/1/1.5/2/2.5/3 minutes, respectively, so as to span the entire meeting duration, were considered for analysis. Then, the mean and standard deviation for each of the above features, computed for every subject, were used in the classification experiments. Now, we briefly describe the techniques employed for automatically extracting the aforementioned features.

3.1.1 Speaking Activity

The long-term spectral divergence (LTSD) algorithm [29] is used to distinguish between speaking and nonspeaking frames. We assume that the most significant information is contained in the time-varying spectral magnitude of the signal for vocal activity detection. The long-term spectral envelope (LTSE) and the long-term spectral divergence are estimated in order to formulate the decision rule for voice activity detection. Let $x[n]$ be the input signal which is sampled using fixed-size overlapping windows, resulting in $l = 1 \dots L$ frames. Let $X(k, l)$ be the amplitude spectrum for frequency band k in frame l. The N-order LTSE for frame $x[l]$ is defined as

$$LTSE_N(k, l) = max\{X(k, l + j)\}_{j=-N}^{j=N}. \quad (1)$$

The N-order LTSD between speech and noise is defined as the deviation of LTSE with respect to the average noise spectrum magnitude $N(k)$ for the kth frequency band, with $k = 0, 1, \dots N_F - 1$, where N_F is the length of the Fast Fourier Transform:

$$LTSD_N(l) = 10log_{10}\left(\frac{1}{N_F}\sum_{k=0}^{N_F-1}\frac{LTSE_N^2(k, l)}{N^2(k)}\right). \quad (2)$$

The decision rule for speech activity detection is based on the LTSD between speech and noise, where the threshold distinguishing speech from nonspeech regions is adjusted to maximize the accuracy with respect to the manually annotated ground-truth data. Employing this decision rule, speaking frames of all participants are labeled as "1" and nonspeaking frames as "0."

3.1.2 Gaze Direction

The gaze direction of a subject, a reliable indicator of his/her social attention, is computed by fusing head pose and eye location information by means of the eye-localization scheme proposed in [30], which is able to accurately extract head-pose and eye-center locations from a monocular video sequence. The method combines a robust cylindrical head model (CHM)-based pose tracker [31] and an isophote-based eye-center locator [32] to obviate the limitations of both methods when considered independently. The system integrates the eye locator with the CHM by interleaving the transformation matrices obtained by both systems. This way, eye-center locations are estimated given the pose and, conversely, pose is adjusted given the eye-center locations: Feedback from the CHM-based pose tracker is used to refine the eye locator's output in extreme head-pose situations, while the detected eye locations are employed to recover the head pose when pose-tracking fails. An illustration of the CHM with the eye-center locator is presented in Fig. 2a.

For estimating the gaze direction, we adapt [30] as follows: The 2D eye-center locations detected on a frontal face are used as reference points to initialize the CHM and used for eye-center estimation in subsequent video frames.

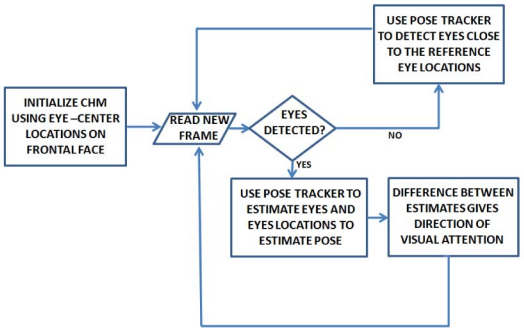

Fig. 1. The gaze-direction estimation process.

The estimated eye centers are projected onto a view-normalized face model and therefore the displacement between the reference and detected eye-center locations is independent of head pose. We approximate the gaze direction as horizontal and vertical shifts of the eye centers from their resting positions on the head surface. The displacement vectors for the two eyes are averaged to compute the gaze-based social attention estimate as a differential of the head pose. The eye-gaze direction estimation process is outlined in Fig. 1.

The eye-gaze direction provides a more accurate estimate of social attention, as shown in Figs. 2b and 2c. Upon computing the gaze direction, we map it to social attention targets "L," "O," "R," or "S" (as for the manual annotations) by employing a Bayesian approach as in [33]. Possible attention targets for each participant are represented as states in a Hidden Markov Model. By exploiting a priori knowledge of others' positions, the observation probability distributions for various states are modeled using Gaussians centered at the expected viewing angle for different targets (including self-attention). Once the social attention states are determined for each meeting participant over the entire meeting duration, the aforementioned social attention features are computed automatically.

3.1.3 Accuracy of Automated Speech and Eye-Gaze Annotation

Comparing the computed label for each speaking frame against the ground truth, the speech detection algorithm is found to be 92 percent accurate, which is sufficient for our purposes. Regarding social attention, employing only head pose to determine the target's attention direction (i.e., social attention feature AG_p), we obtained a mean accuracy of 76.3 percent with respect to observers' annotations for the 10 meetings. The accuracy improved to 82.4 percent when eye-gaze orientation computed relative to the head pose is considered as the social attention cue. Therefore, social attention estimation improves by about 8 percent when eye-gaze direction is employed in lieu of the head pose for our data set.

Even as the performance of gaze-based attention estimation is only moderately high, the obtained accuracy values need to be considered in the light of the challenging nature of the data set (sudden head-tilting, varying illumination, etc.). A closer analysis of the obtained results reveals that a majority of the estimation errors occur as both the pose tracker and eye-center detector fail when there is sudden

Fig. 2. (a) Combining cylindrical head model with eye-center locator. Yellow dots denote eye locations estimated by the view-normalized CHM, while green dots denote detected eye centers. (b), (c) Examples where gaze direction (green circle) differs from head pose (blue normal).

downward tilting of the head as participants turn to examine their lists, and when targets wear thick glasses, resulting in unreliable eye-center detection owing to specular reflections. Nevertheless, social attention features computed from automatic gaze estimation are still sufficient to replicate most results obtained from manual annotation, as discussed in Section 6.

4 Corpus Analysis

This section presents descriptive analyses of the corpus employing manually extracted speech and social attention features. We will first investigate the relationships between the features discussed in the previous section and the Extraversion score and then turn to analyzing the relationships between behavioral features of the target (ST, $AG1$, and $AG2$) and the global attention patterns of peers ($AR1$ and $AR2$).

We analyzed the relationships between behavioral features and Extraversion scores by means of a series of backward linear regression analyses, with the Extraversion score as the dependent variable and the mean (μ) and standard deviation (σ) of speaking time (ST), attention given while speaking ($AG1$) and while silent ($AG2$), and attention received while speaking ($AR1$) and while silent ($AR2$) as predictors. The μ and σ measures are computed over windows obtained by uniformly dividing the slice-length into 30 second segments. We restricted our attention to slices of length 2 ($T2$), 4 ($T4$), and 6 ($T6$) minutes. The first row in Table 1 reports the variables that were discarded by the backward procedure as they did not significantly contribute to explain the dependent variable's variance. The best model is the one for T4, where about 16 percent of the total variance for Extraversion is explained.

The partial correlations between the predictors and the Extraversion score for the $T4$ linear regression model are reproduced in Table 2. Partly supporting our expectations, at $T4$, extraverts give and receive more attention while

TABLE 1
Backward Linear Regression Analyses with Slices of Length 2, 4 and 6 Minutes

	$T2$	$T4$	$T6$
Variables removed		$AG2_\sigma$	$ST_\mu, AG1_\mu, AG2_\mu$
R	.301	.405	.270
R^2	.091	.164	.073

TABLE 2
Partial Correlations for the Best Linear Model at $T4$

ST_μ	ST_σ	$AG1_\mu$	$AG1_\sigma$	$AG2_\mu$	$AR1_\mu$	$AR1_\sigma$	$AR2_\mu$	$AR2_\sigma$
-.127	.101	.159	.166	-.167	.224	-.212	-.188	.086

speaking than introverts, but they give and receive less attention while silent, a pattern that is reminiscent of findings in [34] concerning the visual behavior of dominant versus nondominant people. Quite unexpected though is the weak inverse relationship between speaking time (ST) and the Extraversion score.

In summary, the analysis of features extracted from manual annotations seems to comply with some of the expectations derived from the social attention view of Extraversion. Extraverts actually give and receive more attention while speaking; when silent, in turn, they give and receive lower amounts of attention as compared to introverts. The analysis of the relationships between the target's behaviors and the attention patterns of his/her peers suggests that 1) the effects of ST and AG on the attention received do not compound, 2) ST is positively associated with both $AR1$ and $AR2$, while 3) $AG1$ and $AG2$ are positively associated only with $AR1$. We also found an unexpected trend: the weak inverse relationship between speaking time and Extraversion level.

We conclude this section with a note of caution: The linear model we have just discussed (for $T4$) only explains a small portion (16.4 percent) of variance. Moreover, the partial correlation values reported in Table 2, though significant, never indicate a strong linear relationship. It might well be the case, therefore, that other important relationships have gone unnoticed in the current analyses.

5 Automatic Classification of Extraversion

In this section, we describe the setup employed for Extraversion classification. First, the Extraversion scores were quantized into two classes, LOW and HIGH, of equal size based on the median score (42). For the target, the speaking time and the two forms of attention given to others ($AG1, AG2$) are considered.

According to [13], extraverts enjoy being the target of other's attention and, assuming that extraverts are successful in attracting their peers' attention, it is important to consider how much the others reciprocate the target's attempts by directing their social attention to him/her in terms of $AR1$ and $AR2$. As stated earlier, we also consider the possible contribution of social context (in the form of each peer's ST and AG) to Extraversion prediction. This move is motivated by the well-known observation that, generally, the social context modulates behavioral expression of personality traits. For example, we might expect that, besides being influenced by his/her Extraversion level, the attention a person devotes to others also varies according to the group behavior. Finally, we also investigate the effect of the slice length on Extraversion classification. The resulting experimental design considers the following factors:

- *Time*—With six different thin-slice durations, $T1$, $T2$, $T3$, $T4$, $T5$, and $T6$, corresponding to the 1, 2, 3, 4, 5, and 6 minute-long slices described above.
- *Target*—We begin with six basic levels, each consisting of one of the following features: ST, $AG1$, $AG2$, $AR1$, and $AR2$. We then consider all combinations of these basic levels, resulting in a total of 31 possible conditions. We use the mean (μ) and standard deviation (σ) values for each of the aforementioned features, computed from relevant slices, for our analysis.
- *Others*—With levels ST, AG, $ST + AG$, and No_Feat. The first three levels are meant to account for the modulating effect of the social context on the behavioral expression of Extraversion. $Others = No\_Feat$, in turn, denotes the case where only features pertaining to the target are considered. Finally, the order of features corresponding to the other participants is standardized, and it is based on their seating relative to the target around the table—first the features of the subject to the left, then those of the person in front, and finally the person on the right.

To illustrate with an example, $(T2, ST, ST + AG)$ indicates the condition where 2 minute-long slices are considered, the information concerning the target is limited to (μ and σ of) ST, and the features for "Others" include (μ and σ of) ST and AG for all the peers. Notice that by considering combinations of features, the described experimental design allows for a simple form of multimodal fusion.

The bound-constrained SVM with RBF kernel was used for Extraversion classification. The cost parameter C and the kernel parameter γ were estimated through an inner leave-one-meeting-out cross validation on the training set of the first fold (i.e., the first nine meetings); the obtained parameters were kept fixed for the outer cross validation.[1] For testing, cross validation was performed using a leave-one-meeting out procedure: At each fold, the slices for one meeting were left out for testing, while slices corresponding to all other meetings were used for training.

6 Results

In this section, we discuss the classification results from different points of view. We first consider results obtained from features extracted from manual annotations, and then compare them with results obtained by employing automated methods. Furthermore, we investigate possible effects of the hierarchical nature of our data, where subjects are nested within groups, on the classification results. Finally, we discuss the distribution of correct classification instances employing Pearson residuals.

6.1 Results from Manually Annotated Data

Our first step involved filtering away results that were not significantly higher than those of the baseline classifier. For the latter, we chose the classifier that assigns slices to "Low" or "High" according to their prior probabilities (0.50),

1. We used the BSVM tool available at http://www.csie.ntu.edu.tw/~cjlin/bsvm.

TABLE 3
Classification on Ground-Truth Data: Average Accuracy Values for the Conditions of Type $(*, *, No\_Feat)$ Retained After the Selection Procedure

'Target'	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$	
$AR2$	.392	**.549**	**.571**	**.591**	**.583**	**.618**	.55
ST+$AR1$	.446	.478	**.575**	.547	**.564**	.548	.53
ST+$AR1$ +$AR2$	.506	**.55**	**.598**	**.58**	**.585**	.556	.56
	.45	.53	.58	.57	.58	.57	

Bold *figures mark accuracies higher than the criterion. Outer row/column contains marginal accuracies for the corresponding conditions.*

yielding an expected accuracy of 0.50. The filtering procedure was as follows:

- Regarding accuracy as the probability of success in a single Bernoulli trial (correct versus incorrect outcomes), we formulated the Null Hypothesis, according to which the distribution of hits and errors for each experimental condition follows a binomial distribution with expected success probability of 0.5. We employed the binomial test to verify the Null Hypothesis.
- Given that this test is sensitive to the number of trials (number of available slices) and the latter varies depending on the slice size, six series of tests were performed—one for each level of the "Time" factor. For each run, the significance level was fixed at $\alpha = 0.05$, with Bonferroni correction for multiple comparisons.
- The tests were one tailed and conducted only for conditions whose accuracy (Acc) values appeared higher than or equal to the expected value. The identified thresholds were $Acc_{T1} > 0.525$, $Acc_{T2} > 0.536$, $Acc_{T3} > 0.548$, $Acc_{T4} > 0.558$, $Acc_{T5} > 0.562$, $Acc_{T6} > 0.570$.

Upon applying the filtering procedure, no conditions of types $(*, *, ST)$ and $(*, *, ST + AG)$ passed the test. Of the remaining conditions, we retained only those that yielded significant accuracy values for at least two levels of the "Time" factor. The results are reported in separate tables, Table 3 for $(*, *, No\_Feat)$ and Table 4 for $(*, *, AG)$.

Starting from the conditions where the social context is not considered, only three assortments of features passed the significance test and the amount of attention the target receives from others plays a major role in all of them (Table 3). In particular, $AR2$ (attention received while silent) is alone capable of yielding significant classification performances. The amount of attention received while speaking ($AR1$), in turn, needs to be combined with speaking time in order to achieve significant performances.

None of the target's behavior descriptors, ST, $AG1$, and $AG2$, are effective when employed alone. This is consistent with the results in Section 4 where these features had low partial correlations with Extraversion (Table 2). It is interesting to note, however, that ST acquires predictive force when used in conjunction with the attention the target receives in conditions $(*, ST + AR1, No\_Feat)$ and $(*, ST + AR1 + AR2, No\_Feat)$.

TABLE 4
Classification on Ground-Truth Data: Average Accuracy Values for the Conditions of Type $(*, *, AG)$ Retained After the Selection Procedure

'Target'	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$	
ST	.523	**.559**	**.57**	**.658**	**.596**	**.601**	.59
ST+$AG1$	**.543**	.516	**.595**	**.611**	**.606**	.562	.57
ST+$AG2$	**.543**	**.554**	**.563**	**.663**	**.573**	**.607**	.58
ST+$AR1$	.524	**.548**	**.557**	**.639**	**.587**	**.621**	.58
ST+$AR2$	.515	.539	**.551**	**.649**	**.628**	**.581**	.58
ST+$AG1$+$AR1$	**.54**	.515	**.586**	**.592**	**.564**	**.59**	.56
ST+$AG1$+$AR2$	**.538**	.521	**.571**	**.605**	**.654**	**.551**	.57
ST+$AG1$+$AG2$	**.535**	**.541**	**.579**	**.620**	**.589**	.537	.57
ST+$AG2$+$AR1$	**.532**	**.565**	**.607**	**.643**	.546	**.601**	.58
ST+$AG2$+$AR2$	**.532**	**.561**	**.575**	**.667**	**.606**	**.621**	.59
ST+$AR1$+$AR2$	**.529**	**.568**	**.578**	**.649**	**.61**	**.61**	**.59**
ST+$AG1$+$AG2$ +$AR1$	**.541**	.527	**.604**	**.652**	**.583**	**.579**	**.58**
ST+$AG1$+$AG2$ +$AR2$	**.527**	.528	**.555**	**.612**	**.644**	.5511	.57
ST+$AG1$+$AR1$ +$AR2$	**.543**	.526	**.566**	**.589**	**.617**	.565	.57
ST+$AG2$+$AR1$ +$AR2$	**.533**	**.567**	**.583**	**.687**	**.633**	**.638**	**.61**
ST+$AG1$+$AG2$ +$AR1$+$AR2$	**.538**	**.539**	**.567**	**.625**	**.642**	**.593**	.58
	.53	.54	.58	.64	.61	.59	

Bold *figures mark accuracies higher than the criterion. Outer row/column contains marginal accuracies.*

Turning to the conditions where detailed forms of social context encoding are considered in Table 4, only the information about the distribution of the other parties' social attention ("Others"$= AG$) turns out to be effective. Moreover, the target's speaking activity: "Target"$= ST$ is present in all the conditions listed in Table 4. In other words, combining the target's speech activity with the attention patterns of peers provides a complex and reliable cue to Extraversion classification.

This result extends the datum reported in Table 3 concerning conditions such as $(*, ST + AR1, No\_Feat)$ and $(*, ST + AR1 + AR2, No\_Feat)$. Both $AR1$ and $AR2$ refer to amounts of social attention the target receives from the others; however, "Others"$= AG$ carries much more detail, capturing how each peer distributes his/her attention, and is not restricted only to the attention devoted toward the target. In order to further deepen our understanding of the retained data, we need methodologies that satisfy a number of requirements:

1. They should not rely on the assumption that the data are normally distributed. Indeed, our data are generated by a Bernoullian process and even if summary statistics such as the means of binomial distributions have a normal distribution in the limit, the underlying data are not normal as validated by normality tests. The nonreliance on the normality

TABLE 5
Generalized Estimating Equation Results for Automatically Annotated Data

Conditions	Effect	df	Wald χ^2
$(*,*,No\_Feat)$	*Time*	5	38.709^{***}
	*Target*Time*	10	46.929^{***}
$(*,*,AG)$	*Time*	5	19.868^{***}
	*Target*Time*	40	$4.114 \times 10^{9***}$

Superscript *** *denotes* $p < .001$. ***df*** *denotes the degrees of freedom.*

assumption eliminates many popular tests such as the t-test.

2. They must avoid the common problem of Type I error underestimation that is typical when many pairwise comparisons are performed. One hundred fourteen conditions survived the selection process, making for a very high number of possible pairwise comparisons. Setting an $\alpha = 0.05$ confidence level for each single test practically ensures that at least one of them will reject the null hypothesis when it is actually true. Although various kinds of adjustments for multiple comparisons (e.g., Bonferroni's) have been proposed to secure an acceptable overall confidence level, they tend to decrease the statistical power. It would be advisable to avoid abusing pairwise comparisons, limiting them only to cases which previous omnibus tests have shown to be promising.
3. Statistical dependencies across the various conditions of our design must be accounted for. The adoption of a leave-one-meeting-out strategy for validation makes it possible to limit dependencies to:
   a. those arising from the repeated usage of the same data across the various levels of the "Time" factor—e.g., conditions of type $(*, ST, AG)$ share the same data and
   b. dependencies stemming from (partial) data overlap across levels of "Target" and "Others," e.g., $(T3, ST + AG1, No\_Feat)$ and $(T3, ST, No\_Feat)$ share the data referring to speaking time features. We deal with those dependencies by treating "Time" and "Target" as repeated measure factors.
4. We should be able to assess whether classification performances are dependent on "Time" and/or "Target" and/or if there are any specific combinations of those factors yielding higher (or lower) performances. This should be achievable directly, without having to reconstruct those effects from multiple pairwise comparisons.

ANOVA satisfies requirement 2: By providing tests of significance for the main effects of various factors and for their interactions, it allows limiting the number of multiple comparisons to a few selected ones. Furthermore, the decomposition of effects into main and interaction ones satisfies requirement 4. Finally, ANOVA can easily accommodate within subject (repeated measures) designs, thereby addressing the concerns in requirement 3. It does not comply with requirement 1 though because of its normality assumption. Generalized Linear Models [35] allow for the required flexibility: They can accommodate various data distributions (including binomial, Poisson, multinomial, etc.); they make omnibus tests available; they also allow for an ANOVA-like effect-decomposition. Moreover, a variant of the generalized linear models known as Generalized Estimating Equations [36] can deal with within subject designs, hence addressing requirement 3 as well.

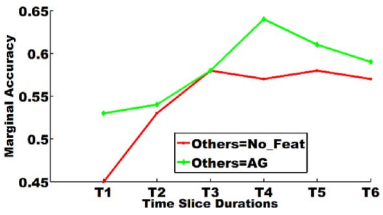

Fig. 3. Marginal accuracies for "Others"$= No\_Feat$ and "Others" $= AG$.

Roughly, Generalized Estimating Equations work by 1) transforming data according to the so-called "link" functions, 2) modeling the transformed data through suitable distributions, and 3) applying techniques like maximum likelihood or reduced maximum likelihood to simulate traditional linear, ANOVA-style modeling. For our analysis, we used the hits per subject, i.e., the number of correctly classified slices, as the dependent variable.

Below, we present two separate analyses for the retained Extraversion data: one for conditions of type $(*,*,No\_Feat)$, and the other for the conditions of type $(*,*,AG)$. Each analysis has two factors, "Time" and "Target," with six and three levels, respectively, for conditions of type $(*,*,No\_Feat)$, and six and 16 levels, respectively, for conditions of type $(*,*,AG)$. In both cases, "Time" and "Target" are treated as repeated measures factors. As the underlying process is Bernoullian, we exploit the logit as the link function, and the transformed data are modeled by means of the logistic distribution. The resulting model is $Count = Intercept + Time + Target + Target * Time$. In both cases, the results are discussed in terms of accuracy $(=$ average probability of hits per subject$)$.

As one can expect by inspecting the marginal accuracies in Tables 3 and 4, "Target" never yields a significant effect, irrespective of the value of "Others." Hence, the feature combinations for "Target" listed in Table 3 produce more or less identical performances, and the same is true for the "Target" levels listed in Table 4. For the latter case, it can be concluded that "Target"$= ST$ and "Others"$= AG$ are the feature combinations that jointly determine the responses obtained, even when more features are added to "Target."

"Time" has a pronounced effect, as shown in Table 5. A trend analysis reveals significant linear and quadratic trends both when "Others" $= AG$ (linear: $p < .05$, quadratic: $p < .05$) and when "Others" $= No\_Feat$ (linear: $p < .0001$, quadratic: $p < .0001$). As shown in Fig. 3, with conditions of type $(*,*,AG)$, the classification performance increases with slice length up to $T4$, and then slightly decreases. For $(*,*,No\_Feat)$, the increase stops at $T3$ and then performance remains largely stable for longer slices.

Analysis of the interactions in the two models does not convey anything interesting, and therefore we omit the

TABLE 6
Results from Generalized Estimating Equation Analyses

Conditions	Effect	df	Wald χ^2
$(*,*,AG)$	*Time*	5	11.524^{*}
	*Target*Time*	40	$3.756 \times 10^{9***}$

Superscripts ***, * denote $p < .001$, $p < .05$, respectively. ***df*** *denotes degrees of freedom.*

details here. Commenting on the results so far, we observe that although extraverts devote more attention to others than introverts while speaking ($AG1$) and less attention while not speaking ($AG2$), as seen in Table 2, neither cue alone is powerful enough to yield significant classification performance. Moreover, the initial expectation concerning speaking time is not born out: In our data, there is only a weak negative correlation between ST and the Extraversion level, and ST alone is not predictive for classification.

The attention received by the target, in turn, is indeed informative for predicting his/her Extraversion level. As seen from Table 2, an inverse correlation exists between the amounts of attention the target receives when silent ($AR2$) and his/her Extraversion level, and a positive correlation between Extraversion and $AR1$. Only the first relationship is reflected in the classification results: As seen for condition $(*, AR2, No\_Feat)$ in Table 3, a mere consideration of the amount of attention the target receives when silent is enough to attain 62 percent accuracy at $T6$. On the contrary, knowing only about the attention the target receives while speaking ($AR1$) is not enough to yield significant performance.

Detailed information about the attention behavior of *each of the others* has a significant impact on Extraversion classification. When "Others" $= AG$, the speaking time becomes a powerful predictor and maintains its power even when combined with other target-related features. In summary, of the two types of behaviors enacted by the target, only the speaking time is effective, although its efficacy manifests only in the presence of information concerning the attention behavior of others. Such a connection holds across the board: We have encountered it in conditions of type $(*, ST + AR1, No\_Feat)$, and $(*, ST + AR1 + AR2, No\_Feat)$, where only the information about the attention devoted by others to the target is considered, as well as in conditions of type $(*, ST, AG)$, where the details about the manner in which each peer distributes his/her social attention are incorporated (even if the actual attention direction is not considered).

These results comply with the social attention view of Extraversion, confirm its importance for automatic Extraversion prediction, and also provide new insights. It is not just that some systematic relation exists between people's speaking behavior and the attention they receive; the joint efficacy of "Target" $= ST$ and of "Others" $= AG$ suggests that the extraversion/introversion distinction *affects* others' attention patterns in a more general and complex way.

6.2 Results from Automated Analysis

In this section, we repeat the above procedures for automatically labeled speech and attention data. Virtually all the conditions of type $(*,*,AG)$ that survived the filtering procedure for manually annotated data are also retained with automatically annotated data, with the only exception of $(*, ST + AG2 + AR2, AG)$. On the contrary, none of the conditions of type $(*,*,No\_Feat)$ yield significant results in this case (see the results in Table 7).

TABLE 7
Classification on Automatically Extracted Features: Average Accuracy Values for the Conditions of Type $(*,*,AG)$ Retained After the Selection Procedure

'Target'	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$	
ST	.519	**.544**	**.572**	**.649**	**.594**	**.626**	.58
ST+$AG1$	**.544**	**.545**	**.587**	**.592**	**.628**	**.598**	**.58**
ST+$AG2$	**.534**	**.547**	**.553**	**.596**	.541	.562	.56
ST+$AR1$	.518	.530	**.566**	**.645**	**.569**	**.593**	.57
ST+$AR2$	.513	.500	**.553**	**.625**	.548	**.610**	.56
ST+$AG1$+$AR1$	**.535**	**.563**	**.583**	**.596**	**.608**	**.593**	.58
ST+$AG1$+$AR2$	**.531**	**.546**	**.564**	**.569**	**.592**	.559	.56
ST+$AG1$+$AG2$	**.532**	.529	**.583**	**.614**	**.569**	.553	.56
ST+$AG2$+$AR1$	**.525**	**.546**	**.555**	**.563**	.55	.556	.55
ST+$AR1$+$AR2$	.521	.518	**.551**	**.614**	**.564**	**.621**	.57
ST+$AG1$+$AG2$ +$AR1$	.535	.529	**.566**	**.591**	**.576**	**.612**	.57
ST+$AG1$+$AG2$ +$AR1$	.524	.525	**.556**	**.591**	.555	**.576**	.55
ST+$AG1$+$AR1$ +$AR2$	.528	.535	**.561**	**.6**	**.606**	.567	.57
ST+$AG1$+$AR1$ +$AR2$	**.53**	.526	.508	**.614**	.546	.565	.55
ST+$AG1$+$AG2$ +$AR1$+$AR2$	**.531**	.518	.533	**.592**	.548	**.584**	.55
	.53	.53	.56	.60	.57	.59	

Bold *figures mark accuracies higher than the criterion. Outer row/column contains marginal accuracies.*

The analysis of the retained data through generalized estimating equations shows a main effect of "Time" and a "Target*Time" interaction effect, as reported in Table 6. Polynomial contrast analysis on "Time" reveals a significant linear trend (Wald $\chi^2_1 = 3.912,\ p < .05$), indicating that the performance tends to increase with slice length. The peak we observed with manually annotated data at $T4$ is now less pronounced (0.60 versus 0.64 overall accuracy) though the difference between the two is not statistically significant. Further tests have not revealed other differences between manually and automatically annotated data; hence, we can conclude that the transition to automated features has not significantly affected performance obtained from manual annotation as far as conditions of type $(*,*,AG)$ are concerned.

6.3 Hierarchical Nature of the Data

The assessment of personality traits we have pursued exploits contextually determined behavioral manifestations to isolate contextually independent characterizations of people—i.e., their personality traits. In view of this dependence on the context, we should check whether the

performance of our classifier is affected by group membership: For a given set of features, does accuracy vary according to membership (the group subjects belong to)? Equivalently, are subjects from one group easier to classify than those of another group? Ideally, the accuracy should only depend on intersubject differences. If, on the contrary, classification is affected by unaccounted-for dependencies on group membership, then a significant portion of variability in accuracy should arise at the between group level, causing our confidence in the relevant personality assessments to decrease.

Although often neglected in the computational literature, this issue is quite general, arising any time subjects are nested within groups, with group membership possibly affecting classification accuracy. We should also be wary of an additional problem arising from the validation schema adopted in this paper. By leaving one whole meeting out at each fold, a possibly unbalanced training set (either in terms of personality labels or feature value distribution) can be produced at each fold. For instance, if the left-out group consists exclusively of extraverts, the High label could be underrepresented in the training set, possibly inducing a classification bias.

Both problems are due to the hierarchical structure of our data—the first problem can arise because of the posited substantive relationships between individual behaviors and the social context they take place in. The second problem is a bias induced by the particular validation schema chosen and the limited sample size. If either or both of those problems are present, we expect to measure significant between group variability in accuracy.

In order to ascertain whether this was the case, we ran a number of component variance analyses on the ground-truth restricted data set, fitting random intercept regression models with "group" treated as a random variable and "Time" as a repeated measure factor. We used a variance component (scaled identity) structure for the random part (groups), while no specific assumptions were made for the repeated measure ("Time"). The model was applied for each retained condition of types $(*, *, AG)$ and $(*, *, No\_Feat)$ to quantify the variance arising from group membership and check its statistical significance. The treatment of "group" as a random variable allows us to generalize the conclusion to the population [37]. In no case was the between group variance significant. Hence, the obtained classification accuracies are not affected by the hierarchical nature of our data.

6.4 Distribution of Hits (and Errors)

In order to further characterize the predictive power of our features and the behavior of the classifiers, it is important to understand whether they work equally well on the two classes. To this end, we start from the confusion matrix of each retained condition and compare the hits for the High and Low classes with those expected from the baseline classifier. The comparison is conducted by using Pearson residuals standardized scores $(N(0,1))$ that measure the difference between observed and expected outcomes (counts). Regarding hits, the absolute value of a Pearson residual measures how much better (positive sign) or worse (negative sign) the classifier performs than the baseline. For errors, the reverse is true. Here, we focus our attention on

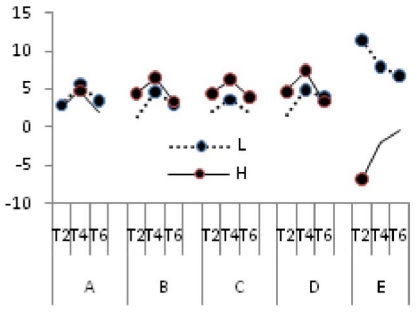

Fig. 4. Pearson residuals for the five selected conditions (respectively denoted by A-E). Dots denote values that are significantly different from the criterion ($\alpha = .05$, two tailed).

Pearson residuals for hits, interpreting them as a sort of recall measure standardized with respect to the baseline.[2]

Because of their standardization, Pearson residuals easily account for the difference in statistical significance that the same recall values can have, e.g., when generated from different sample sizes. For instance, the thresholds for a recall performance higher than the baseline ($\alpha = .05$, two-tailed) for the Low class is 0.54 at $T2$ and 0.58 at $T6$; this difference is due to the sample sizes. In terms of Pearson residual, the thresholds are the same, 2.8, meaning that the better than criterion performances must be 2.8 standard deviations higher than that of the baseline.

Fig. 4 reports the Pearson residuals at $T2, T4$, and $T6$ for the conditions yielding the highest accuracy figures in Tables 3 and 4: $(*, ST, AG)$, $(*, ST + AG2 + AR2, AG)$, $(*, ST + AR1 + AR2, AG)$, $(*, ST + AG2 + AR1 + AR2, AG)$, and $(*, AR2, No\_Feat)$. The trend is very similar for all conditions where "Others" $= AG$. Almost all the values for the High class (10 out of 12) are significantly better than the criterion, with the exception of $(T2, ST, AG)$ and $(T6, ST, AG)$. The number of significant conditions is lower for the Low class (eight cases out of 12). At $T4$, the Pearson residuals for both High and Low are always better than the criterion.

The picture is quite different for those conditions where social context is not considered, such as $(*, AR2, No\_Feat)$: Using only the $AR2$ feature produces a much higher Pearson residual for introverts than for extraverts; the disadvantage reduces with longer slices, but remains significant even at $T6$. Moreover, at $T2$, the standardized recall figure is strongly negative for High, implying that at this slice length, $AR2$ performs much worse than the baseline for slices corresponding to extraverts. These data are consistent with the inverse relationship between Extraversion score and $AR2$ as obtained from the manually annotated features.

In conclusion, conditions of type $(*, ST + *, AG)$ yield performances that are well balanced between the High and the Low classes, meaning that the features used are equally good for identifying extroverts and introverts. Conditions like $(*, AR2, No\_Feat)$ in turn perform very well with introverts, while resulting in bad recall for extraverts.

2. Assuming that the hits are generated by a Binomial process, we compute Pearson residual as $\frac{(c_o - c_e)}{\sqrt{mp(1-p)}}$ where c_o and c_e are the observed and expected number of hits, m is the marginal (i.e., number of extraverts in the example), and $p = 0.5$ is the "hit" probability for the baseline.

7 Discussion

The results discussed in the previous section suggest that the amount of attention received from the rest of the group while the target is not speaking (AR2) is alone sufficient to achieve significant classification performance, although this feature produces very good results with introverts while not being indicative of extraverts. Quite generally, none of the target's behaviors or combinations thereof have ever yielded statistically significant performance unless others' attentional behavior is also considered, echoing findings in [23]. In particular, the attention behavior of the target never provides statistically significant results when employed alone, as in $(*, AG1, No\_Feat)$ or $(*, AG1, AG)$. Good performance is attained only in conditions such as $(*, ST + AG1, AG)$, $(*, ST + AG2, AG)$, but that these results are seen mainly due to the contribution of ST rather than $AG1$, $AG2$.

Also, the verbal activity (ST) of the target is ineffective when used alone, but gains a good predictive power when combined with a detailed description of the peers' attention behavior, as in $(*, ST, AG)$, or to a lesser extent, when combined with summary information regarding attention received by the target, as for $(*, ST + AR1, No\_Feat)$. Conditions like $(*, ST, AG)$ provide the additional advantage of a substantially balanced distribution of the performances between extraverts and introverts. Taken together, these observations suggest that the manner in which attention patterns of peers are captured by "Target" $= AR2$ and by "Others" $= AG$ are not equivalent for classification purposes. A further piece of evidence to this effect is that while $AR2$ can be effective both alone and in conjunction with ST, fine-grained attention patterns encoded by "Others" $= AG$ need "Target" $= ST$.[3]

In view of these differences between "Others" $= AG$ and "Target" $= AR2$, one might suggest that: 1) The target's speaking time not only affects the social attention he/she receives, but extends its influence to the manner in which *each peer* distributes his/her social attention; 2) this extended influence is responsible for the better performances of conditions such as $(*, ST, AG)$ compared to $(*, ST + AR1, No\_Feat)$; 3) however, "Others" $= AG$ produces significant performances only with "Target" $= ST$ and has a less-specific relationship with Extraversion than "Target" $= AR2$. There are, of course, alternative explanations, e.g., the differences are simply an artifact of the way in which features $AR1$ and $AR2$ are computed (as the amount of attention *all* peers devote to the target). Further work is needed to investigate the precise effects of these two scenarios.

The study of the dependence of classification accuracy on slice length provides new insights to ongoing discussions on first impression formation: increasing slice length improves classification performance up to 4 minute-long slices. Then onward, accuracy either declines for conditions of type $(*, *, AG)$, or remains stable, as for conditions $(*, *, No\_Feat)$. The convergence of the classification performance of manually and automatically annotated data in conditions of type $(*, *, AG)$ increases our confidence in the study's outcomes and demonstrates that the automatic feature extraction methods have good robustness. However, the inability to replicate performance for conditions of type $(*, AR2 + *, No\_Feat)$ indicates the instability of such features when computed automatically.

Another control, conducted only on manually annotated data, was aimed at checking for the possibility that group membership affects accuracy. The absence of a significant between group variance indicates that accuracy variability is entirely due to differences among subjects, as expected, and not from grouping or procedural artifacts. Testing for significant between group variance components is an expedient that we recommend be adopted in any study addressing characteristics of individuals nested within groups.

Our study also suffers from a number of shortcomings which need to be acknowledged. The first possible problem is the sample size. We have considered 10 meetings comprising a total of 40 subjects for analysis. Depending on the length of the considered thin slice, the sample sizes varied from 2,316 at $T1$ to 356 at $T6$. It is quite possible that the limited size, especially with longer slices, might have limited the statistical power of our analyses, preventing us from detecting subtler differences—e.g., main effects of the "Target" factors in Table 4; in this respect, our results should be regarded as conservative.

Another possible limitation concerns the representativeness of the behavioral samples we considered. Although related to the above discussion, this problem is more substantial and can be readily illustrated by comparing the results reported in this study to those discussed in [38]. In [38], a subset (four meetings, 16 subjects) of the data used here was employed for Extraversion classification, using similar feature sets and experimental design. The results were remarkably different, both in terms of the accuracy scores attained (generally higher) and of the effectiveness of the exploited features. Apart from the fact that only 16 subjects were considered in [38], the behavior of those subjects did not vary enough to prevent overestimation. It is therefore important that future studies exploit a greater number of subjects, longer meeting times, and/or a larger sampling of situations.

8 Conclusion

Recent work in socio-psychology has emphasized the importance of social attention for Extraversion; Ashton et al. [13] argued that the core of Extraversion is in the tendency to behave in a way so as to engage, attract, and enjoy social attention. According to this view, extraverts invest time and energy in activities that attract others' attention. We have investigated the applicability of this social attention view for automatic prediction of Extraversion in small group meetings. We considered two kinds of behaviors capable of attracting others' attention: speaking time and the amount of social attention devoted to others. The latter was further distinguished into the attention devoted while speaking and while silent, based on the hypothesis that extraverts differ from introverts based on how their social attention behavior is contingent on their speaking status.

3. Specific tests, not discussed in this paper, show that the accuracy of conditions of type $(*, No\_Feat, AG)$ never reached statistical significance.

Together with the target's behaviors, we have also considered a number of cues concerning his/her peers (i.e., the target's social context), including: 1) the amounts of attention the target receives while he/she is speaking and is silent, 2) a detailed description of how each of the peers distributes his/her social attention among the group members, 3) information about the speaking behavior of each peer. We modeled the problem as a classification task over thin slices of target's and peers' behaviors and have systematically investigated the effects of slice length on classification accuracy. The main findings of this paper are:

- The verbal activity (ST) of the target is ineffective when considered alone, but gains good predictive power when combined with (global or detailed) descriptions of the peers' attention behavior.
- The attention behavior enacted by the target is substantially ineffective, be it alone or in combination with other features.
- The attention received from the rest of the group while silent ($AR2$) is sufficient by itself to bring about statistically significant performances.
- Among the various conditions concerning the social context, only those informing about the others' social attention behavior ("Target" =$AR1$, "Target" =$AR2$, and "Others" =AG) are effective.

All these findings are compatible with the social attention theory and support the idea that this view can be fruitfully exploited for the automatic prediction of Extraversion. Our study has paid much attention to securing the reliability of the obtained results. Comparisons between results obtained from manually and automatically extracted features demonstrate the approach's stability. The absence of significant between group variance shows that the hierarchical nature of the data does not produce undesired effects, due either to group membership or to artifacts arising from the validation procedure. Although this study suffers from a number of limitations discussed in the previous section, we hope that it has indicated interesting avenues for future research in the automatic detection of personality traits.

The accuracy levels attained, though interesting, are far from being exceptionally high: The maximum value, obtained on manually annotated data with condition ($T4, ST + AG2 + AR1 + AR2, AG$), is 0.69. Importantly, related works have reported similar accuracies. Mairesse et al. [39], who exploited several machine learning methods and acoustic features on a corpus collected by means of the EAR device [40], reported accuracies ranging from 0.50 to 0.68. Pianesi et al. [23] reported much higher accuracies (up to 0.97) on a three class (Low, Medium, and High) classification task for Extraversion using a variety of acoustic features and SVM. Their baseline, however, was much higher than ours (0.67); moreover, their usage of a leave-one-subject-out procedure for validation might have produced overfitting. Kalimeri et al. [25] reported a value of 0.46, for a three-class classification problem with Bayesian networks, employing acoustic features as in [23] and a leave-one-meeting-out validation procedure. In the end, the accuracy results obtained in this work are in line with those discussed in other works exploiting audio-visual features from meetings.

It is important to notice that in several respects the results from automatic methods as well as their variability are comparable to those obtained from thin-slice-based human judgments. For instance, Dabbs et al. [41] reported an average correlation of 0.185 between personality judgments based on thin slices and NEO [42] self-rating. Borkenau et al.'s [43] correlations ranged between 0.17 (Agreeableness) and 0.39 (Openness) for an experimenter's personality ratings against NEO self-ratings, and between 0.15 (Neuroticism) and 0.31 (Extraversion) for a confederate's ratings versus self-ratings. McCrae [42] reported a 0.50 level of consensus with thin-slice judgments and the description of close acquaintances of the targets (e.g., spouses). In other words, when other people's judgments on standard personality scales are compared to target's self-assessment, correlation rarely exceeded 0.40, and only with close acquaintance's descriptions were values slightly higher.

Although accuracy on a classification task is not directly comparable to correlations between ratings on continuous personality scales, it seems fair to conclude that an accuracy of 0.69 has similar theoretical and practical import as an agreement correlation of 0.38. In both cases, we have autonomous systems—humans and machines—inferring information about people's personality from short behavioral sequences to use as a folk-psychological construct in the prediction and explanation of people's behavior. That information is far from being always correct and, for its deployment, humans continuously confront (and possibly correct) it with situational knowledge.

While the above considerations do not diminish the importance of pursuing higher accuracy levels by exploiting larger samples, different assortments of features, better classification algorithms, etc., they contribute to framing the problem of personality prediction within actual usage scenarios. We conclude by discussing the possibility that there might be limits to the accuracy attainable for Extraversion (and, perhaps, any personality trait) classification that are intrinsic to the manner in which the problem has been defined and addressed.

The common procedure of all works on automatic personality prediction so far has been to take excerpts of a person's behavior and then provide the machine-equivalent of judgments about his/her personality. Though varying on the distal cues exploited, all relevant works, including the present one, have been more or less consistent with a Brunswik lens model for personality attribution as developed by Scherer [44]. The major modification was that no role was given to intermediate entities as proximal percepts, but a direct path was pursued to go from distal cues to attributions, usually by means of statistical modeling. The thin slices approach adds to the picture the idea that personality modeling can be pursued even by resorting to short behavioral sequences.

A possible problem with this approach is within individual behavioral variability. People do not always behave the same way: An extravert might, on occasions, be, e.g., less talkative or attempt less to attract social attention. One might suspect circumstantial effects to be at play here and endorse Fleeson's [45] position that people routinely express all levels of a given trait depending on situational characteristics. The tension between the invariance of personality traits and behavioral variability in concrete situations has sometimes been resolved by considering the latter as noise that has to be canceled out by, e.g., employing larger behavioral samples.

Although larger samples may produce better results, it can be argued that within individual variability is not just noise to be canceled out, but, stemming from the interaction between enduring traits and variable situational properties, it can give a valuable contribution to personality classification [45]. In other words, it can be argued that 1) the neglect of the informative value of within individual behavioral variation is probably one important factor limiting the performances obtained from current automated approaches and 2) an alternative approach should be investigated that exploits the interplay between personal and situational characteristics in a profitable way.

For instance, one might follow [45], and target *personality states*, that is, specific behavioral episodes wherein a person behaves, e.g., more or less introvertly/extravertly, depending on situational characteristics. The distribution of personality states conditioned on situation characteristics can then be used to reconstruct the relevant trait, reconciling the traditional focus on between person variability with the meaningfulness of within individual variability. Such an approach would not only circumvent the difficulties discussed above and possibly yield higher recognition performances, it would also advance the ultimate goal of endowing machines with the capability of explaining and predicting people's behavior by adopting a dynamic framework.

REFERENCES

[1] P. Costa and R. McCrae, "Four Ways Why Five Factors Are Basic," *Personality and Individual Differences,* vol. 13, pp. 653-665, 1992.

[2] G. Matthews, "Extraversion," *Handbook of Human Performance,* vol. 3, pp. 95-126, Academic Press, 1992.

[3] L. Eaton and D. Funder, "The Creation and Consequences of the Social World: An International Analysis of Extraversion," *European J. Personality,* vol. 17, pp. 375-395, 2003.

[4] B. Roberts, N. Kuncel, R. Shiner, A. Caspi, and L. Goldberg, "The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status and Cognitive Ability for Predicting Important Life Outcome," *Perspective Psychological Science,* vol. 2, pp. 313-345, 2007.

[5] T. Trull and K. Sher, "Relationship between the Five-Factor Model of Personality and Axis I Disorders in a Nonclinical Sample," *J. Abnormal Psychology,* vol. 103, pp. 350-360, 1994.

[6] D. Goren-Bar, I. Graziola, F. Pianesi, and M. Zancanaro, "Influence of Personality Factors on Visitors' Attitudes towards Adaptivity Dimensions for Mobile Museum Guides," *User Modelling User-Adaptation Interaction,* vol. 16, no. 1, pp. 31-62, 2006.

[7] J. Sigurdsson, "Computer Experience, Attitudes toward Computers and Personality Characteristics in Psychology Undergraduates," *Personality and Individual Differences,* vol. 12, no. 6, pp. 617-624, 1991.

[8] B. Reeves and C. Nass, *The Media Equation.* Univ. of Chicago Press, 1996.

[9] J. Cassell and T. Bickmore, "Negotiated Collusion: Modeling Social Language and Its Relationship Effects in Intelligent Agents," *User Modelling User-Adaptation Interaction,* vol. 13, pp. 89-132, 2003.

[10] M. Komarraju and S. Karau, "The Relationship between the Big Five Personality Traits and Academic Motivation," *Personality and Individual Differences,* vol. 39, pp. 557-567, 2005.

[11] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal Support to Group Dynamics," *Personal and Ubiquitous Computing,* vol. 12, no. 2, pp. 185-195, 2008.

[12] K. Andrews, "It's in Your Nature: A Pluralistic Folk Psychology," *Synthese,* vol. 165, no. 1, pp. 13-29, 2008.

[13] M. Ashton, K. Lee, and S. Paunonen, "What Is the Central Feature of Extraversion? Social Attention versus Reward Sensitivity," *J. Personality and Social Psychology,* vol. 83, pp. 245-252, 2002.

[14] Y. Iizuka, "Extraversion, Introversion and Visual Interaction," *Perceptual and Motor Skills,* vol. 74, pp. 43-50, 1992.

[15] M. Argyle, *The Social Psychology of Everyday Life.* Routledge, 1999.

[16] D. Rutter, I. Morley, and J. Graham, "Visual Interaction in a Group of Introverts and Extraverts," *European J. Social Psychology,* vol. 2, pp. 371-384, 1972.

[17] N. Ambady and R. Rosenthal, "Thin Slices of Expressive Behaviors as Predictors of Interpersonal Consequences: A Meta Analysis," *Psychological Bull.,* vol. 111, pp. 156-274, 1992.

[18] M. Blackman and D. Funder, "The Effect of Information on Consensus and Accuracy in Personality Judgment," *J. Experimental Social Psychology,* vol. 34, pp. 164-181, 1998.

[19] D. Carney, C. Colvin, and J. Hall, "A Thin Slice Perspective on the Accuracy of First Impressions," *J. Research in Personality,* vol. 41, pp. 1054-1072, 2007.

[20] S. Argamon, S. Dhawle, M. Koppel, and J. Pennbaker, "Lexical Predictors of Personality Type," *Proc. Ann. Meeting of the Interface and the Classification Soc. of North Am.,* 2005.

[21] F. Mairesse and M. Walker, "Automatic Recognition of Personality in Conversation," *Proc. Human Language Technology Conf. NAACL,* 2006.

[22] D. Olguin, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior," *IEEE Trans. Systems, Man, and Cybernetics B—Cybernetics,* vol. 39, no. 1, pp. 43-55, Feb. 2009.

[23] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal Recognition of Personality Traits in Social Interactions," *Proc. Int'l Conf. Multimodal Interfaces,* 2008.

[24] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling Personality of Participants during Group Interaction," *Proc. Int'l Conf. User Modeling, Adaptation, and Personalization: Formerly UM and AH,* 2009.

[25] K. Kalimeri, B. Lepri, and F. Pianesi, "Causal Modelling of Personality Traits: Extraversion and Locus of Control," *Proc. Int'l Workshop Social Signal Processing,* 2010.

[26] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions," *Proc. Int'l Workshop Social Signal Processing,* 2010.

[27] J. Hall and W. Watson, "The Effects of a Normative Intervention on Group Decision-Making Performance," *Human Relations,* vol. 23, no. 4, pp. 299-370, 1970.

[28] M. Perugini and L.D. Blas, "Analyzing Personality-Related Adjectives from an Eticemic Perspective: The Big Five Marker Scale (BFMS) and the Italian ab5c Taxonomy," *Big Five Assessment,* pp. 281-304, Hogrefe and Hufe, 2002.

[29] J. Ramirez, J. Segura, A. Benitez, A. de la Torre, and A. Rubio, "Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information," *Speech Comm.,* vol. 42, nos. 3/4, pp. 271-287, 2004.

[30] R. Valenti, N. Sebe, and T. Gevers, "Combining Head Pose and Eye Location Information for Gaze Estimation," *IEEE Trans. Image Processing,* vol. 21, no. 2, pp. 802-815, Feb. 2012.

[31] J. Xiao, T. Kanade, and J. Cohn, "Robust Full Motion Recovery of Head by Dynamic Templates and Re-Registration Techniques," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2002.

[32] R. Valenti and T. Gevers, "Accurate Eye Center Location and Tracking Using Isophote Curvature," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[33] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues," *IEEE Trans. Neural Networking,* vol. 13, no. 4, pp. 928-938, July 2002.

[34] R.V. Exline, S.L. Ellyson, and B. Long, "Visual Behavior as an Aspect of Power Role Relationship," *Advances in the Study of Comm. and Affect,* vol. 2, pp. 21-52, Plenum Press, 1975.

[35] A. Agresti, *Categorical Data Analysis.* Wiley, 2002.

[36] J. Hardin and J. Hilbe, *Generalized Estimating Equations.* Chapman & Hall, 2001.

[37] H. Goldstein, *Multilevel Statistical Modeling.* Wiley, http://www.bristol.ac.uk/cmm/team/hg/multbook1995.pdf, 1995.

[38] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Employing Social Gaze and Speaking Activity for Automatic Determination of the Extraversion Trait," *Proc. Int'l Conf. Multimodal Interfaces and Workshop Machine Learning for Multimodal Interaction,* 2010.

[39] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artificial Intelligence Research,* vol. 30, pp. 457-500, 2007.

[40] M. Mehl, S. Gosling, and J. Pennbaker, "Personality in Its Natural Habitat Manifestations and Implicit Folk Theories of Personality in Daily Life," *J. Personality and Social Psychology*, vol. 90, pp. 862-877, 2006.
[41] J. Dabbs, F. Bernieri, R. Strong, R. Campo, and R. Milun, "Going on Stage: Testosterone in Greetings and Meetings," *J. Research in Personality*, vol. 35, pp. 27-40, 2001.
[42] R. McCrae, "Consensual Validation of Personality Traits: Evidence from Self-Reports and Ratings," *J. Personality and Social Psychology*, vol. 43, pp. 293-303, 1982.
[43] P. Borkenau, N. Maurer, R. Riemann, F. Spinath, and A. Angleitner, "Thin Slices of Behavior as Cues of Personality and Intelligence," *J. Personality and Social Psychology*, vol. 86, pp. 599-614, 2004.
[44] K. Scherer, "Personality Markers in Speech," *Social Markers in Speech*, pp. 147-209, Cambridge Univ. Press, 1979.
[45] W. Fleeson, "Towards a Structure- and Process-Integrated View of Personality: Traits as Density Distributions of States," *J. Personality and Social Psychology*, vol. 80, pp. 1011-1027, 2001.

Bruno Lepri received the PhD degree from the Department of Information Engineering and Computer Science, Trento University, in 2009. He is currently a Marie Curie postdoctoral fellow at MIT's Human Dynamics Laboratory, Cambridge, and at the FBK Research Centre, Trento, Italy. His research interests are human-behavior understanding, social signal processing, computational social science, social network analysis, and multimodal interaction.

Ramanathan Subramanian received the PhD degree in electrical and computer engineering from the National University of Singapore in 2008. He is currently a postdoctoral researcher in the Department of Information Engineering and Computer Science, University of Trento, Italy. His research interests span human-centered computing, human behavior understanding, computer vision, and multimedia processing, especially studying and modeling various aspects related to human visual and emotional perception. He is a member of the IEEE.

Kyriaki Kalimeri is currently working toward the PhD degree at the University of Trento. She is a visiting PhD student at the MIT Media Lab. Her current research interests are in the computational social science domain, focusing on the automatic analysis of personality integrating methods both from multimedia signal processing, machine learning, and social sciences. She is a student member of the IEEE.

Jacopo Staiano is currently working toward the PhD degree from the Department of Information Engineering and Computer Science, University of Trento, Italy. His recent works include automated multimodal analysis of human behavior, social signal processing, automatic personality recognition, and social network analysis.

Fabio Pianesi is the vice director for Research of Trento RISE, a joint partnership between FBK and the University of Trento, and colocation manager of the Trento Node of EIT ICT Labs. In the past, he served as head of the Cognitive and Communication Technologies Division of ITC-irst and as head of the Computational Cognition Laboratory of FBK. He has been very active in many of the disciplines that contribute to human computing, including computational and formal linguistics, human-computer interaction, multimodal interaction, social signal processing, and human-behavior understanding. He is the chair of the Advisory Board of the ACM International Conference on Multimodal Interaction (ICMI). He is a member of the IEEE.

Nicu Sebe is with the Faculty of Cognitive Sciences, University of Trento, Italy, where he is leading research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He was involved in the organization of major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which he served as a general cochair of the IEEE Automatic Face and Gesture Recognition Conference (FG '08), ACM International Conference on Image and Video Retrieval (CIVR '07 and '10), and WIAMIS '09 and as one of the initiators and a program cochair of the Human-Centered Multimedia track of the ACM Multimedia '07 conference. He is the general chair of ACM Multimedia '13 and was a program chair of ACM Multimedia '11. He is a senior member of the IEEE.