Annotation and Recognition of Personality Traits in Spoken Conversations from the AMI Meetings Corpus

Fabio Valente, Samuel Kim, Petr Motlicek

Idiap Research Institute, CH-1920 Martigny, Switzerland
{*fabio.valente,samuel.kim,petr.motlicek*}*@idiap.ch*

Abstract

Recognition of personality traits is a well studied problem in psychology while only recently it has been addressed by speech and language technology research. This paper describes annotation and experiments towards automatically inferring speakers personality traits in spontaneous conversations. In the first part, the work describes the annotation framework based on the Big-Five personality traits model (Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness) applied to 128 speakers from the AMI corpus. As the corpus contains rich annotations, those data can generalize previous studies based on enacted speech or dialogues. In the second part, the paper describes experiments based on various features including prosody, words n-gram, dialog acts and speech activity. Results reveal that high/low extraversion, consciousness and neuroticism traits can be automatically recognized with accuracy rate of 74.5%, 67.6% and 68.7%, respectively, while agreeableness and openness classification error rates are not statistically better than chance. Non-linguistic features (prosody, speech activity, overlaps and interruptions) outperform linguistic features (words n-gram and dialog acts) in this setup.

1. Introduction

Studying personality and its manifestation has been a very active research field in social psychology for long time however only recently the topic has gained attention in the speech research community. In contrary to other traits like emotions, charisma or mood, personality is considered as a longer term and more stable aspect of individuals [1]. Personality traits have shown to correlate with leadership, job performance, effectiveness in accomplishing tasks and on how people interacts with machine and interfaces (see [2] for a review). Furthermore studies like [3] suggest that matching users' personality increases the effectiveness of many human-machine interface technologies. For those reasons, several recent works have addressed the problem of automatically recognizing personality traits from speech features.

Markers of personality traits, especially the Extraversion trait, have been found in speech features and speaking style [4], lexical categories and choice of vocabulary [5], words n-grams [6], speech acts [7] and part of speech [8]. Most of those studies are based on text data (email, blogs and other written texts) or short spoken utterances/conversations.

On the other hand, works towards automatic personality trait recognition from speech has started much recently, focusing mainly on para-linguistic cues. Data includes corpora of enacted speech (see [9, 10]) where actors produce short utterances simulating a given personality trait reading a paragraph of text, or corpora of short utterances from broadcast recordings [11].

Recognizing speakers personality traits in natural conversations add several challenges compared to acted or zero-aquintance scenarios, due to language-specific and culture specific cues. For instance in [12], authors investigated this problem in spontaneous dialogs from tourist call-center calls involving a user and an agent showing promising rates for Extraversion and Conscientiousness. Whenever naturalistic spoken conversations are considered, other difficulties come from obtaining reliable precise annotations for the various features and phenomena happening during the conversations (disfluencies, overlap, turn-taking).

In this work, we introduce personality traits annotations and initial results on their automatic recognition in the AMI meeting corpus [13]. The rationale for performing personality studies on those data is, from one side, the large amount of spontaneous multi-party speech available (100 hours) from over 120 speakers recorded both with microphone and cameras. On the other hand, meetings are composed of short presentations, dialogues, multi-party discussions and represents a rich setting for generalizing previous works in controlled scenarios [9, 10, 11] or two-side dialogues [12]. Furthermore the availability of very precise and rich annotations for speaking time, words, dialog acts and topics allows to study several phenomena that happens during conversations.

The remainder of the paper is organized as follows: section 2 briefly describes the personality trait theory used in this work, section 3 describes the dataset, the annotation process and the statistics, section 4 describes the experimental setup and results analysis and the paper is then concluded in section 5.

2. Big-Five Personality Traits

Several theories and definitions of personalities have been proposed in psychology. Among those different paradigms, in this work we adopt the "Personality traits scheme" [14] which defines personality as patterns of behavior that manifests itself in terms of measurable traits and in particular the Big Five model which describes human personality as a vector of five values corresponding to five bipolar traits. This theory has been empirically validated and already formed the basis of a number of studies on automatic personality recognition [2, 9, 10, 11, 12] based on text and speech as well as synthesis of personality traits [15]. The five traits can be summarized as follows:

- Extroversion : high extraversion is characterized by positive emotions, surgency, the tendency to seek out stimulation while low extraversion is characterized by lack the social exuberance and activity levels of extroverts.
- Agreeableness: high agreeableness is related to the tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others as in case of low agreeableness.

- Conscientiousness: high conscientiousness is a tendency to show self-discipline, act dutifully, and aim for achievement while low conscientiousness is the tendency to be careless and indifferent.
- Neuroticism: high neuroticism is the tendency to experience negative emotions, such as anger, anxiety, or depression. It is sometimes called emotional instability, or is reversed and referred to as emotional stability.
- Openness: high openness indicates general appreciation for art, emotion, adventure, unusual ideas, imagination and curiosity while low openness characterize conservative people.

Several questionnaire exist for estimating the Big Five traits where raters have to answer a number of questions in a 5-point Likert scale ranging from 'strongly agree' to 'strongly disagree'. This work makes use of the 10-item short version of the Big Five Inventory (BFI) questionnaire in English language [16] also known as BFI-10. While highly correlated with the original full version of the BFI, the BFI-10 allows to assess personality in a shorter time. Two types of personality assessment are generally studied, the first based on self-assessment from the person under study while the second is based on assessment obtained by external raters. In this study, the second approach is used.

3. Dataset and Annotation

The AMI Meeting Corpus is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The video comprises both overview and closeup cameras. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team composed of *Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID)* tasked with designing a new remote control. The meeting is supervised by the Project Manager who follows an agenda with a number of items to be discussed with other speakers. The corpus is manually transcribed at different levels (roles, speaking time, words, dialog act).

In order to annotate how participants' personalities were perceived, the following annotation procedure has been implemented. A subset of 32 meetings (in the corpus notation, they are designated with the letter *d*) containing 128 different speakers (84 male and 44 female participants) is selected from the entire corpus. The age of participants is between 20 and 60 years with a median value of 26 years. Each meeting has been segmented into short clips based on the presence of long-pauses, i.e., pauses longer than one second. Segments are later subsampled in order to cover uniformly the entire recoding thus comprising various part of the meeting including the opening, the presentations, the discussions and the conclusion. In order to provide annotators both with video and audio information, a sequence of video clips is created merging the closeup cameras and the overview camera (depicted in Figure 1) together with the audio from the individual headset microphones that each speaker wears. The total amount of audio/video selected for each meeting is approximatively *12 minutes* covering both monologues, dialogues and multi-party discussions. Each annotator is requested to watch and listen entirely to all the 12 minutes before answering to the BFI-10 questionnaire for each of the meeting participants. During the annotation process, a number of simple control questions based on the speech activity of each speaker in the clips are used to identify and reject annotators that do not carefully listen while performing the annotation.

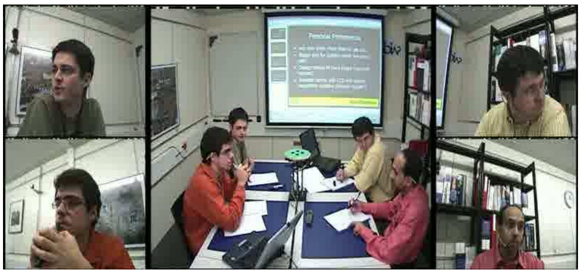

Figure 1: Screen-shot of the clip annotators are provided with. It includes both closeup cameras for each of the four participants and the overview camera. The audio track is obtained by merging the audio from the individual headset microphones.

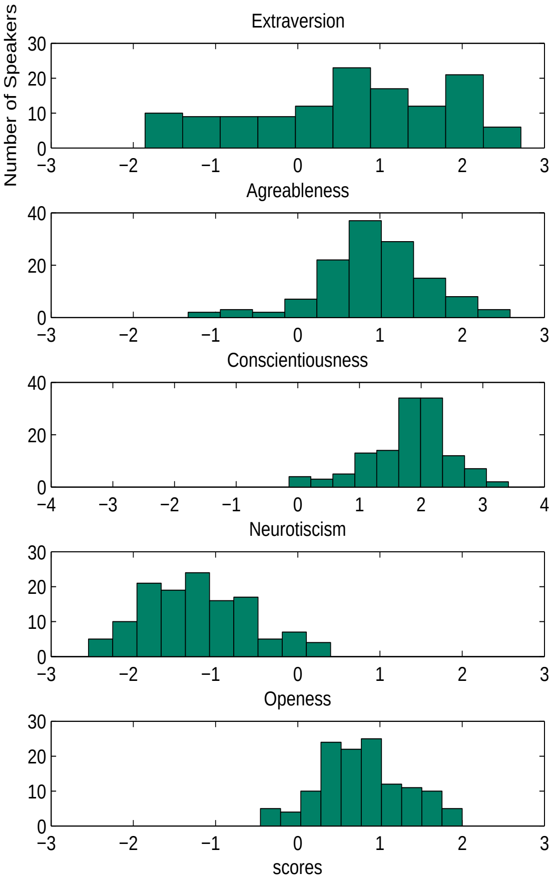

Figure 2: Personality trait score histogram for the 128 speakers (84 male and 44 female participants) in the AMI scenario meetings.

At the end of the process, each meeting (thus the personality of each participant) is assessed by 11 different annotators.

The total scores for the personality traits are then obtained by averaging the scores from the 11 assessors providing a vector of five continuously distributed values in the range [-4,4]. The histograms of those scores are depicted in Figure 2. It can be noticed that extroversion scores range from positive to negative having a quite flat distribution while scores for agreeableness, conscientiousness and openness have positive median value and neuroticism scores have a negative median value.

4. Setup and Experiments

The rich annotations of the AMI corpus allow us to study the automatic recognition of those traits using para-linguistic information as in [9, 10, 11, 12] as well as linguistic features (words, dialog acts) as in [2]. Furthermore, given the multi-party nature of the meetings, several other features like overlaps, floor grasping and conversational behavior can be included in the investigation. The following experiments report initial results based on those features from the corpus. Similarly to what proposed in [12], the five personality scores are converted in two labels: *L* for low scores and *H* for high scores. The splitting is done according to the median value of the scores so that a random classifier has 50% chance of correctly labeling the trait. In order to combine this information in a simple discriminative fashion, this work makes use of boosting algorithms. The principle of boosting is to combine many weak learning algorithms to produce a single accurate classifier. The algorithm generates weak classification rules by calling the weak learners repeatedly in series of rounds. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The version of Boosting algorithm used was multi-class Boosting defined in [17] and implemented using Boostexter. The weak learners are one-level decision trees. This algorithm provides a very simple and effective way to combine continuous features as well as discrete features.

The tests are run in a meeting and speaker independent fashion by using leave-one-out approach where 31 meetings out of the 32 are used for training the boosting and the test is done on the remaining one. It is important to notice that, in this way, none of the speakers in the testing has been seen during the training. In first place, boosting is applied on features obtained from speech from each participant, discarding information on overlaps and interruptions which is included later.

4.1. Single speaker features

Audio manually segmented from the Independent Headset Microphones (IHM) is force-aligned for obtaining precise speech/non-speech segmentation. This segmentation is used to extract a sequence of speaker turns defined as in [18], i.e., speech regions from a single speaker uninterrupted by pauses longer than 300 ms. After this processing, based on the force-aligned segmentation, the following features are extracted for each of the four speakers that take part in the meeting.

1 Speech Activity Features: the total and relative amount of speech time per speaker, statistics on turns and sentences (total number, average duration, maximum duration) as well as the average duration of pauses per speaker. Those values are included in the booster as continuous features.

2 Prosodic Features: the fundamental frequency (F0) is computed from the headset microphones using 30ms long windows shifted by 10ms. After that, speaker statistics like f0 mean, maximum, minimum, median and the standard deviation are computed. The same statistics are also computed on the intensity. Furthermore, average speech rate per speaker is included. Those features undergo a gender-depended Z-norm and are then used in the booster as continuous features.

3 N-gram of words have already been successfully applied in recognized personality from text [2] thus word trigrams are used in the booster.

4 Dialog Act Tags: Dialog Acts (DA) aim at capturing the speaker's intention in the discussion. AMI corpus is annotated in terms of 14 broad DA classes that includes statements, questions and back-channels. DA tags have been shown to correlate with personality traits (especially extraversion) [7], thus the per-speaker DA counts (14 per each speaker) are included in the booster.

As auxiliary information, also gender and age of the speakers are included in the booster. Results are reported in Table 1 per trait and per feature set in terms of recognition accuracy (first four rows). The asterisk beside the accuracy designates the fact the results are not statistically significant based on a binomial test.

It can be noticed that recognition rates are statistically significant only for extraversion and conscientiousness traits. Extraversion recognition is particularly high for speech activity and dialog act features while conscientiousness recognition is higher for prosodic and speech activity features. Simple lexical features, included as word n-gram, provide overall poor performances. Furthermore it can be noticed that accuracies obtained for remaining traits (agreeableness, neuroticism and openness) are not statistically better than chance according to a binomial test with rejection of the null hypothesis at the 5%.

4.2. Feature from participants interactions

The previously described features mainly capture speaker behaviors regardless of their interactions with other participants. In order to model interactions that typically happen when negotiating the floor of the conversation also statistics from overlapping speech regions are included, i.e., the amount of time a speaker overlaps with the other meeting participants, the average overlap duration, the number of time a speaker interrupts (estimated as holding the floor after an overlap), the number of times a speaker is interrupted and the centrality of the speaker in the conversation. Those estimates are included in Booster as continuous features. Results are reported in Table 1 (fifth row) and show comparable results with other features for detecting high/low levels of extraversion and conscientiousness. It is interesting to notice that they hold much better results for detecting high/low neuroticism. As last experiments, the various features are used jointly into the booster to investigate how much improvements can be obtained by merging together different type of informations. Results are reported in Table 1 (sixth row), showing that the combination marginally improves over the best feature set.

4.3. Recognition of extreme high/low traits

Analysis of results reveals that most of the errors in recognizing the high/low traits happen mainly in correspondence of median values similarly to what reported in [12, 11, 19]. In order to investigate how well extreme cases are recognized in this dataset, median values are removed from the setup, i.e., all scores in the range $[m-0.5, m+0.5]$ are not considered where m represents the median score for a given trait. The number of speakers remaining in the setup is reported in Table 2 and corresponds to two-thirds of the total for the extraversion trait and approximatively a third of the total for the other traits. As the number of samples becomes small, a speaker based leave-one-out approach is implemented instead of the previous meeting based leave-one-out. Results are reported in Table 2 showing that, upon removal of median values, recognition rates are above 70% for all the five traits.

5. Discussion and Conclusion

Studying personality and its manifestation has been a very active research field in social psychology while only recently sev-

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Prosody	62.3%	51.5% *	67.1%	45.3% *	48.5% *
Speech Activity	73.5%	55.4% *	64.6%	52.3% *	51.5% *
Dialog Acts	69.2%	51.5% *	61.7% *	57.0%*	52.3% *
Words n-gram	58.0% *	48% *	53.9% *	49.2% *	54.3% *
Participant Interactions	68.1%	55.4% *	66.4%	68.7%	53.9% *
All Features	74.5%	55.4% *	67.6%	68.7%	57.1%*

Table 1: Accuracy for High/Low recognition of personality traits based on speech activity, prosodic, dialog acts and lexical features. The asterisk beside the accuracy designate that the result is not statistically significant according to a binomial test with rejection of the null hypothesis at the 5%.

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Number of speakers after median removal	84/128	42/128	49/128	49/128	44/128
All Features	84.5%	70.5%	80.7%	78.3%	78.3%

Table 2: Accuracy for extreme values of High/Low personality traits based on speech activity, prosodic, dialog acts and lexical features. The number of speakers is reduced to two thirds of the total for the Extraversion trait and to one-third of the total for other traits.

eral works have addressed the problem of automatic recognition of personality traits from speech features. This paper introduces personality traits annotations on the AMI meeting corpus. The choice of the corpus comes from the large amounts of spontaneous conversational data and the very rich transcriptions that have been done over years (speaking time, words, dialog-acts, topics). Those annotations can allow general studies based on para-linguistic information as in [9, 10, 11, 12], linguistic information [2] as well as other phenomena like overlap and interruptions among speakers.

Annotations are done using external raters based on the Big-Five 10 item questionnaire [16] producing personality traits scores (Extraversion, Agreeableness, Consciousness, Neuroticism, Openness) for 128 speakers. Annotators were provided with audio and video information from approximatively 12 minutes of meeting excerpts. Initial studies based on a simple boosting classifier have been carried showing that traits like extraversion, conscientiousness and neuroticism can be automatically recognized. Speech activity statistics provide the best performance for the extraversion trait, prosodic features for the conscientiousness trait and interestingly, overlapping speech statistics provide best performances in case of neuroticism. Dialog act features correlate well only with the extraversion trait. Agreeableness and openness traits are not recognized above chance levels. Those preliminary, the results confirm what observed on other natural conversations [12]. Furthermore they provide some novel insight like the possibility of recognizing neuroticism from conversational behavior and especially overlap and interruptions with other participants in the meeting. On the other hand, in contrary to what observed before, word n-gram do not provide satisfactory performances thus in future plan we will investigate with LWIC and MRC dictionaries as already proposed in [2]. Last but not least, also visual cues like gesture and expression will be studied in future works[1].

6. References

[1] Klaus R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[2] Franois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research, Vol*, vol. 30, pp. 457–501, 2007.

[3] Clifford Nass and Scott Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, The MIT Press, July 2005.

[4] K. R. Scherer, *Personality markers in speech.*, Cambridge University Press, 1979.

[5] Pennebaker JW and King LA., "Linguistic styles: language use as an individual difference.," *Journal of Personality and Social Psychology*, vol. 77(6), 1999.

[6] J. Oberlander and A.J Gill, "Language with character: A stratified corpus comparison of individual differences in e-mail communication," *Discourse Processes*, vol. 42, 2006.

[7] Thorne Avril, "The press of personality, a study of conversations between introverts and extroverts," *Journal of Personality and Social Psychology*, vol. 53(4), 1987.

[8] Jon Oberlander and Scott Nowson, "Whose thumb is it anyway?: classifying author personality from weblog text," in *Proceedings of the COLING/ACL on Main conference poster sessions*, Stroudsburg, PA, USA, 2006, COLING-ACL '06, pp. 627–634, Association for Computational Linguistics.

[9] Tim Polzehl, Sebastian Moller, and Florian Metze, "Automatically assessing personality from speech," in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, Washington, DC, USA, 2010, ICSC '10, pp. 134–140, IEEE Computer Society.

[10] Tim Polzehl, Sebastian Moller, and Florian Metze, "Automatically assessing acoustic manifestations of personality in speech," in *Proceedings of the 2010 IEEE Spoken Language Technology Workshop*, 2010.

[11] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of ACM Multimedia Workshop on Social Signal Processing*, 0 2010.

[12] A.V. Ivanov, G. Riccardi, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Proceedings of Interspeech*, 2011.

[13] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[14] Jerry S. Wiggins, *The Five-Factor Model of Personality - Theorical Perspectives*, The Guilford Press, 1996.

[15] Francois Mairesse and Marilyn Anne Walker, "Towards personality-based user adaptation: Psychologically informed stylistic language generation," *User Modeling and User-Adapted Interaction*, vol. 20, 2010 2010.

[16] Rammstedt B. and John O., "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of Research in Personality*, vol. 41(1), 2007.

[17] Schapire R. and Singer Y., " BoosTexter: A boosting-based system for text categorization," *Machine Learning*, 2000.

[18] Shriberg E. et al., "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *in Proceedings of Eurospeech 2001*, 2001, pp. 1359–1362.

[19] Black Matthew and al., "Automatic classification of married couples' behavior using audio features.," in *INTERSPEECH*, Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, Eds., 2010.

[1] This work was supported by the Swiss National Science Foundation under the NCCR IM2 grant and the EU NoE SSPNet