Acoustic Features in Dialogue Dominate Accurate Personality Trait Classification

Maria Koutsombogera, Parth Sarthy and Carl Vogel
School of Computer Science and Statistics
Trinity College, The University of Dublin
Dublin 2, Ireland
{koutsomm,sarthyp,vogel}@tcd.ie

***Abstract*—We report on experiments in identifying personality traits from the dialogue of participants in the MULTISIMO corpus. Experiments used audio and linguistic features from participants' speech and transcripts, using both self- and observer personality reports. Contrary to our expectations that the linguistic content would best predict traits, the results highlight the multimodal nature of personality computing, suggesting that the content is less important than acoustics: except for two cases, models based on acoustic features only, or combined with linguistic features, outperform models based on linguistic features alone; results also show that there is no optimal choice of a single model or feature set for the prediction of a trait across personality reports, as different models work best for different traits.**

***Index Terms*—personality computing, big five traits, self-assessment, informant-assessment**

I. Introduction

Automatic personality recognition and perception tasks are relevant to computing areas involving understanding, prediction or synthesis of human behavior [1], [2]. In daily social or workplace interactions, human behavior is driven by people's own personality and the way their personality is perceived. Automatic personality perception and synthesis are important in the HCI domain, as personality characteristics designed for and attributed to machines enable associations with the users' attitude [3]. Approaches to personality computing have been applied in a variety of settings, an extensive review of which is presented in [1]. Automatic personality recognition considers self-assessed personality scores, in combination with linguistic and non-verbal features. Automatic personality perception exploits external judgements, i.e. personality traits assessed by informants, and focuses mainly on non-verbal cues.

The Five-Factor Model has become the dominant paradigm in personality research [4]. The Big Five traits are: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (OCEAN). Self-assessments are usually captured via first-person questionnaires, while informant-assessments are attributed with third-person questionnaires. The models that most effectively predict measurable personality aspects have been proven to be those based on traits [5], and are widely used in the computing community, as they represent personality in terms of numerical values [1]. Several works have compared the accuracy and predictive validity of personality judgement, reporting that traits perceived by knowledgable others have better validity over self-judgments [6], [7] or that both types of reports have equal levels of accuracy and provide unique insight into how a person typically behaves [8].

We approach personality computing from both recognition and perception aspects. We conducted experiments to classify the OCEAN personality traits of group dialogue participants within the MULTISIMO corpus [9], using audio and linguistic features extracted from their speech and transcripts respectively, and based on two types of personality reports: self-assessments by the participants themselves, and reports provided by informants unacquainted with the participants. Recognition experiments employ the self-reports, and personality perception experiments involve the informant reports.

We explore the implemented models performance for both types of reports, and test our hypothesis that the linguistic content is a strong predictor of personality traits; this hypothesis is rooted in the assumption that people's words reflect their personality, but also in that reports are based on lexical approaches on personality-related vocabulary [10], [11]. Contrary to our hypothesis, the results suggest that, except for two cases, models based on acoustic features only, or combined with linguistic features, outperform models based on linguistic features alone. Results also show no optimal choice of a single model for trait prediction, as different models work best for different traits, and most informative features are different among self- and informant reports; thus, trait-specific modeling seems more appropriate than attempting to identify a single best model across traits and reports. This work is novel in that it reports experimental results for both recognition and perception of the OCEAN traits, i.e. using both self and informant ratings of personality observed in a new dataset of dialogues and involving two expressive modalities of the dataset subjects, i.e. their audio extracts and full transcripts.

II. Materials

A. Data Set

We used the MULTISIMO corpus [9], which consists of 23 sessions of collaborative group interactions where two players need to provide answers to a quiz and are guided by a facilitator. Out of the total 49 participants (mean age: 30, 25 female), 46 were assigned the role of players and were

The research leading to these results has received funding from the ADAPT Centre for Digital Content Technology(Grant 13/RC/2106), and the EU H2020 programme under the Marie Sklodowska-Curie grant No 701621.

randomly paired, and 3 participants shared the role of the facilitator. The sessions were carried out in English with a mean duration of 10 minutes; the players' task was to discuss, provide the 3 most popular answers to each of 3 questions and rank them from the most to the least popular.[1]

B. Audio Features

To extract acoustic features from audio we followed the thin slice approach. Thin slices of behavior are small behavior samples (i.e. varying from 1 second to several minutes) that raters exploit to infer accurate judgments about other people's states, traits and personal characteristics, and are proven to carry reliable information [12]–[15]. We randomly extracted audio clips of 4-10 seconds from the individual participants' audio files, two clips on average per speaker. The selection criteria addressed solely the quality of the audio, i.e. clips including continuous speech from a sole participant, clean of noise and overlapping talk. The temporal location of the slices within the dialogue was not considered. In total, 79 audio clips were selected, corresponding to 36 speakers.

The audio features fed to the machine learning classifiers were extracted with the openSMILE toolkit [16] and the eGeMAPS extended minimalistic acoustic parameter set configuration [17], which contains low-level descriptors and their functionals, loudness and pitch functionals, cepstral parameters and their functionals, resulting in a total of 88 parameters.

C. Linguistic Features

Audio files were transcribed by 2 annotators using `Transcriber`[2] and were cleaned of disfluencies. Linguistic Inquiry and Word Count (LIWC) [18], a psychologically oriented text analysis tool, extensively used to associate linguistic features with the OCEAN traits, was applied to the final 36 clean transcripts to analyse the participants' verbal content through 88 predefined dimensions (word categories), resulting in a set of values for each of the dimensions, per speaker.

D. Personality Traits: Self and Informants' Reports

Before the recordings participants completed the Big Five Inventory (BFI-44), a self-report inventory measuring the OCEAN traits [19], [20]. The test consists of 44 items (statements) that participants rated to indicate the extent to which they agree with those. After a list of scores per trait and per participant was created, the percentile rank of each participant per trait was calculated upon the groups population (local norms), as opposed to general population norms.

In addition to self-reports, a perception experiment was run after the recordings to collect ratings from 8 independent informants (5 female, average age: 39), who listened to the 79 audio clips (cf. § II-B) and responded to a 10-scale questionnaire (i.e. the BFI-10 questionnaire [21], an abbreviated version of the BFI-44 scale). Informants were not acquainted with the people whose audio they assessed. The mean score per speaker per trait was computed. Again, percentile ranks across the five personality traits were computed on the basis of local norms.

III. Methodology

The machine learning modelling was performed in Weka.[3] Correlations between the personality ratings (dependent variable) and the audio and linguistic feature sets (independent variables) were calculated to select the most significantly informative features. We adopted a binary approach regarding personality ratings: percentile scores for both assessments were classified in two categories, *high*, for scores equal or greater to the 50th percentile, and *low*, for scores below the 50th percentile.The models used were AdaBoost, Naive Bayes, Logistic Regression and Random Forest. 10-fold cross-validation was performed for each model. Each exploited 3 distinct feature sets: linguistic features (LIWC); audio features (eGeMAPS); and a combination of LIWC and eGeMAPS. The measures employed to assess the performance of the models are accuracy, precision, recall, F1 and AUC ROC.

IV. Results

Best performance results for the prediction of the self- and informant-assessed traits are reported in tables provided in supplemental materials.[4] The results are presented per feature set (LIWC, eGeMAPS and combination of LIWC and eGeMAPS).Below, we highlight the dominant results.[5]

For prediction of self-reported traits, the best performance results for Openness are provided by Random Forest on the eGeMAPS feature set (84.31%). AdaBoost holds the best results for both Conscientiousness and Extraversion on the eGeMAPS feature set (82.35% for both cases). Agreeableness is best predicted by AdaBoost (82.35%) on the combined feature set (LIWC& eGeMAPS). Finally, the best performance for Neuroticism is given by the logistic regression model (94.11%). The best results for all traits are obtained from models trained on the eGeMAPS feature set, with the exception of Agreeableness, where the model on the combined set has the best performance.

For prediction of informant-assessed traits, openness is best predicted with AdaBoost on the eGeMAPS set (100%). Logistic regression provides the best results for Conscientiousness, again, on the eGeMAPS set (88.63%). The best performance for Extraversion is provided by Naive Bayes on the combined set (LIWC& eGeMAPS, 79.54%). Agreeableness and Neuroticism are best predicted with the LIWC feature set (Log. Regression (79.54%) and AdaBoost (77.27%) respectively). Unlike the self-reports performance scores, here the predictive validity is shared between linguistic features (LIWC) for Agreeableness and Neuroticism, and audio features (eGeMAPS) for Openness and Conscientiousness, with Extraversion getting the best results from a combined set.

[1] All experiments were supervised by the SCSS Research Ethics Committee at TCD. With complete compliance with participants' consent, 18 dialogues are available at http://multisimo.eu/datasets.html

[2] http://trans.sourceforge.net/, last accessed 02.01.2020

[3] https://www.cs.waikato.ac.nz/ml/weka/, last accessed 02.01.2020

[4] See http://multisimo.eu/personality/personality_experiments_results.pdf.

[5] Further, the full list of significantly correlated features can be found at: http://multisimo.eu/personality/Feature_description.pdf

V. Discussion

Out of the 10 models implemented per personality trait (5 for recognition, 5 for perception), 6 of them performed best on the acoustic feature set, 2 on combined acoustic and linguistic feature sets, and 2 on the linguistic set, suggesting that, in most cases, acoustic features are powerful predictors of personality.

Results on automatic personality recognition (prediction of self-reported traits) are indicative of that the performance is significantly better with acoustic parameters alone or combined with linguistic features. Related work reports best performances related to Openness measured on linguistic, audio, and on combined linguistic and audio features, each achieved with different classifiers [22], while [23] report on encouraging results for Conscientiousness and Extraversion with openSMILE features, and [24] obtained accuracy of 45%-100% based on word n-grams and various feature combinations.

In automatic personality perception (prediction of informant-assessed traits) the linguistic parameters are equally important to the acoustic ones; contrary to what is reported in the personality perception literature, where best accuracies are achieved on acoustic features alone, linguistic features were the best predictors for perceived Agreeableness and Neuroticism. Related approaches on perception report on accuracies between 60% and 73.5% depending on the traits [14], and [25] obtained accuracy of about 60% in all traits prediction, both exploring acoustic features. A dedicated INTERSPEECH challenge on trait prediction using the standard openSMILE feature set showed that no approach clearly outperforms the others [26].

The above findings reveal the difficulty to generalise over a single best model for both recognition and perception experiments, as there is no optimal solution for a trait prediction and different models account for different traits. Also, we noted that features with strongest correlations per trait are different when comparing self- and informant reports; thus, this does not allow generalisations about most informative features within each feature set applying to both prediction and perception experiments.

References

[1] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[2] A. Esposito, A. M. Esposito, and C. Vogel, "Needs and challenges in human computer interaction for processing social emotional information," *Pattern Recognition Letters*, vol. 66, pp. 41–51, 2015.

[3] C. Nass and S. Brave, *Wired for speech : how voice activates and advances the human-computer relationship*. Cambridge, Mass.: MIT Press, 2005.

[4] R. McCrae, "The five-factor model of personality," in *The Cambridge handbook of personality psychology*, J. J. Corr and G. Matthews, Eds. New York: Cambridge University Press, 2009, pp. 149–161.

[5] I. J. Deary, "The trait approach to personality," in *The Cambridge handbook of personality psychology*, P. J. Corr and G. Matthews, Eds. New York: Cambridge University Press, 2009, pp. 89–109.

[6] D. W. Kolar, D. C. Funder, and C. R. Colvin, "Comparing the accuracy of personality judgments by the self and knowledgeable others," *Journal of Personality*, vol. 64, no. 2, pp. 311–337, 1996.

[7] S. Balsis, L. D. Cooper, and T. F. Oltmanns, "Are informant reports of personality more internally consistent than self reports of personality?" *Assessment*, vol. 22, no. 4, pp. 399–404, 2015.

[8] S. Vazire and M. Mehl, "Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior," *Journal of Personality and Social Psychology*, vol. 95, no. 5, pp. 1202–1216, 11 2008.

[9] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus," in *Proc. LREC 2018*, N. C. et al., Ed. Paris, France: ELRA, 2018, pp. 2945–2951.

[10] L. R. Goldberg, "From ace to zombie: Some explorations in the language of personality," in *Advances in Personality Assessment*, C. D. Spielberger and J. N. Butcher, Eds. Hillsdale: Erlbaum, 1982, pp. 203–234.

[11] G. Saucier and L. R. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The five-factor model of personality: Theoretical perspectives*, J. S. Wiggins, Ed. New York: Guilford Press, 1996, pp. 21–50.

[12] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," in *Advances in Experimental Social Psychology*, ser. Advances in Experimental Social Psychology. Academic Press, 2000, vol. 32, pp. 201 – 271.

[13] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *Journal of Research in Personality*, vol. 41, no. 5, pp. 1054 – 1072, 2007.

[14] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[15] L. S. Nguyen and D. Gatica-Perez, "I would hire you in a minute: Thin slices of nonverbal behavior in job interviews," in *Proc. ICMI 2015*. New York, NY, USA: ACM, 2015, pp. 51–58.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proc. 18th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[18] J. Pennebaker, R. Booth, R. Boyd, and M. Francis, "Linguistic inquiry and word count: LIWC2015," www.liwc.net/, 2015.

[19] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory versions 4a and 54," University of California, Berkeley, Institute of Personality and Social Research, Tech. Rep., 1991.

[20] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy," *Handbook of personality: Theory and research*, vol. 3, pp. 114–158, 2008.

[21] B. Rammstedt and O. P. John, "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, no. 1, pp. 203 – 212, 2007.

[22] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Int. Res.*, vol. 30, no. 1, pp. 457–500, Nov. 2007.

[23] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Proc. INTERSPEECH 2011*. ISCA, 2011, pp. 1549–1552.

[24] J. Oberlander and S. Nowson, "Whose thumb is it anyway?: Classifying author personality from weblog text," in *Proc. COLING-ACL '06*. Stroudsburg, PA, USA: ACL, 2006, pp. 627–634.

[25] T. Polzehl, S. Moller, and F. Metze, "Automatically assessing personality from speech," in *Proc. IEEE ICSC 2010*, Sep. 2010, pp. 134–140.

[26] B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, B. Felix, R. v. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," *Proc. INTERSPEECH 2012*, 2012.