On Speaker-Independent Personality Perception and Prediction from Speech

Tim Polzehl[1]*, Katrin Schoenenberg*[1]*, Sebastian Möller*[1]*, Florian Metze*[2]
Gelareh Mohammadi[3]*, and Alessandro Vinciarelli*[3,4]

[1]Quality and Usability Lab, TU-Berlin / Telekom Innovation Laboratories; Germany
[2]Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA; USA
[3]Idiap Research Institute; Martigny; Switzerland
[4]Department of Computing Science; University of Glasgow; UK

{tim.polzehl,katrin.schoenenberg,sebastian.moeller}@telekom.de
fmetze@cs.cmu.edu, gelareh.mohammadi@idiap.ch, alessandro.vinciarelli@glasgow.ac.uk

Abstract

In this paper, we present ongoing experiments and insights regarding automatic and human assessment of perceived personality. While within the INTERSPEECH Speaker Trait Challenge participants will train systems in order to recognize binary targets along the Big 5 personality trait, we will analyze and discuss properties of the data, the labeling scheme and the predictive quality. Conducting factor analyses, estimating reliability, and building regression models capturing dimensions of personality we compare all results to our own work and introduce a new extension of our personality database. In conclusion, this paper contributes in methodology and understanding on how to asses the perceived personality from an unknown speaker by humans and machines.

Index Terms: extra-linguistic speech properties, personality modeling from speech, speaker characteristics

1. Introduction

According to [1] humans assign personality rapidly and automatically. Humans presume personality from the very first perception of "the other", which takes mostly only few seconds but guides our behavior and attitudes towards that person effectively. We analyze the assessments and experiments for prediction of such speaker-independent perceptions from speech making a two-fold contribution at this point in time. First, regarding the INTERSPEECH Speaker Trait Challenge [2] where participants will train automatic classifiers in order to recognize binary targets along the Big 5 personality traits, we provide insights by analyzing and discussing properties of the data and the labeling scheme used in the challenge. Second, we contribute to the general methodology and understanding on the reliability of personality assessment by humans and machines. Following up on previous work ([3, 4]), we introduce a new extension of our database and compare all results between the two corpora.

We follow the widely acknowledged "trait theory", which sees personality as a defined set of habitual patterns of behavior, thoughts, and emotions, that manifests itself in term of measurable traits. Attempting to asses these traits, most of the inventories measure the so called "Big 5" traits, e.g. the NEO-FFI [5] and the BFI-10 [6] used in this study. The five traits are:

- **O** Openness: Attitude towards new experiences in every day life, curiosity, e.g. open-minded vs. conservative
- **C** Conscientiousness: Quality of diligence, e.g. accurate, careful, reliable vs. carelessness, indifferent
- **E** Extroversion: Attitude towards outside, e.g. reserved, contemplating, vs. sociable, energetic, independent
- **A** Agreeableness: Ability of social reflection and trust, e.g. egocentric, competitive vs. sympathetic, trustful
- **N** Neuroticism: Emotional stability, e.g. calm, not easily agitated vs. unstable, unsure, easily driven by feelings

While the NEO-FFI comprises 60 items, i.e. 12 items per trait, the BFI-10 includes 10 items, i.e. two items per trait. Both tests are using 5-point Likert scales ranging from "strongly disagree" to "strongly agree", and are designed for both, self-assessment and observer's-assessment. Focusing on speech, we apply the tests to capture vocal manifestations of perceived personality in speech as presumed by a listener or interlocutor. In traditional application, raters have a multitude of cues on which to base their personality judgment, such as previous knowledge and experience. In our experiments, judgments are based on zero-acquaintance and auditory impression exclusively.

2. Related Work

Apple [7] finds that prosodic speech characteristics, such as pitch and speaking rate, can influence the attribution of truthfulness, empathy, and "potency". Scherer [8] analyzes personality traits and observes that extroverted speakers speak louder, and with fewer hesitations. He concludes, that extroversion is the only factor that can be reliably estimated from speech. Mairesse [9] also confirms prominence of prosodic properties for modeling extroversion, and that extroversion can be modeled best, followed by emotional stability (neuroticism) and openness.

Using a former version of the TPDB analyzed in this work (cf. Section 3), we resent results of an automatic prediction and classification of all the Big 5 traits on a subset of restricted, i.e. acted speech, consisting of fixed text passages in [3]. Factor analyses reveal a 5 factor structure capturing vocal personality from all traits but openness. Also automatic prediction and classification show most confusions with openness. Classifying one out of 10 high and low targets along the scales we reach around 60% accuracy, based on a large number of prosodic features. In [4], we analyze text-independent personality assessment of spontaneous speech, addressing also time-dependency and trait interplay. Showing consistency in labeling over time, we find that neuroticism and extroversion are inversely associated, i.e. increasing one causes significant decrease in perception of the other. Similarly, decreasing agreeableness, e.g. becoming more egocentric, causes a less open impression, increasing it acts the

opposite way. Conducting cluster analyses we see that neurotic stimuli are perceived clearly distinct from all others, while open and extroverted stimuli are perceived as similar.

Using a former version of the SPC corpus analyzed in this work (cf. Section 3), Mohammadi [10] classifies high and low targets along the Big 5 traits. She concludes, that only extroversion and conscientiousness can be classified satisfactorly from the non-acted, speaker-independent SPC data. In [11] she reports on two categories of speakers within the collection, i.e. professional and non-professional speakers, who are expected to differ along the conscientiousness trait. Result show between 60% and 72% accuracy for trait-dependent binary classification.

3. Databases and Labeling

Following up of [3] and [4], we introduce an new extension to the T-Labs Personality Database (TPDB), which consists of 65 German speakers (mostly students, 28 year avg., 56% male) recorded at short conversation tests, e.g. role-playing ticket reservations or food orders. High quality recordings (24bit, 44.1kHz sample rate) were done in two disjoint scenarios. First, 2 participants sitting in two separate sound proofed cabins were recorded using microphones. Second, the participants were sitting far from each other in a large anechoic chamber, separated by heavy curtains. In this part, recordings were done using headsets. To generate human personality labels, we followed the procedure described in [3]. 42 unique German raters (mostly students at Berlin Universities, mean age 29 years, 53% male) listened to 10-20 seconds speech excerpts in random order through high-quality headsets as often as they wanted to, while completing a series of NEO-FFI questionnaires about their first impression of the speaker's recordings. Taking one sample from each speaker, each stimulus was assessed by minimum 15 raters based on randomized orders.

The Speaker Personality Corpus (SPC) consists of 96 randomly extracted news bulletins in French, broadcast by Radio Suisse Romande, at a quality of 16bit, 8kHz sample rate. The collection comprises audio clips of about 10 seconds length from 322 unique speakers. Out of the 640 audio clips in total, 307 are produced by professional speakers, i.e. journalists that talk regularly on the radio. 333 samples from unique 210 speakers are non-professional speakers. Focusing on vocal personality perception, the stimuli were selected not to contain words, such as places or well known people, that might be understood by individuals who do not speak French. In Addition, raters did not speak French. Importantly, and despite the fact that the actual language spoken is French, the raters assigned personality from an language unknown to them, which could in principle be any language. Judgments thus were made out of their acquired extra-linguistic perspective, since the raters were of English origin or were well acquainted with English extra-linguistic stereotypes. For each of the randomly presented clips, 11 independent raters filled out a BFI-10 questionnaire. The raters were not aware of any proficiency split in the corpus.

4. Experiments

In our experiments we analyze the reliability of personality assessments, and the consistency of the applied questionnaires. Further, automatic prediction from prosodic features will be analyzed. As presented in Section 3, the databases predominantly differ with respect to a) questionnaires used for assessment; b) speaker proficiency; c) the languages spoken; d) the neutralization of linguistic interference, and e) the conversational situation of the speaker. Experimentation and interpretation will address these factors whenever possible.

Table 1: Consistencies and correlations (diagonal cells) between traits for TPDB (top) and SPC (bottom) databases.

	O	**C**	**E**	**A**	**N**
O	(.79)	.17	.46	.53	-.29
C		(.93)	.12	.13	-.39
E			(.88)	.32	-.57
A				(.90)	-.03
N					(.90)

	O	**C**	**E**	**A**	**N**
O	(.01)	.22	.18	.13	-.08
C		(.60)	.16	.26	-.19
E			(.64)	-.07	.03
A				(.46)	-.39
N					(.61)

4.1. Reliability

Lilliefors tests, resembling Kolmogorov-Smirnov tests for normality with mean and variance unknown, reveal overall normal distributions for TPDB ratings on all traits ($p<0.05$). The average share of non-normal distributions results below 9%. The top part of Table 1 therefore shows intra-trait correlations according to Pearson. Averaging at 0.30, the highest correlations were found between O and E as well as between O and A. While being weak in magnitude, this association links speakers perceived as more open to speakers perceived as more extroverted and more agreeable. Another weak association inversely links N and E, i.e. the more extrovert, the less neurotic the speakers are perceived. Estimating consistency we calculate Cronbachs Alpha, which results in between 0.79 for O as lowest, and 0.93 for C as highest figure. Further, sample-dependent consistency for O shows more than twice the variation of all other traits. Consequently, answering items on openness reveals most equivocal and sample-dependent for the raters. However, the average value of 0.88 corresponds to good, almost excellent reliability.

Analyzing the SPC collection, we first select only non-professional speakers, to be comparable to TPDB. Lilliefors tests show, that ratings distributions for virtually every third sample do not resemble normal distributions ($p<0.05$), O and A being most pivotal. Accordingly, Table 1 shows inter-trait correlations due to Spearman. Averaging at 0.17, the overall magnitude is very low. The highest correlations is found to associate A and N inversely, i.e. the more neurotic, the more egocentric and distrustful the speakers are perceived. Another correlation higher than average is found between A and C, however, all these correlations are weak in magnitude. Since Cronbachs Alpha cannot be calculated on basis of two item per trait only, we calculate reliability by Spearman correlations corrected by the Spearman-Brown prediction formula, which is commonly used with altering test length. As a result, ratings on O seem not to be reliable at all. Weak reliability is observed for A, while the remaining traits result on a low-moderate level.

Inter-trait correlations in original NEO-FFI and BFI-10 application result very weak on average, i.e. 0.14 and 0.11 respectively. While this is also the case for SPC personality assessments, the German assessments show generally higher inter-

dependency. However, the relative correlation pattern matches previous findings ([3, 4]) very well.

We see a different picture when looking at the SPC results, where raters disagree to a larger extent, or even disagree completely, as for O. We can only speculate about whether this is caused by the lack of linguistic information in SPC annotations, or whether the high agreement on German data is caused by the nature of German extra-linguistic stereotypes. Also, while TPDB consists of students doing conversational test cycles, SPC speakers might have been affected by emotional reactions, nervousness or excitement, due to being broadcasted "on Air". Still, consistencies are generally much higher than inter-dependencies, consequently raters were well able to assess the different traits as independent traits in both languages.

When analyzing professional speakers from SPC we see, that consistencies increase only for E and A to 0.69 and 0.50 respectively. The overall inter-trait correlation increases moderately to 0.22 affecting all traits equally. We also see these finding from acted speech in TPDB and [3].

4.2. Factor Analysis

To explore latent factors in our data, we conduct an exploratory factor analysis, hypothesizing the presence of 5 factors in the ratings. We apply maximum likelihood component extraction to reveal any latent variables that cause the hypothesized factors to covary and rotate them using orthogonal Varimax rotation. Disregarding loadings below 0.4 we obtain well-structured item loadings as shown in Table 2. The first column shows the extracted factors decreasing by explanatory power from row to row. The second and fourth columns show the items that load on these factors, with descending order of loading magnitudes from left to right. For example, the most powerful factor extracted from the NEO-FFI ratings is assembled from items originally designed to load on C factor. More accurately, all of the 12 items ($C1 : C12$) that are to load on C prove to load on just one factor in our data. Hence, the first and most important factor $F1$ can be titled C for NEO-FFI, N for BFI-10 respectively. Accordingly, $F2$ can be titled N for NEO-FFI and C for BFI-10. Columns 3 and 5 show the percentage of explained variance by the factors, so it seems that the confusion between N and C is of minor importance. $F3$ can be entitled A. Note, that although the second BFI-10 item which is to load on this factor does actually not contribute to it, the remaining item explains a share of variance almost comparable to $F3$ on NEO-FFI. $F4$ can be seen as E, explaining less variance for NEO-FFI than for BFI-10. $F5$ can clearly be seen as O for the NEO-FFI, but for BFI-10 a cross-loading item from E actually contributes to a higher extent. Consequently, no O factor can be found from BFI-10. The fifth factor reveals minor importance for both questionnaires, and cross-loadings are only few in number.

Eventually, the sequence and explanatory power of the extracted factors, and the revealed factor structure proves very similar for both questionnaires, even given the inherent differences in language, questionnaire length, and linguistic information overlay at hand. While the overall goodness-of-fit of the 5 factor models cannot cover all of the variance seen in our data (71.0% for SPC, 86.5% for TPDB), the revealed latent structures clearly support the application of the chosen inventories for personality assessment from speech. Our intuitive explanation why openness could not be extracted using the BFI-10 on SPC is that openness was not consistently rated in the data. However, the proof of successful extraction of openness in general when using BFI-10 remains to be shown.

Table 2: Latent factor structures for the chosen questionnaires.

Factor	NEO-FFI	% Var	BFI-10	% Var
F1	C1:C12,O	25.0	N,N	23.7
F2	N1:N12,A,E,E	23.5	C,C	22.4
F3	A1:A11,O,E,E	23.5	A	20.5
F4	E1:E8,O,E,E	16.8	E,E	19.9
F5	O1:O6	11.3	E,O	13.5

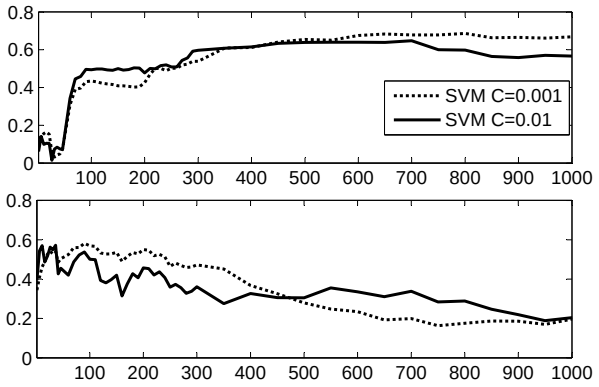

Figure 1: Correlation between automatically predicted extroversion values and human annotations for SPC (top) and TPDB (bottom). Feature space dimensionality used in SVM regression expands along IGR ranking on x-axis, solid and dashed lines show complexity parameter settings.

Also when factor-analyzing professional speakers in SPC, results are similar, and openness could not be extracted. Compared to [3] we observe a similar loading structure. Also here, O items show diffuse loadings patterns.

4.3. Automatic Prediction of Personality

In order to analyze automatic personality prediction we conduct an experiment using the mean of all ratings for a given sample as ground-truth. For prediction, we use SVM regression.

Models are trained with sample-level features that are automatically calculated from audio descriptors extracted at a 10ms frame shift. The descriptors can be sub-divided into 7 groups, i.e. *intensity, pitch, loudness, formants, spectrals, MFCC* and *other* features like duration and rhythm related characteristics. Statistics like means, moments, extrema and ranges are then applied to the descriptors. To span up the feature space we ranked the features according to their Information-Gain-Ratio (IRG) after discretization into 10 nominal bins. To obtain general estimates, we generate the ranking and all further classification results using 10-fold cross-validation. Overall, we generate about 1.5k features, which we have introduced in more detail in previous work on emotion recognition [12].

The results of the prediction experiments for extroversion are given in Figure 1, showing correlation between predicted scores and human labels along IGR expansion of the feature space for up to 1k features on the x-axis. Interestingly, the curve on top for SPC prediction reaches a maximum of almost 0.7 when including a high number of features, although the biggest jump occurs at around 50 features. Looking at the included features we predominantly find loudness, intensity and MFCC related features in high ranks. For TPDB we reach lower corre-

lation of almost 0.6, but at the costs of 35 features only, which are again mostly loudness statistics. Further, adding to many features seem to harm the performance, since the curves are decreasing. As a conclusion, the degree of perceived extroversion follows the dynamics and intensity-related perceptions for both our corpora. Prediction for other traits reveals more challenging. Generally, most of the remaining models resulted in correlations below 0.4, with the exception of neuroticism and conscientiousness on TPDB, which results in roughly 0.5 but allows no systematic insights from the rankings.

However, two interesting observations can be made, when building models on the professional speakers from SPC. Here we reach correlations to neuroticism ratings of almost 0.6 with roughly 100 features on dynamics as well. Hence, in order to express degrees of neuroticism, professional speakers seem to use dynamics in more predictable ways. On the opposite side, when it comes to expressing extroversion, correlations based on dynamics are much lower for professional speakers (0.4) than for non-professional speakers. Accordingly, professional speakers might not express themselves in terms of extroversion as overtly or as prototypical as non-professional speakers do.

As expected, the overall prediction quality decreases when comparing results to text-dependent work in [3]. However, the high value of dynamic-related features for prediction of extroversion shows congruence for all German and English data and other literature, e.g. [8, 9].

5. Summary and Outlook

This paper reports on the assessment and prediction of speaker-independent perceived personality from speech and makes a contribution in two ways. First, analyzing the data provided with the INTERSPEECH Speaker Trait Challenge (SPC) we observe considerably low labeling consistencies for assessment of openness. While also agreeableness labels show lower consistencies compared to other traits, the analysis of inter-trait dependencies shows that traits have generally been assessed as individual traits. When factor-analyzing the ratings from speech, we find the applied personality assessment scheme capable of capturing all traits but openness, that is, as capable as when applied to psychological personality assessment in English. Further, we extract signal-based prosodic features in order to predict the actual personality annotations. As a result, we obtain moderate correlations to human annotations for extroversion.

As s second contribution, we compare all results from SPC to a newly introduced extension to our previous personality data collection comprising professional speakers. Consisting of non-professional speakers, our new data shows overall high consistency. More inter-dependencies have been observed, e.g. speakers presumed to be more open are also presumed to be more extroverted and more agreeable. Further, the more extrovert, the less neurotic the speakers are perceived. Factor structures resulted congruent to SPC data and to former results on text-dependent and speaker-dependent analysis. Predicting the actual personality values results in moderate correlations for extroversion, neuroticism and conscientiousness. Interestingly, results suggest that professional speakers use prosodic dynamics such as loudness and intensities in more predictable ways when expression degrees of neuroticism. On the opposite side, they might not express themselves as prototypical as non-professional speakers do on extroversion.

In this work we compare results for two languages, using different assessments schemes. A closer look reveals that while the SPC corpus contains French speech, the labelers were selected not to understand French. This way the assessment of SPC can be expected to be considerably less affected by linguistic content than for TPDB, which could possibly lead to lower openness perceptions as well. The question how English-acquainted raters translate French speech gestures into English personality stereotypes and how this relates to cultural differences and stereotypes remains open. With respect to the prediction experiment, the shown IGR curves shows some ripple including notches, which indicates that the feature selection can be further improved, e.g. by wrapper based subset selection. Also the difference in signal quality and the absolute number of speakers could have affected the overall performance.

Future work will be bound to the availability of data. In addition, capturing emotional states on top of personalities will be beneficial to estimate a joint behavior from speech. Also, the agreement on personality stereotypes might not be as high as for, e.g. , emotions. Therefore, knowing the personality characteristics of the person giving judgment will contribute to an comprehensive understanding on who judges about who. Finally, we emphasize that this work does not attempt to assess personality as common in psychology, but the personality impression humans leave when speaking to each other.

6. References

[1] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction," *Journal of Experimental Psychology*, pp. 171–181, 2001.

[2] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The Interspeech 2012 Speaker Trait Challenge," in *Proc. Interspeech 2012*, 2012.

[3] T. Polzehl, S. Möller, and F. Metze, "Automatically assessing acoustic manifestations of personality in speech," in *Workshop on Spoken Language Technology*. Berkeley, U.S.A.: IEEE, 2010.

[4] T. Polzehl, S. Möller, and F. Metze, "Modeling speaker personality using voice," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2011)*. Florence, Italy: ISCA, 2011.

[5] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual*. Psychological Assessment Resources, 1992.

[6] B. Rammstedt and O. 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.

[7] W. Apple, L. A. Streeter, and R. M. Krauss, "Effects of pitch and speech rate on personal attributions," *Journal of Personality and Social Psychology*, vol. 37, no. 5, pp. 715–727, 1979.

[8] K. R. Scherer and U. Scherer, "Speech Behavior and Personality," *Speech Evaluation in Psychiatry*, pp. 115–135, 1981.

[9] 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," *Journal of Artificial Intelligence Research (JAIR)*, vol. 30, pp. 457–500, 2007.

[10] G. Mohammadi, M. Mortillaro, and A. Vinciarelli, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of the International Workshop on Social Signal Processing*, 2010, pp. 17–20.

[11] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. to appear, 2012.

[12] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication, Special Issue: Sensing Emotion and Affect - Facing Realism in Speech Processing*, 2011.