PERSONALITY PREDICTION BASED ON INTONATION STYLIZATION

Uwe D. Reichel

Institute of Phonetics and Speech Processing, University of Munich
reichelu@phonetik.uni-muenchen.de

ABSTRACT

This study's aim is to predict speaker personality traits from intonation patterns in spoken dialogs. Intonation patterns were extracted by a parametric superpositional stylization approach that allows for pattern description on a parametric as well as on a categorical level. Based on features derived from these representations we trained support vector machines and fitted generalized linear regression models to predict speaker personality with respect to the four dimensions *acting, extroversion, other-directedness*, and *sensitivity*. The personality classification accuracies ranged from 79 to 91%.

Keywords: intonation, stylization, personality, machine learning

1. INTRODUCTION

In search of correlates of a speaker's personality in spoken utterances various acoustic and linguistic parameters have been addressed in previous studies. Among the most common and partly overlapping personality category schemes are the "big five" traits proposed by [9] *openness to experience, conscientiousness, extra version, agreeableness, neuroticism*, and the four dimensions used in the extended self-monitoring scale of [25, 27] *acting, extroversion, other-directedness, sensitivity*. Commonly examined acoustic features are: pitch mean, range, and variance [6, 8, 19], speaking rate [24, 8], intensity [6, 8], and voice quality [21, 6]. Linguistic features comprise amongst others lexical cues [8, 19], type/token counts and part of speech usage [8]. These features were employed for automatic personality classification e.g. by [8, 13, 6] as well as in expressive speech synthesis [26] to systematically vary the generated personality.

This study's focus is on intonation correlates of personality aspects. Instead of examining coarse pitch features mentioned above, namely global F0 mean, range, and variance, we aim to examine intonation personality in a more fine-grained way in terms of intonation stylization parameters and contour classes.

2. DATA

Corpus We used the GECO corpus, which was recorded, orthographically transcribed, signal-text aligned, and automatically annotated on the segment and syllable level at the Institute for Natural Language Processing (IMS) Stuttgart, Germany by [22, 23]. It contains 46 German spoken dialogs, each of approximately 25 minutes length, between 13 previously unacquainted female subjects. The total duration amounts about 20 hours. Signal and text were aligned on the phone, syllable, and word level by the aligner of [14]. Moreover, the corpus contains mutual ratings and self-monitoring information about the interlocutors. The latter is used for the current study.

Self-monitoring scale In GECO the participants' personality aspects were tested by a questionnaire developed for the self-monitoring scale of [27], which is a German adaption of the scale of [25]. Self-monitoring is defined as a person's ability to adapt his/her behavior to external situational factors, and can be quantified along four personality dimensions introduced below. The questionnaire comprises 35 items (25 from [25] as well as 10 additional items from [27]) that were presented in the same random order to all subjects. The items were posed as statements, about which the subjects had to state whether they "agree" or "disagree". Each item was designed to be indicative for one of four aspects of personality:

- *acting (AC)*, i.e. self-manifestation in front of others; 11 items; example: "I can make impromptu speeches even on topics about which I have almost no information"; supporting answer: "agree",
- *extroversion (EV)*, i.e. active outward behavior; 7 items; "In a group of people I'm rarely the center of attention"; "disagree",
- *other-directedness (OD)*, i.e. orientation towards others' behaviors and opinions; 9 items; "When I am uncertain how to act in a social situation, I look to the behavior of others for cues"; "agree", and

- *sensitivity (SN)* to expressive behavior and social cues; 8 items; “When with a group of people, I can normally foresee the others’ reactions to my behavior”; “agree”.

One subject did not answer these items, so that the two dialogs this subject took part in were dismissed from further analyses.

3. INTONATION STYLIZATION

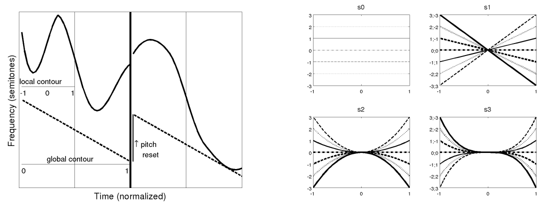
For intonation stylization we adopt the parametric CoPaSul approach of [17], which is illustrated in the left half of Figure 1. Within this framework intonation is stylized as a superposition of linear global contours, and third order polynomial local contours. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. The domain of local contours is determined by placing boundaries behind each content word determined by POS tagging [15]. Thus these local contour domains roughly correspond to syntactic chunks [1] and generally contain at most one pitch accent. As in [10, 17] the global and local contour parameter vectors are clustered to derive intonation contour classes. Intonation patterns thus can be described in parametric as well as in category terms.

Preprocessing F0 was extracted by autocorrelation (PRAAT 5.3.16 [2], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [20]. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

Parameterization The global linear component is given by the F0 baseline. Following [18] a time window is shifted along the F0 contour, and within each window the median of all values below the 10th percentile is calculated. The baseline then is fitted to this sequence of medians. [18] have shown, that this median-based method is less error-prone than fitting a line through local F0 minima.

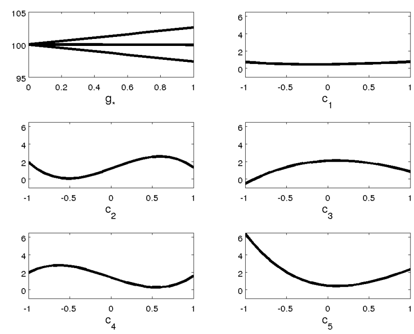
The baseline is then subtracted from the F0 contour, and a third order polynomial is fitted to the F0 residual within each local segment. Time is normalized to the range from -1 to 1 so that time 0 is placed in the mid of the content word’s syllable bearing the lexical stress. Lexical stress is identified by the BALLOON toolkit [16]. This normalization allows for capturing pitch peak alignment with respect to the accented syllable.

Figure 1: Left: CoPaSul stylization of global and local intonation components. **Right:** Variation of 3rd order polynomial coefficients to capture local contour shapes.



Contour clustering To allow for an additional categorical description, the slopes of the global contours as well as the polynomial coefficients of the local contours are clustered by the Kmeans [7]. Following [17] the optimal number of contour classes was initialized by subtractive clustering [3], whose parameters were optimized by the Nelder-Mead [11] method. In [17] this way of determining initial cluster centers turned out to yield stable clustering results on disjunct data subsets.

Figure 2: Global and local contour classes.



Parameter level features The linear global contour coefficient represents the declination slope. As can be seen in the right half of Figure 1 the local contour polynomial coefficients are related to several aspects of local F0 contours. Given the polynomial $\sum_{i=0}^{3} s_i \cdot t^i$, s_0 is related to the local F0 level relative to the baseline. s_1 and s_3 are related to the general F0 trend (rising or falling) and to peak alignment. s_2 determines the peak shape (convex or concave) and its acuity. This parameterization thus allows to relate means and variances of distinct F0 aspects as level, trend, peak shape, and alignment to personality aspects. For high-level AC and EV speakers a more extrovert speaking style is expected. Previous findings summarized in [8] revealed a positive correlation between extrovert speech and F0 variability. In terms of our proposed stylization extrovert speaking style and thus high-level AC and EV is expected to be characterized by more pronounced F0 movements for example reflected in higher s_0 coefficient values, and by more variable F0 movements reflected by higher variances of all coefficients.

Class level features On the categorical level contour class probabilities show, whether F0 movements tend to be more or less pronounced. Local contour class 2 with a relatively prominent peak height is an example for the former, whereas the flat class 1 stands for the latter. Variability is measured in terms of contour class bigram entropy. The entropy for all class bigram types B observed for a speaker within a dialog is given by $H(B) = -\sum_{b \in B} p(b) \cdot \log_2 p(b)$. $p(b)$ denotes the conditional probability of a contour class given the preceding one. The higher the entropy the less predictable a contour class given the preceding class, and thus the more variable the intonation unit sequence. Therefore, the greater F0 variability expected for high-level AC and EV can be expressed in categorical terms by higher class bigram entropy values.

All parameter and class level features are summarized in Table 1. These features are examined with respect to their discriminatory power between the high and low level for each personality dimension. Furthermore, based on these features personality dimension classifiers and regression models are trained.

Table 1: Intonation features.

Features	Description	Number
Class-level features		
$H(G), H(C)$	global and local class bigram entropies	2
$P(g_*), P(c_*)$	global and local class probabilities	8
Parameter-level features		
$\mu(u_1), \sigma(u_1)$	mean and variance of the baseline slope	2
$\mu(s_*), \sigma(s_*)$	means and variances of the local contour coefs	8

4. PREDICTION OF PERSONALITY ASPECTS

Features and targets In this study we did not address self-monitoring as a whole, but focused on each of the four personality dimensions listed in section 2 in isolation. For each participant and personality aspect the proportion of matches between the participant's and the aspect supporting answers was calculated yielding numbers between 0 and 1. Then for each interlocutor in a dialog, we aimed to predict for each of the four personality aspects (1) whether the speaker's match to this aspect is high or low, and (2) the matching score itself. (1) is a binary classification task, and (2) a regression task.

Except of AC the variability of the personality match proportions was low, for OD and SN all matches were above 0.61, and for EV even above 0.72. To make use of the entire data, for the classification task we thus distinguish the two classes "above" and "below the respective match median", and for the regression task, the target values were normalized (stretched) to the interval from 0 to 1.

For both tasks the feature vector consists of 20 variables introduced in Table 1. Their values were calculated for each speaker over a whole dialog tier. All predictors were z-transformed and orthogonalized by a principal component analysis.

Prediction methods For the two-category classification tasks *high vs. low personality level*, we employed support vector machines (SVM) [4] with a third order polynomial kernel function. The separating hyperplane was derived by sequential minimal optimization.

The regression task to predict the personality matching scores was accomplished by generalized linear models (GLM) [12] using a binomial distribution. The output was mapped to the interval from 0 to 1 by a logit link function.

5. RESULTS

Intonation patterns As illustrated in Table 2 especially the personality dimensions AC and EV are well distinguishable by the intonation variables. For AC the tendencies are in line with our expectation, that a high AC level is reflected in high class- and parameter-level variabilities as well as in pronounced F0 movements. This is expressed in significantly higher class entropies H, coefficient variabilities σ, offset coefficient values $\mu(s_0)$, and by higher probabilities of pronounced local intonation classes c_{2-5} and a lower probability of the flat class c_1. However, for EV the reverse pattern emerged: High-level EV is related to significantly lower entropies and variances than low-level EV. As expected, for the dimensions OD and SN a fewer number of significant intonation differences is observed.

Table 2: Relations between stylization variables and personality dimensions: significantly higher $>$ or lower $<$ variable values for the high level personality group (two-sided Mann Whitney, resp. Welch tests, $p < 0.01$). '–' indicates no significant difference. g_* and c_* stand for global and local contour classes, u_*, s_* for global and local stylization coefficients, respectively. See Table 1 for feature description.

	AC	EV	OD	SN
$P(g_1)$	$<$	$>$	$>$	–
$P(g_2)$	$>$	–	$<$	–
$P(g_3)$	$>$	$<$	–	–
$P(c_1)$	$<$	$>$	$>$	–
$P(c_2)$	$>$	$<$	–	–
$P(c_3)$	$>$	$<$	–	–
$P(c_4)$	$>$	$<$	$<$	–
$P(c_5)$	$>$	–	$<$	$>$
$H(G)$	$>$	$<$	$<$	–
$H(C)$	$>$	$<$	$<$	–
$\mu(u_1)$	–	$<$	$<$	–
$\sigma(u_1)$	$>$	$<$	$<$	–
$\mu(s_0)$	$>$	$<$	$<$	–
$\mu(s_1)$	$>$	$<$	$>$	$<$
$\mu(s_2)$	–	–	–	$>$
$\mu(s_3)$	–	$>$	–	–
$\sigma(s_0)$	$>$	$<$	$<$	–
$\sigma(s_1)$	$>$	$<$	$<$	–
$\sigma(s_2)$	$>$	$<$	$<$	–
$\sigma(s_3)$	$>$	$<$	$<$	–

Personality prediction The classification and regression results of a 10-fold cross validation are shown in Figure 3. The mean classification accuracies range from 78.9% for both EV and OD to 91.1% for AC . The mean accuracy for SN amounts 84.4%. Since the median personality match scores were taken as category boundaries, the baseline accuracy given by random assignment is 50%. All classification accuracies turned out to be significantly higher than this baseline (one sided sign rank tests, $p < 0.01$)

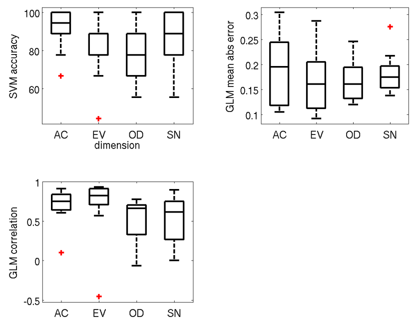
Regression was evaluated in terms of the correlation r between the reference and the predicted scores, and by the mean absolute error e. Sorted by correlation the performance again was best for the dimension AC (mean $r = 0.70$, $e = 0.19$), followed by EV ($r = 0.69$, $e = 0.17$), SN ($r = 0.53$, $e = 0.18$), and OD ($r = 0.52$, $e = 0.17$). All mean correlations differed significantly from 0 (one sided sign rank test, $p < 0.05$).

A sequential feature selection did not further improve the results.

6. DISCUSSION

As to be seen in Table 2, for dimension SN the discriminatory power of the examined features is low.

Figure 3: Prediction performance for each personality dimension after 10-fold cross validation. Classification accuracy of the support vector machines (SVM). Correlation and mean absolute error of the generalized linear models (GLM).



A possible reason is, that SN rather refers to information reception than to reactions to them, so that different levels might to a lesser extent be expressed in observable signals. Alternatively, high-level SN as well as OD might express themselves more in the interaction with the interlocutor. Since in this study all features were extracted within a single speaker, such interactions are not covered. Thus, if the feature pool would be enlarged by intonation convergence measures used in entrainment research [5], GLM performance for OD and SN might increase.

A possible explanation for the puzzling finding, that for EV most intonation variables behaved opposite to the expectations, might be that all subjects matched EV by a proportion of at least 0.72, which indicates, that all subjects were rather extroverted and thus not well dividable into two classes.

Compared to previous rather coarse pitch examinations, the current stylization-based approach allows for a more fine-grained examination of the interplay between personality and intonation since it covers diverse pitch pattern aspects on the parametric and the categorical level. This approach could also be of use in expressive speech synthesis in predicting the derived contour classes not only by linguistic concepts as discourse structure [17] but taking into account personality-related class priors and realization variability.

This study addressed each personality dimension of the self-monitoring scale in isolation. A subsequent aim is thus to predict a speaker's self-monitoring degree as a whole based on an enlarged feature pool that consists of further prosodic, convergence and text-level features.

7. REFERENCES

[1] Abney, S. 1991. Parsing by chunks. In: Berwick, R., Abney, S., Tenny, C., (eds), *Principle-Based Parsing*. Dordrecht: Kluwer Academic Publishers 257–278.

[2] Boersma, P., Weenink, D. 1999. PRAAT, a system for doing phonetics by computer. Technical report Institute of Phonetic Sciences of the University of Amsterdam. 132–182.

[3] Chiu, S. 1994. Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligence & Fuzzy Systems* 2(3), 267–278.

[4] Cortes, C., Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20.

[5] Levitan, R., Hirschberg, J. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proc. Interspeech* Florence, Italy. 3081–3084.

[6] Liu, C.-J., Wu, C.-H., Chiu, Y.-H. 2013. BFI-based speaker personality perception using acoustic-prosodic features. *Proc. APSIPA*.

[7] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability* volume 1 281–297.

[8] Mairesse, F., Walker, M., Mehl, M., Moore, R. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.* 30(1), 457–500.

[9] McRae, R., John, O. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 175–215.

[10] Möhler, G., Conkie, A. 1998. Parametric modeling of intonation using vector quantization. *Proc. 3rd ESCA Workshop on Speech Synthesis 1998* Jenolan Caves, Australia. 311–316.

[11] Nelder, J., Mead, R. 1965. A simplex method for function minimization. *Computer Journal* 7, 308–313.

[12] Nelder, J., Wedderburn, R. 1972. Generalized linear models. *J. Royal Statistical Society* 135(3), 370–384.

[13] Polzehl, T., Schoenenberg, K., Möller, S., Metze, F., Mohammadi, G., Vinciarelli, A. 2012. On speaker-independent personality perception and prediction from speech. *Proc. Interspeech* Portland, US.

[14] Rapp, S. 1995. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov models – An aligner for German. *Proc. ELSNET Goes East and IMACS Workshop Integration of Language and Speech in Academia and Industry* Moscow, Russia.

[15] Reichel, U. 2005. Improving Data Driven Part-of-Speech Tagging by Morphologic Knowledge Induction. *Proc. AST Workshop 2007* Maribor, Slovenia. 65–73.

[16] Reichel, U. 2012. Perma and Balloon: Tools for string alignment and text processing. *Proc. Interspeech 2012* Portland, Oregon. paper no. 346.

[17] Reichel, U. 2014. Linking bottom-up intonation stylization to discourse structure. *Computer, Speech, and Language* 28, 1340–1365.

[18] Reichel, U., Mády, K. 2014. Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian. *Proc. Interspeech 2014* Singapore. 111–115.

[19] Rosenberg, A., Hirschberg, J. 2005. Acoustic/prosodic and lexical correlates of charismatic speech. *Proc. Interspeech* Lisbon, Portugal. 513–516.

[20] Savitzky, A., Golay, M. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627–1639.

[21] Scherer, K. 1978. Personality inference from voice quality: The loud voice of extroversion. *European J. Social Psychology* 467–487.

[22] Schweitzer, A., Lewandowski, N. 2013. Convergence of articulation rate in spontaneous speech. *Proc. Interspeech* Lyon, France. 525–529.

[23] Schweitzer, A., Lewandowski, N. January 2015. IMS GECO database. http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.en.html.

[24] Smith, B., Brown, B., Strong, W., Rencher, A. 1975. Effects of speech rate on personality perception. *Language and Speech* 18, 145–152.

[25] Snyder, M. 1974. Self-monitoring of expressive behavior. *J. of Personality and Social Psychology* 30, 526–537.

[26] Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., Barry, W. 2006. Modelling personality features by changing prosody in synthetic speech. *Proc. Speech Prosody 2006* Dresden, Germany.

[27] von Collani, G., Stürmer, S. 2009. Deutsche Skala zur Operationalisierung des Konstrukts Selbstüberwachung (Self-Monitoring) und seiner Facetten. In: Glöckner-Rist, A., (ed), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. Bonn, Germany: GESIS. ZIS Version 13.00.