Automatic Recognition of Personality in Conversation

François Mairesse
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
F.Mairesse@sheffield.ac.uk

Marilyn Walker
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
M.A.Walker@sheffield.ac.uk

Abstract

The identification of personality by automatic analysis of conversation has many applications in natural language processing, from leader identification in meetings to partner matching on dating websites. We automatically train models of the main five personality dimensions, on a corpus of conversation extracts and personality ratings. Results show that the models perform better than the baseline, and their analysis confirms previous findings linking language and personality, while revealing many new linguistic and prosodic markers.

1 Introduction

It is well known that utterances convey information about the *speaker* in addition to their semantic content. One such type of information consists of cues to the speaker's *personality traits*, typically assessed along five dimensions known as the Big Five (Norman, 1963):

- Extraversion (sociability, assertiveness)
- Emotional stability (vs. neuroticism)
- Agreeableness to other people (friendliness)
- Conscientiousness (discipline)
- Intellect (openness to experience)

Findings include that extraverts talk more, louder, and faster, with fewer pauses and hesitations, and more informal language (Scherer, 1979; Furnham, 1990; Heylighen and Dewaele, 2002; Gill and Oberlander, 2002). Neurotics use more 1st person singular pronouns and negative emotion words, while conscientious people avoid negations and negative emotion words (Pennebaker and King, 1999). The use of words related to insight and the avoidance of past tense indicate intellect, and swearing and negative emotion words mark disagreeableness. Correlations are higher in spoken language, possibly especially in informal conversation (Mehl et al., in press).

Previous work has modeled emotion and personality in virtual agents, and classified emotions from actor's speech (André et al., 1999; Liscombe et al., 2003). However, to our knowledge no one has tested whether it is possible to automatically recognize personality from conversation extracts of unseen subjects. Our hypothesis is that automatic analysis of conversation to detect personality has application in a wide range of language processing domains. Identification of leaders using personality dimensions could be useful in analyzing meetings and the conversations of suspected terrorists (Hogan et al., 1994; Tucker and Whittaker, 2004; Nunn, 2005). Dating websites could analyze text messages to try to match personalities and increase the chances of a successful relationship (Donnellan et al., 2004). Dialogue systems could adapt to the user's personality, like humans do (Reeves and Nass, 1996; Funder and Sneed, 1993). This work is a first step toward individual adaptation in dialogue systems.

We present non-linear statistical models for ranking utterances based on the Big Five personality traits. Results show that the models perform significantly better than a random baseline, and that prosodic features are good indicators of extraversion. A qualitative analysis confirms previous findings linking language and personality, while revealing many new linguistic markers.

2 Experimental method

Our approach can be summarized in five steps: (1) collect individual corpora; (2) collect personality

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 85–88,
New York, June 2006.

ratings for each participant; (3) extract relevant features from the texts; (4) build statistical models of the personality ratings based on the features; and (5) test the learned models on the linguistic outputs of unseen individuals.

2.1 Spoken language and personality ratings

The data consists of daily-life conversation extracts of 96 participants wearing an Electronically Activated Recorder (EAR) for two days, collected by Mehl et al. (in press). To preserve the participants' privacy, random bits of conversation were recorded, and only the *participants'* utterances were transcribed, making it impossible to reconstruct whole conversations. The corpus contains 97,468 words and 15,269 utterances. Table 1 shows utterances for two participants judged as introvert and extravert.

Introvert:
- Yeah you would do kilograms. Yeah I see what you're saying.
- On Tuesday I have class. I don't know.
- I don't know. A16. Yeah, that is kind of cool.
- I don't know. I just can't wait to be with you and not have to do this every night, you know?
- Yeah. You don't know. Is there a bed in there? Well ok just...
Extravert:
- That's my first yogurt experience here. Really watery. Why?
- Damn. New game.
- Oh.
- Yeah, but he, they like each other. He likes her.
- They are going to end up breaking up and he's going to be like.

Table 1: Extracts from the corpus, for participants rated as extremely introvert and extravert.

Between 5 and 7 independent observers scored each extract using the Big Five Inventory (John and Srivastava, 1999). Mehl et al. (in press) report strong inter-observer reliabilities for all dimensions ($r = 0.84$, $p < 0.01$). Average observers' ratings were used as the scores for our experiments.

2.2 Feature selection

Features are *automatically* extracted from each extract (see Table 2). We compute the ratio of words in each category from the LIWC utility (Pennebaker et al., 2001), as those features are correlated with the Big Five dimensions (Pennebaker and King, 1999). Additional psychological characteristics were computed by averaging word feature counts from the MRC psycholinguistic database (Coltheart, 1981). In an attempt to capture initiative-taking in conversation (Walker and Whittaker, 1990; Furnham, 1990), we introduce utterance type features using heuristics on the parse tree to tag each utterance as a command, prompt, question or assertion. Overall tagging accuracy over 100 randomly selected utterances is 88%. As personality influences speech, we also use Praat (Boersma, 2001) to compute prosodic features characterizing the voice's pitch, intensity, and speech rate.

LIWC FEATURES (Pennebaker et al., 2001):

· STANDARD COUNTS:
- Word count (WC), words per sentence (WPS), type/token ratio (Unique), words captured (Dic), words longer than 6 letters (Sixltr), negations (Negate), assents (Assent), articles (Article), prepositions (Preps), numbers (Number)
- Pronouns (Pronoun): 1st person singular (I), 1st person plural (We), total 1st person (Self), total 2nd person (You), total 3rd person (Other)

· PSYCHOLOGICAL PROCESSES:
- Affective or emotional processes (Affect): positive emotions (Posemo), positive feelings (Posfeel), optimism and energy (Optim), negative emotions (Negemo), anxiety or fear (Anx), anger (Anger), sadness (Sad)
- Cognitive Processes (Cogmech): causation (Cause), insight (Insight), discrepancy (Discrep), inhibition (Inhib), tentative (Tentat), certainty (Certain)
- Sensory and perceptual processes (Senses): seeing (See), hearing (Hear), feeling (Feel)
- Social processes (Social): communication (Comm), other references to people (Othref), friends (Friends), family (Family), humans (Humans)

· RELATIVITY:
- Time (Time), past tense verb (Past), present tense verb (Present), future tense verb (Future)
- Space (Space): up (Up), down (Down), inclusive (Incl), exclusive (Excl)
- Motion (Motion)

· PERSONAL CONCERNS:
- Occupation (Occup): school (School), work and job (Job), achievement (Achieve)
- Leisure activity (Leisure): home (Home), sports (Sports), television and movies (TV), music (Music)
- Money and financial issues (Money)
- Metaphysical issues (Metaph): religion (Relig), death (Death), physical states and functions (Physcal), body states and symptoms (Body), sexuality (Sexual), eating and drinking (Eating), sleeping (Sleep), grooming (Groom)

· OTHER DIMENSIONS:
- Punctuation (Allpct): period (Period), comma (Comma), colon (Colon), semi-colon (Semic), question (Qmark), exclamation (Exclam), dash (Dash), quote (Quote), apostrophe (Apostro), parenthesis (Parenth), other (Otherp)
- Swear words (Swear), nonfluencies (Nonfl), fillers (Fillers)

MRC FEATURES (Coltheart, 1981):

Number of letters (Nlet), phonemes (Nphon), syllables (Nsyl), Kucera-Francis written frequency (K-F-freq), Kucera-Francis number of categories (K-F-ncats), Kucera-Francis number of samples (K-F-nsamp), Thorndike-Lorge written frequency (T-L-freql), Brown verbal frequency (Brown-freq), familiarity rating (Fam), concreteness rating (Conc), imageability rating (Imag), meaningfulness Colorado Norms (Meanc), meaningfulness Paivio Norms (Meanp), age of acquisition (AOA)

UTTERANCE TYPE FEATURES:

Ratio of commands (Command), prompts or back-channels (Prompt), questions (Question), assertions (Assertion)

PROSODIC FEATURES:

Average, minimum, maximum and standard deviation of the voice's pitch in Hz (Pitch-mean, Pitch-min, Pitch-max, Pitch-stddev) and intensity in dB (Int-mean, Int-min, Int-max, Int-stddev), voiced time (Voiced) and speech rate (Word-per-sec)

Table 2: Description of all features, with feature labels in brackets.

2.3 Statistical model

By definition, personality evaluation assesses relative differences between individuals, e.g. one per-

son is described as an extravert because the average population is not. Thus, we formulate personality recognition as a ranking problem: given two individuals' extracts, which shows more extraversion?

Personality models are trained using RankBoost, a boosting algorithm for ranking, for each Big Five trait using the observers' ratings of personality (Freund et al., 1998). RankBoost expresses the learned models as rules, which support the analysis of differences in the personality models (see section 3). Each rule modifies the conversation extract's ranking score by α whenever a feature value exceeds experimentally learned thresholds, e.g. Rule 1 of the extraversion model in Table 4 increases the score of an extract by $\alpha = 1.43$ if the speech rate is above 0.73 words per second. Models are evaluated by a ranking error function which reports the percentage of misordered pairs of conversation extracts.

3 Results

The features characterize many aspects of language production: utterance types, content and syntax (LIWC), psycholinguistic statistics (MRC), and prosody. To evaluate how each feature set contributes to the final result, we trained models with the full feature set and with each set individually. Results are summarized in Table 3. The baseline is a model ranking extracts randomly, producing a ranking error of 0.5 on average. Results are averaged over a 10 fold cross-validation.

Feature set	All	LIWC	MRC	Type	Pros
Set size	117	88	14	4	11
Extraversion	0.35•	0.36•	0.45	0.55	**0.26•**
Emot. stability	0.40	0.41	**0.39•**	0.43	0.45
Agreeableness	**0.31•**	0.32•	0.44	0.45	0.54
Conscientious.	**0.33•**	0.36•	0.41•	0.44	0.55
Intellect	0.38•	**0.37•**	0.41	0.49	0.44

• statistically significant improvement over the random ordering baseline (two-tailed paired t-test, $p < 0.05$)

Table 3: Ranking errors over a 10 fold cross-validation for different feature sets (Type=utterance type, Pros=prosody). Best models are in bold.

Paired t-tests show that models of extraversion, agreeableness, conscientiousness and intellect using all features are better than the random ordering baseline (two-tailed, $p < 0.05$)[1]. Emotional stability is the most difficult trait to model, while agreeableness and conscientiousness produce the best results, with ranking errors of 0.31 and 0.33 respectively. Table 3 shows that LIWC features perform significantly better than the baseline for all dimensions but emotional stability, while emotional stability is best predicted by MRC features. Interestingly, prosodic features are very good predictors of extraversion, with a lower ranking error than the full feature set (0.26), while utterance type features on their own never outperform the baseline.

The RankBoost rules indicate the impact of each feature on the recognition of a personality trait by the magnitude of the parameter α associated with that feature. Table 4 shows the rules with the most impact on each best model, with the associated α values. The feature labels are in Table 2. For example, the model of extraversion confirms previous findings by associating this trait with a high speech rate (Rules 1 and 4) and longer conversations (Rule 5). But many new markers emerge: extraverts speak with a high pitch (Rules 2, 6 and 7), while introverts' pitch varies a lot (Rules 15, 18 and 20). Agreeable people use longer words but shorter sentences (Rule 1 and 20), while swear words reduce the agreeableness score (Rules 12, 18 and 19). As expected, conscientious people talk a lot about their job (Rule 1), while unconscientious people swear a lot and speak loudly (Rules 19 and 20). Our models contain many additional personality cues which aren't identified through a typical correlational analysis.

4 Conclusion

We showed that personality can be recognized automatically in conversation. To our knowledge, this is the first report of experiments testing trained models on unseen subjects. There are models for each dimension that perform significantly better than the baseline. Combinations of these models may be useful to identify important personality types in different NLP applications, e.g. a combination of extraversion, emotional stability and intellect indicates leadership, while low intellect, extraversion and agreeableness are correlated with perceptions of trustworthiness.

One limitation for applications involving speech recognition is that recognition errors will introduce noise in all features except prosodic features, and prosodic features on their own are only effective in the extraversion model. However, our data set is relatively small (96 subjects) so we expect that more

[1] We also built models of self-reports of personality, but none of them significantly outperforms the baseline.

#	Extraversion with prosody	α	Emotional stability with MRC	α	Agreeableness with all	α	Conscientiousness with all	α	Intellect with LIWC	α
1	Word-per-sec $\geq$ 0.73	1.43	Nlet $\geq$ 3.28	0.53	Nphon $\geq$ 2.66	0.56	Occup $\geq$ 1.21	0.37	Colon $\geq$ 0.03	0.49
2	Pitch-mean $\geq$ 194.61	0.41	T-L-freq $\geq$ 28416	0.25	Tentat $\geq$ 2.83	0.50	Insight $\geq$ 2.15	0.36	Insight $\geq$ 1.75	0.37
3	Voiced $\geq$ 647.35	0.41	Meanc $\geq$ 384.17	0.24	Colon $\geq$ 0.03	0.41	Posfeel $\geq$ 0.30	0.30	Job $\geq$ 0.29	0.33
4	Word-per-sec $\geq$ 2.22	0.36	AOA $\geq$ 277.36	0.24	Posemo $\geq$ 2.67	0.32	Int-stddev $\geq$ 7.83	0.29	Music $\geq$ 0.18	0.32
5	Voiced $\geq$ 442.95	0.31	K-F-nsamp $\geq$ 322	0.22	Voiced $\geq$ 584	0.32	Nlet $\geq$ 3.29	0.27	Optim $\geq$ 0.19	0.24
6	Pitch-max $\geq$ 599.88	0.30	Meanp $\geq$ 654.57	0.19	Relig $\geq$ 0.43	0.27	Comm $\geq$ 1.20	0.26	Inhib $\geq$ 0.15	0.24
7	Pitch-mean $\geq$ 238.99	0.26	Conc $\geq$ 313.55	0.17	Insight $\geq$ 2.09	0.25	Nphon $\geq$ 2.66	0.25	Tentat $\geq$ 2.23	0.22
8	Int-stddev $\geq$ 6.96	0.24	K-F-ncats $\geq$ 14.08	0.15	Prompt $\geq$ 0.06	0.25	Nphon $\geq$ 2.67	0.22	Posemo $\geq$ 2.67	0.19
9	Int-max $\geq$ 85.87	0.24	Nlet $\geq$ 3.28	0.14	Comma $\geq$ 4.60	0.23	Nphon $\geq$ 2.76	0.20	Future $\geq$ 0.87	0.17
10	Voiced $\geq$ 132.35	0.23	Nphon $\geq$ 2.64	0.13	Money $\geq$ 0.38	0.20	K-F-nsamp $\geq$ 329	0.19	Certain $\geq$ 0.92	0.17
11	Pitch-max $\geq$ 636.35	-0.05	Fam $\geq$ 601.98	-0.19	Fam $\geq$ 601.61	-0.16	Swear $\geq$ 0.20	-0.18	Affect $\geq$ 5.07	-0.16
12	Pitch-slope $\geq$ 312.67	-0.06	Nphon $\geq$ 2.71	-0.19	Swear $\geq$ 0.41	-0.18	WPS $\geq$ 6.25	-0.19	Achieve $\geq$ 0.62	-0.17
13	Int-min $\geq$ 54.30	-0.06	AOA $\geq$ 308.39	-0.23	Anger $\geq$ 0.92	-0.19	Pitch-mean $\geq$ 229	-0.20	Othref $\geq$ 7.67	-0.17
14	Word-per-sec $\geq$ 1.69	-0.06	Brown-freq $\geq$ 1884	-0.25	Time $\geq$ 3.71	-0.20	Othref $\geq$ 7.64	-0.20	I $\geq$ 7.11	-0.19
15	Pitch-stddev $\geq$ 115.49	-0.06	Fam $\geq$ 601.07	-0.25	Negate $\geq$ 3.52	-0.20	Humans $\geq$ 0.83	-0.21	WPS $\geq$ 5.60	-0.20
16	Pitch-max $\geq$ 637.27	-0.06	K-F-nsamp $\geq$ 329	-0.26	Fillers $\geq$ 0.54	-0.22	Swear $\geq$ 0.93	-0.21	Social $\geq$ 10.56	-0.20
17	Pitch-slope $\geq$ 260.51	-0.12	Imag $\geq$ 333.50	-0.27	Time $\geq$ 3.69	-0.23	Swear $\geq$ 0.17	-0.24	You $\geq$ 3.57	-0.21
18	Pitch-stddev $\geq$ 118.10	-0.15	Meanp $\geq$ 642.81	-0.28	Swear $\geq$ 0.61	-0.27	Relig $\geq$ 0.32	-0.27	Incl $\geq$ 4.30	-0.33
19	Int-stddev $\geq$ 6.30	-0.18	K-F-ncats $\geq$ 14.32	-0.35	Swear $\geq$ 0.45	-0.27	Swear $\geq$ 0.65	-0.31	Physcal $\geq$ 1.79	-0.33
20	Pitch-stddev $\geq$ 119.73	-0.47	Nsyl $\geq$ 1.17	-0.63	WPS $\geq$ 6.13	-0.45	Int-max $\geq$ 86.84	-0.50	Family $\geq$ 0.08	-0.39

Table 4: Best RankBoost models for each trait. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, e.g. introversion.

training data would improve model accuracies and might also make additional features useful. In future work, we plan to integrate these models in a dialogue system to adapt the system's language generation; we will then be able to test whether the accuracies we achieve are sufficient and explore methods for improving them.

Acknowledgements

Thanks to Matthias Mehl and James Pennebaker for sharing their data.

References

E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist. 1999. Integrating models of personality and emotions into lifelike characters. In *Proc. of the International Workshop on Affect in Interactions*, p. 136–149.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

M. Coltheart. 1981. The MRC psycholinguistic database. *Quarterly J. of Experimental Psychology*, 33A:497–505.

B. Donnellan, R. D. Conger, and C. M. Bryant. 2004. The Big Five and enduring marriages. *J. of Research in Personality*, 38:481–504.

Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. 1998. An efficient boosting algorithm for combining preferences. In *Proc. of the 15th ICML*, p. 170–178.

D. Funder and C. Sneed. 1993. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *J. of Personality and Social Psychology*, 64(3):479–490.

A. Furnham, 1990. *Handbook of Language and Social Psychology*, chapter Language and Personality. Winley.

A. J. Gill and J. Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proc. of the 24th Annual Conference of the Cognitive Science Society*, p. 363–368.

F. Heylighen and J.-M. Dewaele. 2002. Variation in the contextuality of language: an empirical measure. *Context in Context, Special issue of Foundations of Science*, 7:293–340.

R. Hogan, G. J. Curphy, and J. Hogan. 1994. What we know about leadership: Effectiveness and personality. *American Psychologist*, 49(6):493–504.

O. P. John and S. Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality theory and research*. New York: Guilford Press.

J. Liscombe, J. Venditti, and J. Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proc. of Eurospeech - Interspeech 2003*, p. 725–728.

M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. In press. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *J. of Personality and Social Psychology*.

W. T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *J. of Abnormal and Social Psychology*, 66:574–583.

S. Nunn. 2005. Preventing the next terrorist attack: The theory and practice of homeland security information systems. *J. of Homeland Security and Emergency Management*, 2(3).

J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *J. of Personality and Social Psychology*, 77:1296–1312.

J. W. Pennebaker, L. E. Francis, and R. J. Booth, 2001. *LIWC: Linguistic Inquiry and Word Count*.

B. Reeves and C. Nass. 1996. *The Media Equation*. University of Chicago Press.

K. R. Scherer. 1979. Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social markers in speech*, p. 147–209. Cambridge University Press.

S. Tucker and S. Whittaker. 2004. Accessing multimodal meeting data: Systems, problems and possibilities. *Lecture Notes in Computer Science*, 3361:1–11.

M. Walker and S. Whittaker. 1990. Mixed initiative in dialogue: an investigation into discourse segmentation. In *Proc. of the 28th Annual Meeting of the ACL*, p. 70–78.