FUSION OF ACOUSTIC, LINGUISTIC AND PSYCHOLINGUISTIC FEATURES FOR SPEAKER PERSONALITY TRAITS RECOGNITION

Firoj Alam, Giuseppe Riccardi

Department of Information Engineering and Computer Science, University of Trento, Italy

ABSTRACT

Behavioral analytics is an emerging research area that aims at automatic understanding of human behavior. For the advancement of this research area, we are interested in the problem of learning the personality traits from spoken data. In this study, we investigated the contribution of different types of speech features to the automatic recognition of Speaker Personality Trait (SPT) across diverse speech corpora (broadcast news and spoken conversation). We have extracted acoustic, linguistic, and psycholinguistic features and modeled their combination as input to the classification task. For the classification, we used Sequential Minimal Optimization for Support Vector Machine (SMO) together with Relief feature selection. The present study shows different levels of performance for automatically selected feature sets, and overall improved performance with their combination across diverse corpora.

Index Terms— Affective Computing, NLP, Behavioral Signal Processing, Paralinguistic analysis in Speech.

1. INTRODUCTION

Behavior may be explained based on a person's underlying personality traits as well as the situation in context [29]. The research on personality understanding has attracted attention from several fields, the most notable of which are human-machine interaction, health diagnosis and the newly emerging field of behavioral analytics.

Recent work has shown how people's personality is expressed and how it can be predicted and applied in different contexts. In spoken interaction, a user's personality can be predicted [4], which can increase the possibility of natural interaction. Job performance can be predicted by studying a person's personality [16] (especially, conscientiousness and neuroticism dimensions). Conversational expressions in video blogs can be analyzed to understand a user's personality [19]. Research findings also show that personality is closely associated with romantic relationships [25], preference of genre of music [26], and consumer preference of brands [27].

In spoken language communication, speech signal provides important information for analyzing and modeling human behavior. This speech signal carries rich information about a variety of linguistic and paralinguistic phenomena, which is encoded with different behavioral cues [1], including emotion, intent, traits.

Understanding behavioral cues may help to make automated systems such as robots, embodied virtual agents and animated characters more human-like [28]. In this study, we are interested in understanding one of the behavioral dispositions - speaker personality traits, by analyzing acoustic, linguistic and psycholinguistic descriptors of human speech.

In [21], we studied different feature selection methods along with ensemble classification methods that can tackle the high-dimensionality and variability of the classification problem. Following our previous study, our goal here is (a) to understand the prediction capability of linguistic and psycholinguistic features in addition to acoustic features, (b) analyze the feature fusion technique to get the best prediction and c) evaluate our algorithms across different speech corpora. There are several studies that show how personality manifests in word usage [2,3,5]. This has indeed motivated us to use linguistic and psycholinguistic features in this context.

This paper is organized as follows. Section 2 describes related works. Section 3 describes the corpora used in the experiment and Section 4 defines the experimental method. Details of the classification results and discussions are given in Section 5. Conclusions are provided in Section 6.

2. LITERATURE REVIEW

This section provides an overview of the personality traits theory following a review of the related work in speech and behavioral science. Aristotle was the first to study personality, in the fourth century. Since then, psychologists have been trying to define theories and inventories by analyzing lexical terms or biological phenomena. Personality is defined as the coherent pattern of affect, behavior, cognition and desire over time and space, which are used to characterize unique individuals. Among the various theories of personality traits, Big-5 framework is the most widely used and accepted model [15]. It describes human personality as a vector of five values corresponding to bipolar traits, as defined below.

O (Openness): Artistic, curious, imaginative, etc.

C (Conscientiousness): Efficient, organized, etc.

E (Extraversion): Energetic, active, assertive, etc.

A (Agreeableness): Compassionate, cooperative etc.

N (Neuroticism): Anxious, tense, self-pitying, etc.

There are several rating instruments available for measuring each of these traits, which describe personality: (a) self-report is used to rate oneself; (b) observer-report is used to rate others. The annotation of the two corpora that we used in this study is based on these ratings.

There have been significant studies on predicting personality traits from social media by analyzing text, audio and video: Twitter [14], Facebook [22,23], blog [24] and video-blog [19]. Mairesse [5] used both conversation and essay corpora to study personality traits. Research on personality traits recognition from speech is relatively recent. Major contributions were made in the Interspeech 2012 Speaker Traits Challenge [6]. The outcome of the evaluation campaign indicates that great research effort is needed to understand feature selection and classification approaches in order to develop a better hypothesis. We also observed in our previous study [21], that more investigation is needed in order to produce a usable research outcome. This is why we decided to continue this work with the same datasets.

3. DATA

Two diverse corpora have been studied: (i) Speaker Personality Corpus (SPC), and (ii) Personable and Intelligent virtual Agents (PerSIA) corpus. Different annotation schemes (i.e., self-report, observer-report) have been used in these corpora.

The PerSIA [9] corpus is an Italian human-human spoken conversation, recorded in a simulated tourist information center. Two separate groups of people participated by playing “customer” and “agent” roles over telephone conversations. Each customer was given a tourism task to perform and the agent provided relevant answers. Out of the 24 speakers 12 were customers and 12 were agents. Personality label was assigned based on the self-report during the data collection. Out of 144 (user and agent) calls, 119 calls of the agent sub-corpus were used for the experiment. The total duration of the conversations in the corpus was 2 hours and 14 minutes. The corpus was transcribed manually, which contains ~9K tokens and ~1K token types. The minimum and maximum lengths of utterances were 14 and 283 tokens, respectively. The mean and standard deviation were 76 and 51 tokens, respectively.

The SPC Corpus was obtained from the organizers of the Interspeech 2012 Speaker Trait Challenge [6]. The audio clips of this corpus were randomly collected from the French news bulletins, broadcasted in February 2005, with a quality of 16 bits, 8kHz sample rate, and contained 1 hour and 14 minutes of recordings. Eleven annotators annotated the corpus by listening to all the audio clips using BFI-10 [13]. The annotators did not understand French; therefore, the annotation was based on paralinguistic information. One of the contributions of this study is that we transcribed the corpus manually to extract linguistic and psycholinguistic features[1]. Some descriptive statistics of the transcription are given in Table 1. The training, development and test sets consist of 256, 183 and 201 instances, respectively.

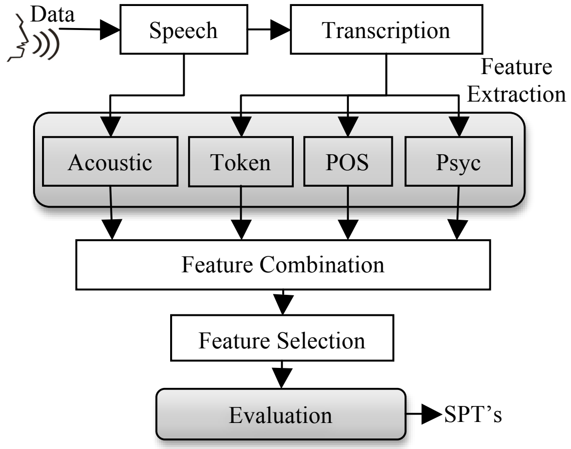
Table 1: Statistics of the SPC transcriptions in terms of tokens

Data	Min	Max	Mean	Std	Tok	T.type
Train	14	47	31.3	5.75	8023	3089
Dev	11	50	32.5	5.70	5954	2373
Test	3	51	30.8	7.58	6193	2512

4. EXPERIMENTAL DESIGN

For our experiment, we first extracted acoustic features from speech, linguistic and psycholinguistic (Psyc) features from the transcription, and then generated and evaluated models for each feature set. We then experimented with the feature fusion techniques and eventually selected a fusion technique where we combine all feature vectors into a single vector and apply feature selection followed by classification. A conceptual design of the system is given in Figure 1.

Figure 1: Conceptual design of our system for the study of SPT



4.1. Features

The following subsections discuss how different types of features were extracted.

4.1.1. Acoustic Features

These features were extracted using openSMILE [7] with the predefined configuration file (IS2012.conf), which is provided in the Interspeech-2012 Speaker Trait evaluation campaign. The low-level acoustic features were extracted with approximately 100 frames per second, with 10-30 milliseconds per frame. These low-level descriptors (LLDs) were then projected onto single scalar values by descriptive statistical functionals [10]. More detail on the acoustic features can be found in [6].

4.1.2. Linguistic Features

Bag-of-words is the most widely used approach in document categorization. It is also commonly used in behavioral signal

[1] The extension of the SPC corpus along with transcriptions will be made publicly available.

processing [20]. In this study, we studied the influence of tokens and parts-of-speech (POS) separately by first representing them into two feature vectors and then by transforming feature values with term-frequency and inverse-document-frequency (tf-idf).

4.1.3. Psycholinguistic (Psyc) Features

Pennebaker & King developed the Linguistic Inquiry Word Count (LIWC) [17] over the past few decades using human judges to design word categories for the most commonly used words. LIWC has since been used to study gender, age, personality, and health and the correlation between these attributes and word uses. There are a total of 81 word categories, a few of which are family, cognitive mechanism, affect, occupation, body, article, and function words. The LIWC analyzes language (in our case utterance) on a word-by-word basis. The system has master dictionaries for different languages. LIWC counts the words in the utterance sample that match each of the categories in the dictionary. Scores for each category are expressed as percentages, or a proportion of words that match the total number of words used. For example, if an utterance used 10 words that fall into the word category *"anger"* and the utterance contained 100 words, then the word category's score for *"anger"* would be 0.10. We used the dictionaries that are available with LIWC for Italian and French with PerSIA and SPC, respectively. Throughout this paper, we use the term *psycholinguistic features* to refer to features that were extracted using LIWC.

4.2. Feature Selection

To understand the contribution of each feature set, before feature combination phase as shown in Figure 1, we tried to reduce dimension for acoustic and token feature vectors. Dimensionality reduction has been used because of the assumption that higher dimension may decrease the performance of the classifier. For dimensionality reduction, we applied the Relief feature selection technique [18]. No feature selection has been applied for the POS and psycholinguistic feature sets. For the study of POS, we used only tags that were extracted from tokens. We used tree-tagger[2] for the PerSIA corpus and Stanford POS tagger[3] for the SPC corpus.

After evaluating each feature set, we combined the different baseline feature vectors into a single vector, which, as a result, introduced high-dimensional problems. We thus applied the same relief feature selection approach as that we used in [21], in order to avoid high variance and over-fitting. Before the feature selection phase, feature values were discretized into equal frequency bins. Therefore for different categories number of bins was different which ranged from 2-20. Eventually, all continuous valued features were transformed into discrete valued features.

4.3. Classification and Evaluation

We generated our classification models using SMO [8] with its linear kernel. SMO is an optimization technique for solving quadratic optimization problems, which arise during the training of SVM. The main reasons for choosing SMO were (a) its higher generalization capability and (b) the fact that we obtained better results using it, compared to Adaboost and Random Forest algorithms in [21]. The linear kernel was chosen in order to alleviate the problem of higher dimensions. In all of the classification settings, we used SMO's defaults parameters, whereas in our previous study we tuned those parameters.

The performance of the system was measured in terms of Weighted Average (WA) and Un-weighted Average (UA), which have recently been used in the paralinguistic tasks [6]. However, for the sake of simplicity, we show only UA in this paper.

To evaluate the performance of the SPC development set (dev), we used the SPC training set (train) to generate the model. To evaluate the performance of the SPC test set, we generated a model by combining the SPC training and development sets (training set: train + dev). In each case, performance was estimated using (15x2 cv) Leave Speaker Group Out (LSGO) cross-validation method on the training set, with macro-averaging. In macro-averaging, UA and WA were calculated for each cross validation fold and their average was computed.

For the PerSIA corpus, we used Leave One Speaker Out (LOSO) cross-validation with micro-averaging to measure the performance of the system. Micro-averaged values were calculated by first constructing a global confusion matrix from each cross-validation fold, and then by computing UA_{micro} and WA_{micro}, as shown in equations 1 and 2. Imbalance class distribution of the PerSIA corpus was the main reason for choosing micro-average.

$$\text{UA}_{\text{micro}}=\frac{1}{2}\left(\frac{\sum_{i=1}^{F} TP_i}{\sum_{i=1}^{F} TP_i+FN_i}+\frac{\sum_{i=1}^{F} TN_i}{\sum_{i=1}^{F} TN_i+FP_i}\right) \quad (1) \qquad \text{WA}_{\text{micro}}=\frac{1}{2}\left(\frac{\sum_{i=1}^{F} TP_i+TN_i}{\sum_{i=1}^{F} TP_i+FP_i+TN_i+FN_i}\right) \quad (2)$$

where i=1…F is the number of folds. TP-true positive, TN-True negative, FP-False positive, FN-False negative.

5. RESULTS AND DISCUSSION

BIG-5 personality traits binary classification models have been evaluated on both corpora. We present the performance of the system for each feature set, their combination and oracle. The results presented on the acoustic and bag-of-words (token) features are obtained after applying feature selection. We obtained the results on the combined feature set by combining the baseline feature vectors into one vector and then applying the feature selection technique. Oracle performance gives an upper bound on our model performance based on current single feature type models. It selects best label from the generated

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[3] http://nlp.stanford.edu/software/tagger.shtml

labels of different models. Oracle label is incorrect only if no model produces the reference label [30].

Classification results on the SPC dev and test sets are given in Tables 2 and 3, respectively. The feature combination provides comparable results with the state-of-the-art, even when using SMO's default parameters. The mean UA results over all Big-5 categories show that the acoustic feature set contributes most to the classification decision, whereas the psycholinguistic feature set appears to contribute the second most.

The annotation of the SPC corpus was based on paralinguistic cues (i.e., annotators did not understand the language). However, it seems that lexical-prosodic information coexists here. This means that words, perhaps salient, representing the prosodic information convey some information. Therefore, the feature sets extracted from transcription show quite improved results when combined with acoustic features.

A closer investigation was done after applying feature selection to understand which types of features are important among feature sets in different Big-5 categories. For Big-5 categories, feature selection method select different ranges of features. However, overall reduction appears to be from 35% to 62% on the SPC train + dev sets, out of ~9.5K features. Study of SPC feature sets reveals that for different Big-5 categories, the feature selection method selects and rank different types of features. For example, in the openness category, MFCC-based features appear to have higher ranking within acoustic features. Within the psycholinguistic feature set, personal pronoun, articles, social and affective categories appear in ranked order. In the POS feature set, it appears that pronouns, verbs and adverbs have greater significance, and in that order.

The results of ensemble method in our previous study [21] on the SPC dev set are comparable with the feature combination results of the same data set. The performance of the present system was improved by 2.5% and 1.1% in openness and extraversion categories, respectively, even with default parameters.

Table 2: UA results on SPC dev set using different feature sets. Tok.: token, POS: parts-of-speech, Psyc: Psycholinguistic, AC: acoustic, Comb: Feature combination with feature selection, Ora: oracle performance.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	51.6	50.0	56.1	59.1	67.7	92.1
C	55.5	54.7	65.2	74.1	73.7	96.1
E	52.0	53.7	63.4	83.6	84.1	98.4
A	52.3	46.8	54.0	64.0	64.9	97.1
N	51.3	50.0	49.7	63.5	66.3	97.9
Mean	**52.5**	**51.1**	**57.7**	**68.8**	**71.4**	**96.3**

The results of the SPC test are comparable with the baseline results presented in Interspeech 2012 paralinguistic challenge [6], noting only one difference – the baseline results were obtained using tuned parameters whereas our results are obtained using SMO's default parameters. However, our results on the SPC test set outperform the baseline results in all categories except the conscientiousness category and overall improvement is 2.1%. We performed cross-validation on the training set (train + dev) and obtained 68.1 ± 2.7 (mean ± standard deviation) in all traits, which shows the statistical variation.

Table 3: UA results on SPC test set of different feature sets.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	49.4	49.6	52.8	63.1	62.5	93.6
C	64.6	48.8	69.8	78.6	79.6	94.0
E	56.5	56.0	61.6	77.2	78.2	97.3
A	48.8	51.4	56.2	62.5	65.1	94.2
N	50.4	49.4	50.1	65.6	66.9	91.0
Mean	**53.9**	**51.0**	**58.1**	**69.4**	**70.4**	**94.0**

In the study of PerSIA corpus we obtained a similar improvement using our feature combination method. Compared to our previous study [21], we obtained an improvement in all categories except conscientiousness category and the overall improvement is 1.2%. An interesting finding here is that in the openness category, using majority voting ensemble method, we obtained 50.2, which is 2.7% better than the feature combination method.

Table 4: UA_{micro} results on PerSIA of different feature sets using LOSO cross validation.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	57.1	44.3	41.5	45.9	47.5	81.6
C	37.5	46.6	37.6	72.6	68.5	81.1
E	37.9	28.6	43.8	52.1	67.3	90.0
A	32.1	48.0	75.1	71.8	74.9	96.9
N	53.7	69.8	64.7	52.1	60.5	96.7
Mean	**43.7**	**47.5**	**52.6**	**58.9**	**63.7**	**89.2**

6. CONCLUSIONS

In this paper, we presented our contribution to the automatic recognition of speaker personality traits from speech using two different corpora – conversation and broadcast news by studying different types of feature sets. In all the experiments, we used SMO's default parameters with its linear kernel. We obtained comparable results by combining these feature sets into a single vector. Psycholinguistic features, extracted using LIWC, give better results when compared with token and POS feature sets, whereas acoustic features outperform the other feature sets. However, oracle performance suggests that there is room for improvement in the feature or decision combination approach, which our future research will address. We also plan to use automatic transcriptions in our future study.

7. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610916-SENSEI.

8. REFERENCES

[1] Shrikanth S. Narayanan and Panayiotis G. Georgiou, “Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language”, in: Proceedings of IEEE, 101:5(1203 - 1233), 2013.

[2] James W. Pennebaker, “The Secret Life of Pronouns: What Our Words Say About Us”, Bloomsbury Press, USA, 2011.

[3] Lisa A. Fast and David C. Funder, “Personality as Manifest in Word Use: Correlations With Self-Report, Acquaintance Report, and Behavior”, Journal of Personality and Social Psychology,Vol. 94, No. 2, p:334 –346, 2008.

[4] T.W. Bickmore and R.W. Picard., “Establishing and maintaining long-term human-computer relationships”, ACM Trans. Comput.-Hum. Interact., 12:293–327, June 2005.

[5] 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. Art. Intelligence Res., 30:457–500, 2007.

[6] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, B. Weiss, “The Interspeech 2012 Speaker Trait Challenge”, Proc. Interspeech 2012, ISCA, Portland, OR, USA, 2012.

[7] F. Eyben, M. Wo¨llmer, and B. Schuller, “openSMILE - The Munich Versa- tile and Fast Open-Source Audio Feature Extractor”, In Proc. ACM Multi- media. Florence, Italy, ACM, pp. 1459–1462, 2010.

[8] John C. Platt, “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines”, Microsoft Research, Technical Report, April 21, 1998.

[9] A. V. Ivanov, G. Riccardi, A. J. Spork and J. Franc. 2012. “Recognition of Personality Traits from Human Spoken Conversations”, In: Proceedings of the 12th Annual Conference of ISCA. p. 1549-1552, 2011.

[10] Björn Schuller, “Voice and Speech Analysis in Search of States and Traits”. In Computer Analysis of Human Behavior, Albert Ali Salah, Theo Gevers (eds.), Advances in Pattern Recognition, Springer, pp. 227-253, 2011.

[11] A. J. Gill and R. M. French. “Level of representation and semantic distance: Rating author personality from texts”, In Proc. of the 2nd European Cognitive Science Conference (EuroCogsci07), 2007.

[12] J. Oberlander and A. J. Gill., “Individual differences and implicit language: Personality, parts-of-speech and pervasiveness”, In Proc. of the 26th Annual Conference of the Cognitive Science Society, Chicago, IL, USA, 2004.

[13] Rammstedt, B. & John, O. P., “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. 41, pp. 203-212, 2007.

[14] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner, “Predicting Personality from Twitter”, In Proceedings of the 3rd IEEE International Conference on Social Computing, Boston, Massachusetts, USA, (pp. 149–156), 2011.

[15] G. Matthews, I. Deary, and M. Whiteman, “Personality traits”, Cambridge University Press, 2003.

[16] Barrick, Murray R.; Mount, Michael K., “The Big Five Personality Dimensions and Job Performance: A Meta-Analysis”, Personnel Psychology; Spring 1991.

[17] Pennebaker, J. W., Francis, M. E., & Booth, R. J. “Linguistic Inquiry and Word Count (LIWC): LIWC2001”. Mahwah, NJ: Erlbaum. http://www.liwc.net/.

[18] Igor Kononenko, “Estimating Attributes: Analysis and Extensions of RELIEF”, In European Conference on Machine Learning, 171-182, 1994.

[19] J.-I. Biel and D. Gatica-Perez, “The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs”, IEEE Trans. on Multimedia, Vol. 15, No. 1, pp. 41-55, Jan. 2013

[20] Björn Schuller, Ronald Müller, Manfred Lang, Gerhard Rigoll, “Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles”, Proc. Interspeech 2005, Special Session: Emotional Speech Analysis and Synthesis: Towards a Multimodal Approach, Lisbon, Portugal, pp. 805-809, 2005.

[21] Firoj Alam, Giuseppe Riccardi, “Comparative Study of Speaker Personality Traits Recognition in Conversational and Broadcast News Speech”, Interspeech-2013, Lyon, France, 2013.

[22] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, et al., “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.” PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791, 2013.

[23] Firoj Alam, Evgeny A. Stepanov, Giuseppe Riccardi, “Personality Traits Recognition on Social Network – Facebook”, WCPR (ICWSM-13), Cambridge, MA, USA, 2013.

[24] Iacobelli F, Gill AJ, Nowson S, Oberlander J, “Large scale personality classification of bloggers”, In: Proc of the 4th int conf on Affect comput and intel interaction. Springer-Verlag, pp. 568–577. 2011.

[25] T. Chamorro-Premuzic. “Personality and Romantic Relationships”, Personality and Individual Differences. Blackwell Publishing, 2007.

[26] P. Rentfrow and S. Gosling, “The do re mi’s of everyday life: The structure and personality correlates of music preferences”. Journal of Personality and Social Psychology, 84(6):1236–1256, 2003.

[27] S. Whelan and G. Davies. “Profiling consumers of own brands and national brands using human personality”. Journal of Retailing and Consumer Services, 13(6):393–402, 2006.

[28] Riccardi, G. and Hakkani-Tur D., “Grounding Emotions in Human-Machine Conversational Systems”, Lecture Notes in Computer Science, Springer-Verlag, pp. 144-154, 2005.

[29] Tomas Chamorro-Premuzic, Adrian Furnham, “Personality and Intellectual Competence”, Lawrence Erlbaum Associates, 2005.

[30] Bouvry, Pascal and Klopotek, Mieczyslaw A and Leprevost, Franck and Marciniak, Malgorzata and Mykowiecka, Agnieszka and Rybinski, Henryk, “Security and Intelligent Information Systems”, International Joint Confererence, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers, Volume: 7053, Springer, 2012.