Mitigating Gender and L1 Differences to Improve State and Trait Recognition

Guozhen An[1,2], *Rivka Levitan*[2,3]

[1]Department of Engineering Technology, Queensborough Community College (CUNY), USA
[2]Department of Computer Science, CUNY Graduate Center, USA
[3]Department of Computer and Information Science, Brooklyn College (CUNY), USA

gan@qcc.cuny.edu, rlevitan@brooklyn.cuny.edu

Abstract

Automatic detection of speaker states and traits is made more difficult by intergroup differences in how they are distributed and expressed in speech and language. In this study, we explore various deep learning architectures for incorporating demographic information into the classification task. We find that early and late fusion of demographic information both improve performance on the task of personality recognition, and a multitask learning model, which performs best, also significantly improves deception detection accuracy. Our findings establish a new state-of-the-art for personality recognition and deception detection on the CXD corpus, and suggest new best practices for mitigating intergroup differences to improve speaker state and trait recognition.

Index Terms: Personality, Deception, L1, Gender, Multi-Task Learning

1. Introduction

Speaker state and trait recognition can be defined as the task of using the speech and language signal to discover paralinguistic information about the speaker: their current *state*, such as emotion or whether they are being deceptive; or their more stable *traits*, long-term attributes such as gender or personality. These tasks are complicated by differences between groups in how states/traits are (a) distributed, and (b) realized in speech and language. For example, [1] found that lower pitch was associated with dullness in American females, but with dominance in American males. Furthermore, the feature priors themselves are distributed differently among different kinds of speakers. Gender, for example, is easily recoverable from speech and language [2, 3], and native language (L1) introduces variations in language use as well [4].

Our corpus includes English speech from female and male speakers who are native speakers of Standard American English (SAE) and Mandarin Chinese (MC). We hypothesize that intergroup differences inhibit the performance of models trained on this heterogeneous data, and that performance can therefore be improved by mitigating these differences using machine learning techniques. We evaluate our approach on two affect recognition tasks: personality recognition and deception detection.

In this study, we fuse a gender and L1 classification task with a personality classifier and evaluate various ways to combine the two tasks, either as a single network with shared layers, or by feeding gender and L1 labels into the personality classifier. We show that including gender and L1 classification as an additional task in the personality classifier improves the performance of personality recognition on the CXD corpus by 4% relative, achieving new state-of-the-art results on in this corpus and demonstrating the capacity for gender and L1 information and multi-task learning to boost personality recognition. These results hold for the deception detection task, also achieving the highest published accuracy on this corpus, and demonstrating the general utility of this approach.

The remainder of this paper is structured as follows. In Section 2, we review previous work. Description of the dataset can be found in Section 3. In Section 4, we describe the feature set and deep learning models. Section 5 presents the results from various experiments. Finally, we conclude and discuss future research directions in Section 6.

2. Related Work

Numerous studies have shown that manifestations of personality are influenced by gender and culture differences. [1] described that the use of nasal and louder voice is perceived as extraverted, and American extraverts tend to make fewer pauses while speaking. Conversely, German extraverts produce more pauses than introverts. [5] found that the “Big Five” personality traits are distributed differently across genders. For example, extraversion is more commonly seen in males, and agreeableness in females. In a study of a group of male and female, [6] found that some linguistic cues to personality vary significantly across genders e.g. males who are marked high in conscientiousness use more filler words, while females don’t.

Similarly, interpersonal differences have been cited as a major obstacle to accurate deception detection [7]. Gender and L1 in particular have been associated with differences in deception *detection* behavior [8]. However, there has been little research on using demographic information to improve deception detection. Levitan et al. [9] found that including gender, native language, and personality scores along with acoustic-prosodic features improved classification accuracy on the Columbia Cross-Cultural Deception (CXD) Corpus [10], supporting the notion that deceptive behavior varies across different groups of people, and that including information about interpersonal differences can improve the performance of a deception classifier. Similarly, in [11], we showed that jointly learning deception and personality labels improved deception detection.

The machine learning approaches in this study build on our previous work on personality recognition [12] and deception detection [11], as well as Mendels et al.’s deception detection in this corpus [13]. Here, we focus on how adding demographic information to our existing models can improve classification accuracy beyond previous results.

3. Data

The collection and design of Columbia X-Cultural Deception (CXD) Corpus analyzed here is described in more detail by [10, 14, 9]. It contains within-subject deceptive and

http://dx.doi.org/10.21437/Interspeech.2019-2868

non-deceptive English speech, collected using a fake resume paradigm, from native speakers of Standard American English (SAE) and Mandarin Chinese (MC). There are approximately 125 hours of speech in the corpus from 173 subject pairs and 346 individual speakers.

Transcripts for the recordings were obtained using Amazon Mechanical Turk[1] (AMT), and the transcripts were force-aligned using Kaldi [15]. The speech was then automatically segmented into inter-pausal units (IPUs) using Praat, and transcripts were manually corrected. The subject key-presses were aligned with the speech as well.

Before the recorded interviews, subjects filled out the NEO-FFI (Five Factor) personality inventory [16], yielding scores for Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). A median split of the Big Five scores is done to divide each of the big five personality groups into two classes, and so the task was five different binary classification tasks, one for each trait.

The unit of segmentation used for personality classification is the turn. Turn boundaries were extracted in the following manner: the manual orthographic transcription was force-aligned with the audio, and the speech was segmented if there was a silence of more than 0.5 seconds. In total, there are 29175 turn-level instances, with an average duration of 9.03s. During training and testing, all turns from a single speaker were contained within a single fold.

For deception classification, the speech was segmented in two different ways: turn and IPU. An IPU is defined as speech from a single speaker separated by at least 50 ms silence, and a turn is defined as speech from a single speaker separated by at least 500 ms silence. Segments were eliminated if their duration is less than 0.05 seconds, resulting in average durations of 1.31 and 4.24 seconds for IPUs and turns, respectively. Finally, there are 79,632 and 30,368 IPU and turn level segments respectively, totaling 110,000 instances. Including instances from both levels of segmentation significantly increased the training size.

4. Methodology

4.1. Features

For our experiments, we use the feature sets described in [17, 12]: acoustic-prosodic low-level descriptor features (**LLD**); word category features from **LIWC** (Linguistic Inquiry and Word Count) [18]; and word scores for pleasantness, activation and imagery from the Dictionary of Affect in Language (**DAL**) [19]. We also use the Gensim library [20] to extract two sets of word embedding features (**WE**) using Google's pre-trained skip-gram vectors [21] and Stanford's pre-trained GloVe vectors [22]. The feature sets used here represent information from both the acoustic and lexical signal, as well have the higher-level psycholinguistic information represented by the LIWC and DAL features.

4.2. Models

Our basic DNN model is a multilayer perceptron (MLP) [23], a simple feed-forward network using the sigmoid activation function. We experiment with the following fusion methods:

Early fusion. Gender and L1 are concatenated with the original feature vectors (Figure 1).

Late fusion. The feature vectors are fed through several fully-connected layers with the sigmoid activation function. Gender and L1 are concatenated with the final hidden layer and fed forward to the output layer for prediction (Figure 2).

Multi-task learning. Gender and L1 are included as *output* nodes that share hidden layers with the state/trait classification task (Figure 3). In a late fusion variant of this approach, once the demographic labels have been predicted, they are concatenated with the final hidden layer and fed into the state/trait prediction layer (Figure 4). It is hypothesized that robustness and accuracy may be improved by giving the classifier more than one task, since the tasks can influence each other through a shared representation.

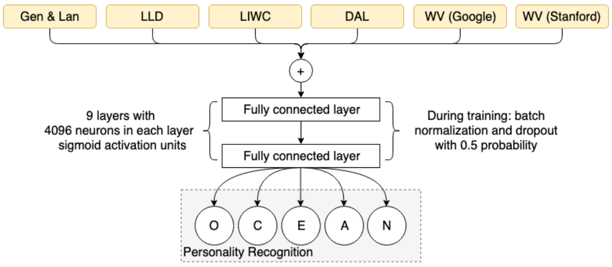

Figure 1: *Early fusion model.*

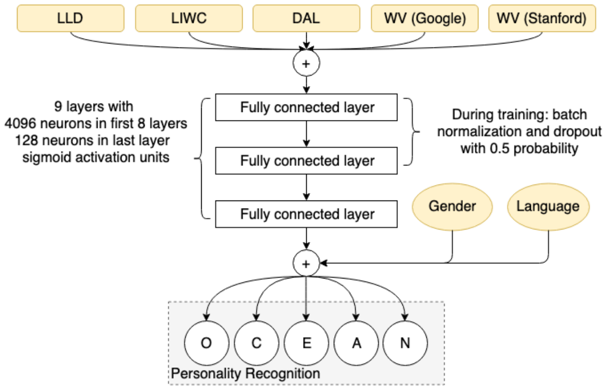

Figure 2: *Late fusion model.*

Details about each network topology are included in the figures. Tuning the learning rate improved performance, and the final model used $\alpha = 0.001$ with 100 epochs. Adam optimizer was used during training, and we used mean squared error as the loss function (all losses were averaged in the multi-task learning setting). During training, we add batch normalization [24] and a dropout layer [25] with 0.5 probability to each hidden layer.

Figures 1- 4 show the networks used for personality classification. We also implemented the MTL-Late fusion model for deception detection, which looked like Figure 4, with the addition of an output layer for deception after the personality prediction. The shared hidden layers were used to predict gender and language. The demographic labels were then concatenated with the final hidden layer to predict personality. The personality labels were then concatenated with the demographic labels and final hidden layer and used to predict deception.

5. Results

All models were trained using the Adam optimizer [26] with learning rate 0.001, decreasing at a rate of 50% for 100 epochs.

[1] https://www.mturk.com

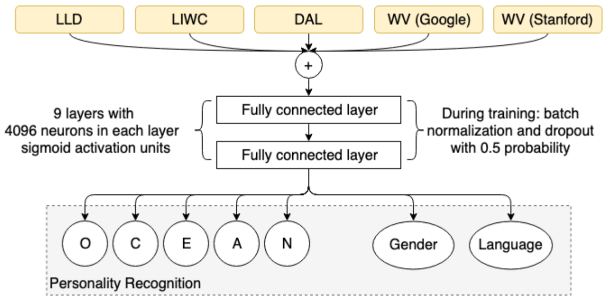

Figure 3: *Multi-task learning model (MTL).*

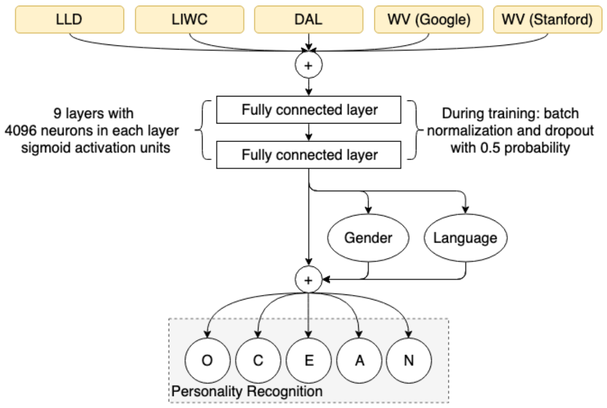

Figure 4: *Multi-task learning + late fusion model (MTL-Late).*

We use the same data setting as [12] for the personality experiment: We split data to 90%/10% train/test, which results in approximately 40 speakers in the test set. The data setting for the deception experiment is same as [11]: Our data was split into train and test sets of 88,000 and 22,000 samples respectively, and it was also split based on speaker id, so no speaker appears in both train and test. We compare to a baseline model which contains no demographic information. For personality recognition, this corresponds to Figure 1 without the demographic feature set. For deception detection, this is the model described in [11], which is similar but includes personality recognition as an additional task.

5.1. Varying amounts of in-group training data

To estimate how much improvement is likely to be obtained from mitigating intergroup variability, we trained multiple models for personality detection, varying the percentage of female or male training data in each one, and tested them on only female or only male speech. We hypothesized that the models with more in-group training data would perform better. Comparing the in-group to the out-group models can bring empirical support to the intuition that personality recognition is impaired by intergroup variability.

Results are shown in Figure 5. They do not show straightforward linear improvement with increasing in-group data, but most of the lines do show an upward trajectory. The average F-measure with 0% in-group data is 0.59, and the average F-measure with 100% in-group data is 0.64: on average, a 4.44% improvement. These results support our hypothesis that there is room for performance improvement through mitigation of intergroup differences.

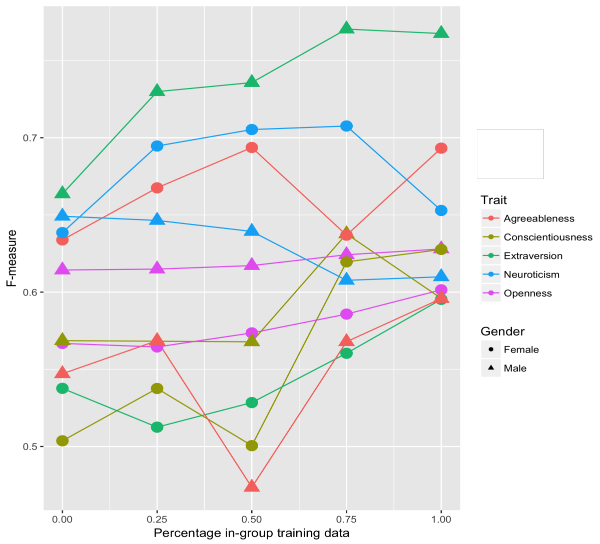

Figure 5: *Personality recognition models with varying amounts of in-group training data.*

5.2. Results with demographic information

Table 1: *Personality Recognition Results*

Model	O	C	E	A	N	Avg
Baseline	.56	.53	.59	.64	.58	.58
Early Fusion	.58	.56	.55	.66	.59	.59
Late Fusion	.58	.62	**.62**	.64	.59	.61
MTL	**.61**	**.68**	.58	**.67**	**.68**	.64
MTL-Late	**.61**	.67	**.62**	.65	**.68**	**.65**

Table 1 shows the results for personality recognition. We first evaluate the basic deep learning structure without feeding any gender and L1 information, and the resulting average accuracy (over all personality traits) is 0.58. We then train both early fusion and late fusion models by feeding gender and L1 labels as features, improving average F1 to 0.59 and 0.61, respectively. Treating gender and L1 prediction as additional tasks improves performance on all traits except Extraversion, for an average accuracy of 0.64. The MTL-Late hybrid model yields high performance on all traits for an average accuracy of 0.65, the highest reported average accuracy on this corpus. This is the case in spite of the fact that the gender and L1 classification accuracies are imperfect, though high (92% and 84%, respectively).

From these results, we can observe that adding L1 and gender information improves the personality recognition accuracy in almost every case. The late fusion model generally outperforms the early fusion model, since gender and L1 information can make the biggest impact in late fusion. Finally, the multi-task learning model with shared layers performs best among the various models. Therefore, we can conclude that gender and L1 information generally can help personality recognition, and multi-task learning is the most effective model for incorporating this information.

Table 2: *Deception Detection Result*

Model	Precision	Recall	F1
Baseline [11]	74.37	74.67	74.38
GLP-MTL	76.53	76.71	76.58

Since the MTL-Late model performed best for personality classification, we implemented it for deception detection. Table 2 shows that including gender and L1 classification tasks improves the performance of the baseline deception classification, from 74.38 to 76.58 on F1, the highest performance reported so far on this corpus. This supports the hypothesis that our approach is generally useful and that demographic information can improve deception as well as personality classification, suggesting that it can be helpful for other tasks as well.

6. Conclusion

We compare several approaches to integrating demographic information into personality and deception classification tasks and find that both early and late fusion approaches improve performance, and multi-task learning performs best, improving performance on personality and deception detection to a new state-of-the-art for this corpus: 65% and 76.58%, respectively.

One question left open by our analysis is whether the improved performance in the MTL condition is due to gender or L1 specifically, or perhaps to the nature of multi-task learning regardless of intergroup information. We do not build models incorporating either demographic label separately, or include tasks unrelated to intergroup variability, such as, for example, automatic speech recognition. We will explore this question in future work, as well as the extension of these findings to other datasets and other related classification problems.

7. Acknowledgements

We thank Julia Hirschberg, Sarah Ita Levitan, and Andrew Rosenberg.

8. References

[1] K. R. Scherer, *Personality markers in speech*. Cambridge University Press, 1979.

[2] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” *Text-The Hague Then Amsterdam Then Berlin-*, vol. 23, no. 3, pp. 321–346, 2003.

[3] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, “Robust gmm based gender classification using pitch and rasta-plp parameters of speech,” in *Machine Learning and Cybernetics, 2006 International Conference on*. IEEE, 2006, pp. 3376–3379.

[4] J. Tetreault, D. Blanchard, and A. Cahill, “A report on the first native language identification shared task,” in *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, 2013, pp. 48–57.

[5] A. J. Stewart and C. McDermott, “Gender in psychology,” *Annu. Rev. Psychol.*, vol. 55, pp. 519–544, 2004.

[6] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, “Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life.” *Journal of personality and social psychology*, vol. 90, no. 5, p. 862, 2006.

[7] A. Vrij, P. A. Granhag, and S. Porter, “Pitfalls and opportunities in nonverbal and verbal lie detection,” *Psychological Science in the Public Interest*, vol. 11, no. 3, pp. 89–121, 2010.

[8] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, “Personality factors in human deception detection: comparing human to machine performance.” in *INTERSPEECH*, 2006.

[9] S. I. Levitan, Y. Levitan, G. An, M. Levine, R. Levitan, A. Rosenberg, and J. Hirschberg, “Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection,” in *Proceedings of NAACL-HLT*, 2016, pp. 40–44.

[10] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg, “Individual differences in deception and deception detection,” 2015.

[11] G. An, S. I. Levitan, J. Hirschberg, and R. Levitan, “Deep personality recognition for deception detection,” *Proc. Interspeech 2018*, pp. 421–425, 2018.

[12] G. An and R. Levitan, “Lexical and acoustic deep learning model for personality recognition,” *Proc. Interspeech 2018*, pp. 1761–1765, 2018.

[13] G. Mendels, S. I. Levitan, K.-Z. Lee, and J. Hirschberg, “Hybrid acoustic-lexical deep learning approach for deception detection,” *Proc. Interspeech 2017*, pp. 1472–1476, 2017.

[14] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, “Cross-cultural production and detection of deception from speech,” in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 1–8.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[16] P. T. Costa and R. R. MacCrae, *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources, 1992.

[17] G. An, S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine, and J. Hirschberg, “Automatically classifying self-rated personality scores from speech,” *Interspeech 2016*, pp. 1412–1416, 2016.

[18] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, 2001.

[19] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, “A dictionary of affect in language: Iv. reliability, validity, and applications,” *Perceptual and Motor Skills*, vol. 62, no. 3, pp. 875–888, 1986.

[20] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[22] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation.” in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[23] M. W. Gardner and S. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” *Atmospheric environment*, vol. 32, no. 14, pp. 2627–2636, 1998.

[24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*, 2015, pp. 448–456.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.