Speaking Style Based Apparent Personality Recognition

Jianguo Yu[1](✉), Konstantin Markov[1](✉), and Alexey Karpov[2](✉)

[1] The University of Aizu, Fukushima, Japan
{d8182103,markov}@u-aizu.ac.jp
[2] SPIIRAS, St. Petersburg, Russia
karpov@iias.spb.su

Abstract. In this study, we investigate the problem of apparent personality recognition using person's voice, or more precisely, the way he or she speaks. Based on the style transfer idea in deep neural net image processing, we developed a system capable of speaking style extraction from recorded speech utterances, which then uses this information to estimate the so called Big-Five personality traits. The latent speaking style space is represented by the Gram matrix of convoluted acoustic features. We used a database with labels of personality traits perceived by other people (first impression). The experimental results showed that the proposed system achieves state of the art results for the task of audio based apparent personality recognition.

Keywords: Automatic Apparent Personality Recognition · First impression prediction · Speaking style representation · Computational Paralinguistics

1 Introduction and Related Works

The interest for Automatic Personality Recognition (APR) has rapidly risen in recent years as it has many important applications [28], such as products, jobs, or services recommendation [8,23], mental health diagnosis [6], computer-assisted tutoring systems [29], social network analysis [2], etc. But since it is very difficult to infer a person's true personality, many researchers started to pay attention to a less complex problem instead: Automatic Apparent Personality Recognition (AAPR), which is the personality perceived by other people (first impression). AAPR also has many practical applications since people constantly estimate other persons personality. For example, if the interviewer's first impression on the job candidate is bad, he has lower chance to get the job; The audiences' first impression on a YouTuber's voice also influences whether they continue watching or close the video.

1.1 The Big-Five Model

The personality, as well as apparent personality, are formally described by five dimensions known as the Big-Five personality traits [19]:

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 540–548, 2019.
https://doi.org/10.1007/978-3-030-26061-3_55

- **EXT**raversion vs. Introversion (sociable, assertive, playful vs. aloof, reserved, shy).
- **NEU**roticism vs. Emotional stability (calm, unemotional vs. insecure, anxious).
- **AGR**eeableness vs. Disagreeable (friendly, cooperative vs. antagonistic, fault-finding).
- **CON**scientiousness vs. Unconscientious (self-disciplined, organized vs. inefficient, care-less).
- **OPE**ness to experience (intellectual, insightful vs. shallow, unimaginative).

For personality recognition, the true labels are usually obtained by self-assessment questionnaire [7], where people rate their own behavior with Likert scales [1]. While for the apparent personality recognition, the labels are obtained by other people's first impression [9].

1.2 Audio Based AAPR

The personality traits can be inferred based on many types of observations, such as text [17,18,30], audio [20,24], video [22,31], or any combination of them, each of which has its own applications, depending on the availability of observations in different situations. For example, the audio based AAPR is very useful for the producers who make education or explainer videos since the audiences' first impression on their voices can largely affect the trustiness and attractiveness of the videos.

The conventional methods of AAPR from audio typically use a large pool of potentially prosody features (e.g. Mel Frequency Cepstral Coefficients, pitch, energy, and their 1st/2nd order temporal derivatives) and "Interspeech 2012 Speaker Trait Challenge" [26] is the first, rigorous comparison of different approaches over the same data and using the same experimental protocol for audio based AAPR, where the performances of most approaches depend heavily on careful feature selection [3,13,21,25]. Many of those features are included in the open-source openSMILE tool [10] and can serve as baseline for audio based AAPR. For example, the winner in the ChaLearn 2017 Job Candidate Screening Competition also used the openSMILE feature configuration that served as challenge baseline in the INTERSPEECH 2013 Computational Paralinguistics Challenge, which is 6373-dimensional feature set and was found to be the most effective acoustic feature set among others for personality trait recognition [12]. In order to learn useful features automatically, deep learning based methods have also been proposed for audio based AAPR. The audio model baseline provided by the organizer is a variant of the original ResNet18 model [9], which was trained on random 3s crops of the audio data and tested on the entire audio data. However, since the general network architecture is not specifically designed for AAPR, it doesn't appear to clearly outperform the conventional methods.

1.3 Neural Style Transfer

The neural style transfer became popular after the paper [11], where the style representation of an image is described as the correlation between different filter responses given by the Gram matrix. The basic idea was developed to classify image style in work [4], where the VGG-19 network [27] trained on the ImageNet dataset was used to obtain filter responses at different layers whose Gram matrix is calculated and transformed into a style vector, which is then classified by an SVM (support vector machine) classifier.

But the characteristics of audio signals are different from those of the images, e.g. speech is a sequential signal while the image is a 3D-tensor, and the duration varies for different utterances. Moreover, the Gram matrix representing styles is usually calculated from pre-trained networks and might not hold the best features for the desired task. In this work, we propose a system that automatically captures speaking styles for apparent personality recognition.

2 System Description

The proposed system evaluates a speech signal and returns 6 scores for the 5 personality traits and an interview variable (whether a candidate will be invited for a job interview).

In our neural network, the Gram matrix is not calculated from any pre-trained networks. Everything is jointly learned from scratch. The overall architecture is illustrated in Fig. 1.

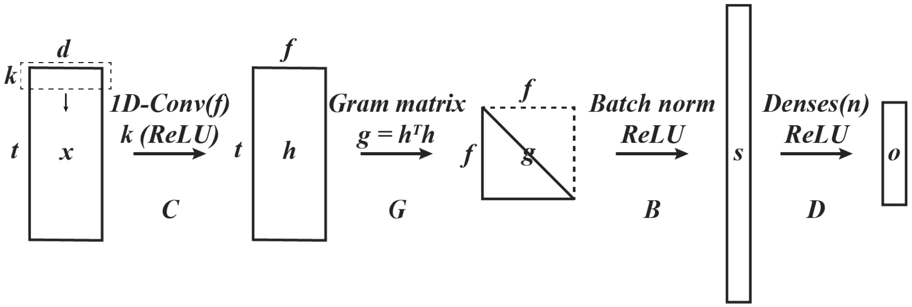

Fig. 1. Neural network architecture used in our system.

- **Input:** the input $x \in R^{t \times d}$ to our network contains d-dimensional speech features obtained at t timesteps.
- **Target:** the learning target $t \in R^6$ is a 6-dimensional vector (representing five traits and the interview variable), whose range is [0,1].
- **Convolutional layer:** the input x is first fed to a convolutional layer with f number of filters, $k \times d$ kernel size, 1 stride, and "same" zero padding, resulting in a feature map $h \in R^{t \times f}$. This is intended to automatically filter

out the silence and extract useful features for computing the speaking styles. A Rectified Linear Unit (ReLU) activation function is then applied to introduce non-linearity.

- **Gram layer:** Gram matrix g is then calculated from the feature map h, where $g = h^T h$. The lower (or upper) triangular matrix and diagonal are flattened into a vector $g* \in R^{(f+1)*f/2}$ for the next layer. A Gram layer actually represents the speaking styles as the correlations between different channels of the feature maps from the previous convolutional layer.
- **Batch norm layer:** since the norms of values in $g*$ are very big, a batch normalization layer with a ReLU activation function is added to solve this issue, resulting in a vector s that represents the speaking styles.
- **Fully connected layers:** the style vector s is then fed to one or more fully connected layers (dense layers) with ReLU activation function that further transforms s to higher level features.
- **Output layer:** finally, an output layer without activation function follows the dense layer(s) to produce an output o with 6 dimensions.
- **Loss function:** We tackle this task as a regression problem, so the mean squared error (MSE) is used as loss function.

3 Experiments and Results

3.1 Dataset

The dataset used in our experiments was the first impressions data set (CVPR 2017) [9], which comprises of 10,000 clips (with an average duration of 15s) extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing a camera and speaking in English. People in videos have different gender, age, nationality, and ethnicity. Each clip is labeled for the Big Five personality traits scores along with an interview variable score that recommends whether a job candidate should be invited for an interview or not.

The train/val/test split used by the CVPR 2017 workshop participants is 6000/2000/2000 and we followed the same protocol (The numbers in parenthesis are the actual number of examples used in our experiments due to data corruption): train the networks on the trainset (5992), tune the networks using validation set (2000) to find the best hyper-parameters, with which the networks are retrained on both train and validation sets (7992), and finally test on the testset (1997).

For each of the five traits and the interview variable, the performance was evaluated by the Mean Absolute Error (MAE) subtracted from 1, which is formulated as follows:

$$E = 1 - \frac{\sum_{i=1}^{N} |target_i - predicted_i|}{N} \tag{1}$$

The score varies between 0 (worst case) and 1 (best case).

3.2 Low Level Feature Extraction

16kHz audio signals are extracted from the video clips and 13 dimensional Mel frequency cepstral coefficients (MFCCs) are computed every 10 ms over a 25 ms window, along with their first and second derivatives and used for our acoustic feature vector $x \in R^{1528 \times 39}$, where 1528 is the number of timesteps.

3.3 Overall Settings

In all the networks to be trained, every hidden dense layer has 512 nodes and is followed by a dropout layer with a drop rate of 40%. The kernel size of every convolutional layer is 3. Each network was trained by 300 epochs using Adam [16] update method with a learning rate of 1e-4 and a batch size of 16. We chose 300 epochs because the networks after 300 epochs perform fairly well on the validation set. The L2 regularization with a rate of 1e-4 is also added to the final loss, which is $10^{-4} \sum(\|\theta\|^2)/2$ and θ is the weights vector of a layer.

3.4 Our Baseline

In order to verify whether the performance improvement is provided by the speaking styles captured by the Gram matrix, we also trained networks without it. We tried recurrent networks with GRU (Gated Recurrent Unit) cell [5] and found they are not as good as convolutional networks for this task. The networks with max pooling layer or more than one convolutional layers didn't show improvement either. We found the best network architecture without Gram layer is the network with one 1D-convolutional layer, one average pooling layer over all timesteps, one dense layer, and the output layer.

3.5 Results and Discussion

The experimental results of testset in terms of 1-MAE are summarized in Table 1. The column "System" denotes different DNN configurations. Thus, C(32) stands for a convolutional layer with 32 filters, P - an average pooling layer over all timesteps, B - a batch normalization layer with ReLU activation and D - a dense layer (2D means 2 consecutive dense layers).

Because it is hard to keep the numbers of parameters in the baseline and proposed architectures the same, we tried many hyper-parameter combinations and found that C(32)+P+2D was the best one among architectures without speaking styles. From the results, we can see that batch normalization layer didn't show any improvement in these cases and could not outperform the ResNet18. However, when the Gram layer along with a batch normalization layer is used, all configurations shows significant performance increase with the C(128)+G+B+2D achieving the best audio based AAPR results.

Table 2 shows the Big-Five traits and the interview score classification results. The ground truth labels and the system predictions were binarized based on the training set mean scores. If a given score is above the corresponding mean, the label or the prediction is considered positive, otherwise - negative. The accuracy results also show that our proposed architecture brings significant improvements for both the personality traits and interview variable.

We also noticed that the Gram layer cannot be jointly trained without a batch normalization layer (e.g. C(32)+G+D didn't converge). The reason might be that the values of the Gram matrix are changing dramatically for each batch when the Gram matrix is not calculated from the pre-trained (fixed) convolutional layer, but from a convolutional layer that is also being trained.

Table 1. $1 - MAE$ results. OPE: openness to experience. CON: conscientiousness. EXT: extroversion. AGR: agreeableness. NEU: (non-)neuroticism. Inter: interview invite variable. Ave: the average score of 5 traits (interview variable is not included).

System	Ave	OPE	CON	EXT	AGR	NEU	Inter
Published Results							
ResNet18 [9]	0.9004	0.9024	0.8966	0.8994	0.9034	0.9000	0.9032
OS_IS13 [14]	0.8996	0.9022	0.8919	0.8980	0.9065	0.8991	0.8999
Models without Speaking Style							
C(256)+B+P+D	0.8996	0.9017	0.8981	0.8980	0.9034	0.8968	0.9013
C(32)+B+P+2D	0.8999	0.9021	0.8970	0.8984	0.9038	0.8981	0.9017
C(32)+P+2D	0.9004	0.9023	0.8964	0.9005	0.9047	0.8983	0.9020
C(128)+P+2D	0.8993	0.9027	0.8948	0.8983	0.9040	0.8967	0.9013
C(256)+P+2D	0.9001	0.9022	0.8967	0.8994	0.9043	0.8979	0.9022
Models with Speaking Style							
C(32)+G+B+D	0.9013	0.9025	0.9008	0.9004	0.9035	0.8993	0.9044
C(128)+G+B+D	0.9050	0.9055	0.9054	0.9040	0.9063	0.9038	0.9083
C(256)+G+B+D	0.9053	0.9058	0.9055	0.9049	0.9068	0.9037	0.9078
C(128)+G+B+2D	**0.9061**	0.9062	0.9072	0.9049	0.9073	0.9049	**0.9101**

Table 2. Big five traits and interview variable F1 score results for different systems.

System	Ave	OPE	CON	EXT	AGR	NEU	Inter
Published Results							
OS_IS13 [15]	67.93	-	-	-	-	-	69.25
Models without Speaking Style							
C(32)+P+2D	68.35	70.15	69.90	68.50	64.79	68.40	69.30
Models with Speaking Style							
C(128)+G+B+2D	**70.92**	70.45	74.16	70.50	66.44	73.05	**72.20**

4 Conclusion and Future Work

In this work, we developed a convolutional neural network with the Gram matrix that is intended to capture the speaking styles for audio based AAPR.

The proposed architecture can learn to capture the speaking styles end-to-end and the experimental results showed that the idea of style capturing also works in the audio domain. The correlation between different dimensions of a speech signal can help to infer the personality traits and interview variable and our proposed system C(128)+G+B+2D achieves the state of the art results for audio based AAPR: the average score of five traits is 0.9061 and the interview variable score is 0.9101.

In future work, we plan to apply this technique on other modalities (e.g. text, video) and merge it with generative adversarial networks (GANs) to generate the voice with particular personality traits scores.

References

1. Boyle, G., Helmes, E.: Methods of personality assessment. In: The Cambridge Handbook of Personality Psychology, p. 110. Cambridge University Press, Cambridge (2009)
2. Celli, F., Rossi, L.: The role of emotional stability in twitter conversations. In: Proceedings of the Workshop on Semantic Analysis in Social Media, pp. 10–17. Association for Computational Linguistics (2012)
3. Chastagnol, C., Devillers, L.: Personality traits detection using a parallelized modified SFFS algorithm. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
4. Chu, W.T., Wu, Y.L.: Image style classification based on learnt deep correlation features. IEEE Trans. Multimed. **20**(9), 2491–2502 (2018)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
6. Cohen, A.S., Elvevåg, B.: Automated computerized analysis of speech in psychiatric disorders. Curr. Opin. Psychiatry **27**(3), 203 (2014)
7. Costa Jr., P.T., McCrae, R.R.: Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. J. Pers. Assess. **64**(1), 21–50 (1995)
8. Denscombe, M.: The Good Research Guide: For Small-Scale Social Research Projects. McGraw-Hill Education, London (2014)
9. Escalante, H.J., et al.: Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. arXiv preprint arXiv:1802.00745 (2018)
10. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM (2010)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)

12. Gürpinar, F., Kaya, H., Salah, A.A.: Multimodal fusion of audio, scene, and face features for first impression estimation. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 43–48. IEEE (2016)
13. Ivanov, A., Chen, X.: Modulation spectrum analysis for speaker personality trait recognition. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
14. Kaya, H., Gürpınar, F., Salah, A.A.: Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9 (2017)
15. Kaya, H., Salah, A.A.: Multimodal personality trait analysis for explainable modeling of job interview decisions. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 255–275. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4
16. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007)
18. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
19. Matthews, G., Deary, I.J., Whiteman, M.C.: Personality Traits. Cambridge University Press, New York (2003)
20. Mohammadi, G., Vinciarelli, A.: Automatic personality perception: prediction of trait attribution based on prosodic features. IEEE Trans. Affect. Comput. **3**(3), 273–284 (2012)
21. Montacié, C., Caraty, M.J.: Pitch and intonation contribution to speakers' traits classification. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
22. Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., Zancanaro, M.: Multimodal recognition of personality traits in social interactions. In: Proceedings of the 10th International Conference on Multimodal Interfaces, pp. 53–60. ACM (2008)
23. Piwek, L., Ellis, D.A., Andrews, S., Joinson, A.: The rise of consumer health wearables: promises and barriers. PLoS Med. **13**(2), e1001953 (2016)
24. Polzehl, T., Moller, S., Metze, F.: Automatically assessing personality from speech. In: 2010 IEEE Fourth International Conference on Semantic Computing, pp. 134–140. IEEE (2010)
25. Sanchez, M.H., Lawson, A., Vergyri, D., Bratt, H.: Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
26. Schuller, B., et al.: The interspeech 2012 speaker trait challenge. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Trans. Affect. Comput. **5**(3), 273–291 (2014)

29. Vinciarelli, A., et al.: Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans. Affect. Comput. **3**(1), 69–87 (2012)
30. Yu, J., Markov, K.: Deep learning based personality recognition from facebook status updates. In: 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), pp. 383–387. IEEE (2017)
31. Zhang, C.L., Zhang, H., Wei, X.S., Wu, J.: Deep bimodal regression for apparent personality analysis. In: European Conference on Computer Vision, pp. 311–324. Springer (2016). https://doi.org/10.1007/978-3-319-49409-8_25