論文 / 著書情報
Article / Book Information

論題	
Title	Personality Recognition on Dyadic Interactions with Representation Learning
著者	Nah Nathania, 越仲 孝文, 篠田 浩一
Authors	Nathania Nah, Takafumi Koshinaka, Koichi Shinoda
出典	電子情報通信学会技術研究報告, vol. 122, no. 389, pp. 241-246
Citation	IEICE technical report, vol. 122, no. 389, pp. 241-246
発行日 / Pub. date	2023, 2
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright(c) 2023 IEICE

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2022-117,SIP2022-161,SP2022-81(2023-03)

Personality Recognition on Dyadic Interactions with Representation Learning

Nathania NAH † Takafumi KOSHINAKA ‡ and Koichi SHINODA †

† Tokyo Institute of Technology 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
‡ Yokohama City University 22-2 Seto, Kanazawa-ku, Yokohama, Kanagawa, 236-0027 Japan
E-mail: nathania@ks.c.titech.ac.jp, koshinak@yokohama-cu.ac.jp, shinoda@c.titech.ac.jp

Abstract Personality computing explores methods of automatically measuring human traits to create a better understanding of the human psyche and thought processes. We examine conversations and interactions in dyadic environments through the perspective of representation learning to capture the psychological traits that compose a target's personality profile. We propose a bimodal speech-text model to predict scores for personality traits at a sentence level for the speakers using disentangled representations on speech and text. Our model outperforms current personality prediction methods using visual features and/or metadata on the UDIVA dataset's English subset.

Keywords Personality recognition, disentanglement representation learning

1. Introduction

Personality computing is a field that connects computer science to personality psychology. Understanding how different people think and act can often be achieved by understanding their personalities [1]. Personality psychology explains why people think and behave in a certain way [2]. It provides an explanatory account of an individual's thoughts, feelings, motivations, and behaviors and their patterning [3].

With personality computing, one of the goals is to develop techniques for artificial intelligence to detect, recognize, and predict human emotions and be able to respond and adapt to them [4]. Advancements within this field can not only improve computational systems but ultimately help further our understanding of human psychology and behavior [5, 6].

In this work, we explore the automatic inference of personality using audio data recorded during face-to-face dyadic interactions using a multimodal method that is effective for emotion recognition. Inspired by the use of autoencoders for disentanglement representation learning, we propose an audio-textual multimodal method for personality score prediction using the UDIVA dataset [7]. Current methods of personality recognition using this dataset propose methods using visual features and/or metadata [8]. However, our approach focuses on using just the audio recorded from these sessions to perform our personality analysis. Doing so not only allows us to reduce the amount of overhead for data collection and processing but also prevents more confounding variables and biases that are known to be associated with visual modalities.

Our work is the first to our knowledge to evaluate speech and text multimodal personality recognition on the UDIVA dataset as well as to perform self-reported personality recognition with a cross-representation autoencoder. We introduce an automatic speech recognition (ASR) component which makes it possible to perform text personality analysis solely with audio from the dataset.

2. Related Works

2.1. Emotion Recognition

This work is inspired by a multimodal audio-textual system [9] which uses data captured in dyadic interactions to perform emotion recognition. Both emotion recognition and personality recognition are tasks within the scope of affective computing and pertain to the automatic evaluation of human affect. As such, we found that a similar approach was effective for an adjacent task in the field.

We aim to capture similar features in the speech data to evaluate its performance from emotion recognition to personality recognition. However, our multimodal system regresses personality scores for each input. We also introduce an ASR component to our system to automatically generate text transcriptions from the speech audio samples and a new weighting policy for the fusion of both modalities.

2.2. Personality Modeling

The Big Five Model of Personality (also known as the Five-Factor Model or OCEAN Model) is one of the most influential models in psychology [10]. The traits present in the Big Five represent those dimensions that have shown up most frequently throughout questionnaires and lexicon

of trait-descriptive items.

The Big Five traits are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. With this trait model, personality can be numerically represented by assigning a value to each dimension on a scale. This personality model and other similar five-factor solutions (e.g., the AB5C model) [2] have been found to be more stable than more complex solutions [10].

For this work, we examine dyadic interactions to perform our personality analysis. These types of interactions are interactions between two parties, which are often used as a source for context-rich scenarios to measure interpersonal constructs and social behaviors [7]. Research in this field often leverages dyadic interactions for detecting and modeling individual and interpersonal social signals and dynamics [1].

2.3. Personality Recognition

The goal of automatic personality recognition is to identify various personality-relevant information in a subject to provide a useful personality assessment [11]. This has been accomplished in previous works by focusing on a single point of view, relying on handcrafted features such as facial landmarks, gaze, and head and body gestures [8]. However, in this work, we aim to perform personality recognition without using such features. Previous studies have shown that apparent personality perception can suffer heavily from bias associated with perceived attributes like gender, ethnicity, age, and face attractiveness [12,13].

It can be time-consuming and exhaustive work to record, transcribe, and synchronize transcriptions in data collection. Furthermore, several methods applied to our chosen UDIVA dataset use handcrafted features that require labor-intensive annotations [7]. In our work, we propose a multimodal method that only requires an audio stream as input to perform personality recognition.

3. Methodology

Our model contains a bimodal structure to perform personality recognition using a separate “Speech” audio modality and a “Text” language modality, as depicted in Figure 1. We develop a model for each modality that predicts personality scores over the five personality traits presented in Big Five model of personality and then combine the results to get our final output. Our speech model uses the audio from the dataset to train an auto-encoder to generate representations disentangled with speaker features, which is used to estimate personality scores, described in Section 3.2. Likewise, our text model uses a CNN to generate embeddings for scores, described in Section 3.3. We fuse these two scores from each model to generate final predictions using our method explained in Section 3.4.

The system uses as inputs audio recordings of each session, text transcriptions of those audio recordings and outputs labels containing the personality scores of the participant for each sample. For the speech model, we first extract wav2vec2.0 features, a Mel-spectrogram, speaker identity embeddings, and a phone sequence from each sample and then use them as input to get the predicted personality scores.

For the text model, we input text features extracted from the transcripts using Transformer-based language models to get predictions for their personality scores. Finally, we combine these two scores with a weighted score fusion to obtain the final personality score predictions.

3.1. Data Preprocessing

We first undergo data preprocessing to obtain speech audio, text transcriptions, and personality labels for each session recording to divide the data based on every `turn' spoken in the recording.

For input into our system, we take the video recordings of the UDIVA dataset and extract the audio from each session. From there, we separate the audio based on each spoken turn and participant. For each speech sample, we generate a label containing the speaker's personality scores obtained from the metadata information. We also use wav2vec2.0 [16] to automatically generate transcripts for each sample. The wav2vec2.0 model used in this component is a pre-trained language model released by huggingface which has been fine-tuned for speech recognition in English [19].

3.2. Audio Modality

The Encoder takes the wav2vec2.0 features and speaker identity embeddings as input and outputs a speech representation with personality information. This representation is then input into the Decoder along with the phone sequence embeddings generated by the phone encoder and speaker identity embeddings to reconstruct the Mel-frequency spectrograms.

We minimize the loss between the reconstructed spectrograms and the original ones generated by the speech samples to leverage both high-level and low-level features of the speech data. Finally, we disentangle features from the speaker identity and phone sequences irrelevant to personality recognition by reducing the size of the encoder's bottleneck. This allows us to generate speech representations with personality cues that are uninfluenced

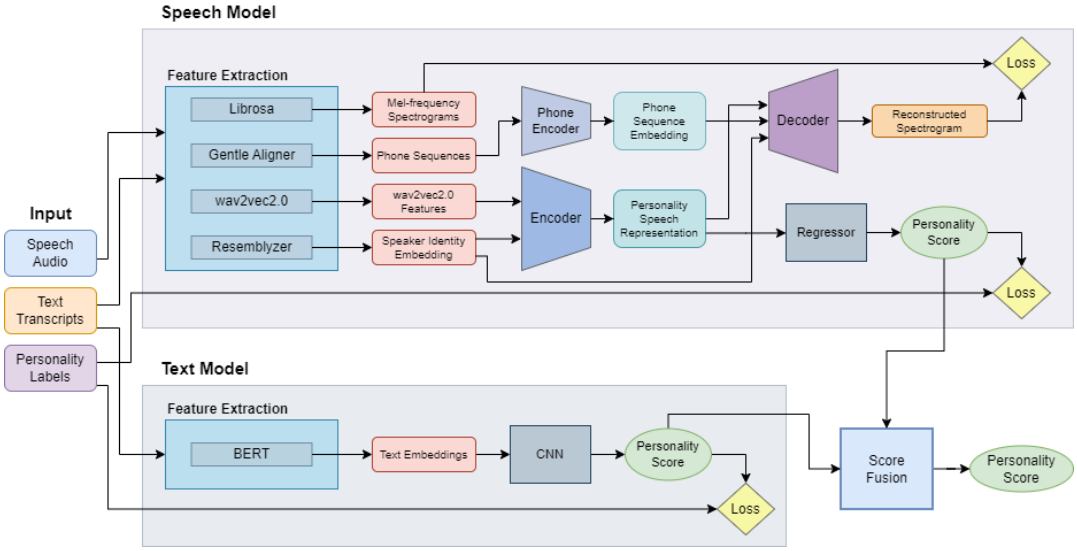

***Figure 1**. High level overview of the architecture of our multimodal personality recognition model. In our speech model, we perform feature extraction using Librosa [14], Gentle Aligner [15], wav2vec2.0 [16], and Resemblyzer [17]. These features are then input into our encoder-decoder as depicted above. The text model uses BERT [18] to generate text embeddings which are processed through a convolutional neural network. Using the output of both models, we are able to perform late fusion to get the final personality predictions.*

by confounders in the speaker's identity, which are then used in our regressor to predict personality scores.

To perform personality recognition using the audio modality, we analyze speech and speech features extracted from the audio and train it through the encoder-decoder model to generate speech representations that contain personality information. This is depicted in the "Speech Model" component of Figure 1 for the architecture of this model. Following the method in [9], we randomly crop our samples to 96-frame speech segments to perform feature extraction on the input to generate speech spectrograms, phone sequences, wav2vec2.0 features, and speaker identity embeddings.

The encoder concatenates the wav2vec2.0 feature and speaker identity embeddings and then processes them through three convolutional layers and two Bidirectional Long Short Term Memory (BLSTM) layers, followed by a downsampler to get the personality speech representation.

This representation is then input into the Decoder along with the phone sequence embeddings generated by the phone encoder and speaker identity embeddings to construct the Mel-frequency spectrograms. First, the decoder takes the speech representation of size from the encoder and upsamples it back to its previous size. This upsampled speech representation is then concatenated with the speaker identity embedding and the phone sequence embedding, then processed through an LSTM layer, three convolutional layers, two more LSTM layers, and finally, a fully connected layer to output a feature array that represents the Mel-frequency spectrogram.

We compare this with the ground-truth spectrograms using mean squared error loss. We minimize the loss between the reconstructed spectrograms, which allows us to the cross-representational aspect of this disentanglement method, with its ability to represent both high-level wav2vec2.0 features and low-level spectrogram features. Finally, we can disentangle features from the speaker identity and phone sequences irrelevant to personality analysis by reducing the size of the encoder's bottleneck.

At inference time, the audio sample is divided into 96-frame speech segments, and personality predictions are made for every segment in the speech sample. The final personality score is determined by taking the average of all the predicted scores within that sample.

3.3. Text Modality

In addition to the speech-based personality recognition, we also analyze the text features of the spoken dialogue during each session to perform text-based personality recognition.

During the data preprocessing, we included an ASR component that generates text transcriptions for each speech sample. Then, we use pretrained Transformer-based models [18] to extract features from each transcription which are used to train our text-based personality recognition model using a convolutional neural network.

Compared to the speech-based component, which randomly selects 96-frame speech segments from each sample, this text-based component generates predictions using the entire text-transcribed sample.

***Table 1**. Comparison of the results of our proposed method with challenge results from [8]. The values on the table represent MSE, in which case a lower value indicates better performance. *Results from the English subset*

Method	**O**	**C**	**E**	**A**	**N**	**Avg↓**
UDIVA Baseline [7]	0.744	0.794	0.886	0.653	1.012	0.818
SMART-SAIR Solution [20]	0.711	0.723	0.867	**0.548**	0.997	0.769
FGM Utrecht Solution [8]	0.752	0.687	0.917	0.671	1.098	0.825
Speech Model* (ours)	0.497	0.989	0.444	0.925	1.518	0.875
Text Model* (ours)	0.223	**0.399**	0.200	0.667	1.272	0.552
Fusion Model* (ours)	**0.195**	0.434	**0.185**	0.674	**0.862**	**0.470**

3.4. Multimodal Fusion

Both our speech modality and text modality produce personality score predictions using their respective models. To leverage both speech and text modes, we perform score fusion to combine the results of these two components.

Rather than assigning a single scalar value to each modality as proposed by [9], we instead introduce a weighted method that uses different weights for each individual trait. For each sample, we use the following:

$$\boldsymbol{p} = \frac{1}{\boldsymbol{w_1} + \boldsymbol{w_2}}(\boldsymbol{w_1} \odot \boldsymbol{p_s} + \boldsymbol{w_2} \odot \boldsymbol{p_t})$$

where $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ represent weight vectors for the different modalities and $\odot$ is the Hadamard or element-wise product.

4. Experiments

4.1. Conditions

In this work, we use the UDIVA dataset [7], which contains video recordings of dyadic interaction sessions and self-reported personality scores for each participant. These sessions were conducted in Spanish, Catalan, or English. In this work, we focus on the sessions of the dataset that were in English, discarding the sessions in Spanish and Catalan.

During our training and evaluation, we perform cross-validation to separate our training and validation sets across five folds. For each session, we randomly assign it to one of the folds such that it is part of the test set, and the remaining sessions make up the training set. This process is necessary to ensure that speakers in the test set are not found in the training set.

In the speech modality of the personality recognition model, we use the Adam optimizer with a learning rate at 1e-4 and default beta coefficients of 0.9 and 0.999. The batch size is set at 2, and the model is trained for 500,000 iterations at around 14 to 15 hours per fold. For the text modality of the personality recognition model, we similarly use the Adam optimizer with a learning rate at 1e-4 and default beta coefficients of 0.9 and 0.999. The batch size is set at 4, and the model is trained for 412,800 iterations for around 30-40 minutes per fold. Our models were implemented in PyTorch and trained on an NVIDIA Telsa P100 GPU.

4.2. Results

We evaluate the performance of our model by comparing the unimodal and fusion results to the challenge results presented in [8].

In Table 1, we observe that our model outperforms existing methods on nearly all personality traits. Our overall average mean squared error for the fusion model is 0.470, much smaller than that of the winning solution [20] of the ChaLearn challenge at 0.769. However, these results may not entirely represent the whole dataset because our model only examines the English sessions.

5. Ablation Studies

We perform a series of ablations on both the speech and text modalities to gain a better understanding of the contributions of each component in our system.

5.1. Bottleneck Experiments

We present several bottleneck configurations for the encoder by manipulating the values of the number of neurons in the BLSTM layers and the downsampling frequency. Our goal is to reduce the size of the encoder's output to be smaller than the size of the wav2vec2.0 feature array $[1024, 96]$. The results of these experiments are shown in Table 2.

Our results reveal that different traits perform differently depending on the bottleneck size, which may imply that the optimal size of the speech representation may differ among traits.

***Table 2**. Ablations on different bottleneck sizes in the speech encoder. The Bottleneck label refers to the size of the speech representation output by the encoder at the specified configuration. The values indicate MSE.*

Bottleneck	O	C	E	A	N	Avg↓
[256, 48]	**0.49**	1.34	0.52	0.84	1.48	0.94
[128, 24]	0.97	1.32	0.99	**0.81**	1.03	1.02
[64, 8]	0.73	1.23	0.83	0.96	**0.98**	0.95
[16, 2]	0.50	**0.99**	**0.44**	0.92	1.52	**0.87**

5.2. Encoder-only Experiments

To evaluate the performance and effectiveness of the cross-representational disentanglement in the speech model, we use a speech model that employs only an encoder to generate speech representations. As a result, this new encoder-only model does not use any disentanglement. For these experiments, we compare the results of our encoder-decoder method to the encoder-only method.

The results for evaluating the performance of the disentanglement method can be found in Table 3. We observe that in nearly all traits, the model with the Encoder-Decoder outperforms that of the Encoder-only architecture. Thus, we confirm that speech disentanglement using the speaker's identity features is effective for personality recognition.

Speech disentanglement on a speaker's identity was originally proposed in [9] to eliminate speaker information from the speech representations for emotion recognition. However, speaker identity plays a different role in personality recognition, as personality is a dimension of the speakers themselves. Thus, when we apply this method to personality recognition, rather than eliminating speaker identity, we instead disentangle speaker identity features from their personality. The results found in Table 3 indicate that this type of disentanglement is indeed effective, and we can eliminate aspects of speaker identity that distract from personality trait estimation.

5.3. ASR Analysis

We compare the performance of the text model trained using the ASR-generated text transcriptions with one trained using the provided transcriptions in the dataset.

From the results depicted in Table 4, we observe that in almost every trait, the model with the ASR transcripts outperforms that of the UDIVA Transcripts.

The transcripts in the UDIVA dataset were obtained through a third-party company and manually reviewed for cleanliness and data protection [8]. However, upon close inspection, the UDIVA dataset transcripts appear to correct grammar and omit filler words such as 'like' or 'uh' that the ASR can sometimes pick up on. Our ASR transcripts were much more consistent in transcribing the speech word-for-word.

Thus, we can conclude that our model can perform well using automatically generated text transcriptions from ASR rather than the given transcripts.

***Table 3**. Ablation on using disentanglement. We compare the disentanglement-employed Encoder-Decoder model and the Encoder-only without disentanglement. The values indicate MSE.*

	O	C	E	A	N	Avg↓
Encoder-Decoder	**0.27**	**0.64**	**0.27**	0.60	**0.81**	**0.52**
Encoder only	0.29	0.74	0.30	**0.59**	0.94	0.57

***Table 4**. Ablation results for the training on the provided text transcripts in the UDIVA dataset compared the transcripts generated using ASR. The values indicate MSE.*

	O	C	E	A	N	Avg↓
UDIVA	0.23	0.42	0.21	0.69	**1.27**	0.56
ASR	**0.22**	**0.40**	**0.20**	**0.67**	1.27	**0.55**

6. Conclusion

In this work, we presented a multimodal system for personality recognition that outperforms current existing methods evaluated on the UDIVA dataset. Employing representation learning for this task was very effective, since performing disentanglement using an autoencoder was able to help our system better pick up on personality cues within the target's speech audio signals. Furthermore, using ASR for text transcriptions helped increase performance with our text model. Finally, our weighting policy allowed us to individually balance the contributions from each modality depending on the trait. Using our system, we produced better results within the English subset of the UDIVA dataset, outperforming current methods that use visual features by overall 0.299 MSE.

7. Acknowledgements

This work was supported by JST CREST JPMJCR1687 and JST COI-NEXT JPMJPF2101, JAPAN.

References

[1] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, et al., "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 69–87, 2011.

[2] Boele de Raad and Marco Perugini, Big Five Assessment, Hogrefe & Huber Publishers, 2002.

[3] Niclas Kuper, Nick Modersitzki, Le Vy Phan, and John F. Rauthmann, "The dynamics, processes, mechanisms, and functioning of personality: An overview of the field," British Journal of Psychology, vol. 112, no. 1, pp. 1–51, 2021.

[4] Geoffrey Gaudi, Bill Kapralos, KC Collins, and Alvaro Quevedo, "Affective computing: An introduction to the detection, measurement, and current applications," in Advances in Artificial Intelligence-based Technologies, pp. 25–43. Springer, 2022.

[5] Pedro Branco and L Miguel Encarnacao, "Affective computing for behavior-based ui adaptation," in Proc. of Intelligent User Interface2004 Conf., Ukita, 2004.

[6] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 3s, dec 2019.

[7] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapes, et al., "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, January 2021, pp. 1–12.

[8] Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, Albert Clap´es, Johnny N´u˜nez, et al., "Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results," in Understanding Social Behavior in Dyadic and Small Group Interactions. 16 Oct 2022, vol. 173 of Proceedings of Machine Learning Research, pp. 4–52, PMLR.

[9] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda, "Multimodal emotion recognition with high-level speech and text features," in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021, pp. 350–357.

[10] Robert R. McCrae, The Five-Factor Model of personality traits: consensus and controversy, p. 148–161, Cambridge Handbooks in Psychology. Cambridge University Press, 2009.

[11] Le Vy Phan and John Rauthmann, "Personality computing: New frontiers in personality assessment," Social and Personality Psychology Compass, vol. 15, 06 2021.

[12] Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Bar´o, and Sergio Escalera, "Person perception biases exposed: Revisiting the first impressions dataset," in 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW), 2021, pp. 13–21.

[13] Ricardo Dar´ıo P´erez Principi, Cristina Palmero, Julio CS Jacques Junior, and Sergio Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," IEEE Transactions on Affective Computing, vol. 12, no. 3, pp. 607–621, 2019.

[14] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, et al., "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, 2015, vol. 8.

[15] Robert M Ochshorn and Max Hawkins, "Gentle aligner," https://lowerquality.com/gentle/, 2017.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[17] Corentin Jemine, "Resemblyzer," https://pypi.org/project/Resemblyzer, 2019.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). June 2019, pp. 4171–4186, Association for Computational Linguistics.

[19] Jonatas Grosman, "Fine-tuned XLSR-53 large model for speech recognition in English," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

[20] Hanan Salam, Viswonathan Manoranjan, Jian Jiang, and Oya Celiktutan, "Learning personalised models for automatic self-reported personality recognition," in Understanding Social Behavior in Dyadic and Small Group Interactions. 16 Oct 2022, vol. 173 of Proceedings of Machine Learning Research, pp. 53–73, PMLR.